Are Two Heads Better Than One?: An Experimental
Analysis of Group vs. Individual Decisionmaking

by

Alan S. Blinder and John Morgan,
Princeton University

CEPS Working Paper No. 70

May 2001

Abstract

We report the results of two sets of experiments comparing decisions make as individuals to those made in groups under majority and unanimity rule. The first setup posed a purely statistical problem devoid of any economic content: Subjects were asked to guess the composition of an (electronic) urn "filled" with blue balls and red balls. The second mimicked the problem faced by central bankers: Subjects were asked to steer an (electronic model of an) economy by manipulating the interest rate. The results were both striking and strikingly consistent. Neither experiment supported the commonly-held belief that groups reach decisions more slowly than individuals. Both experiments found that groups, on average, made better decisions than individuals. Surprisingly, groups outperformed individuals by almost exactly the same margin in each experiment. Finally, there were practically no differences between group decisions made by majority rule and group decisions made under a unanimity requirement.

1. Introduction and Motivation

Do decisions made by groups differ systematically from the decisions of the individuals who comprise them? That is a question infrequently asked by economists, even though economics is often characterized as the science of choice. As a profession, we analyze and glorify the virtues of freely-made, self-interested decisions. But those decisions are almost always individual choices: A consumer with a utility function and a budget constraint decides what to purchase; a firm, modeled as an individual decisionmaker, decides what will maximize its profits; a central banker with a well-defined loss function selects the optimal interest rate. While some economic literature, much of it derived from Arrow's (1963) seminal work, deals with group decisionmaking, it seems fair to say that economics dotes on *individual* choices: Some agent maximizes or minimizes something *by himself*.

Many real-world choices are, in fact, like that--even though the welfare of others may be taken into account. A consumer decides whether to buy milk or wine, mindful of the wants and needs of the family. A sole proprietor decides how many workers to hire, even if her objectives extend beyond her own profits. A central bank governor makes monetary policy decisions on his own, even though he has the best interests of society in mind. Decisions like these are the bread and butter of economic analysis.

But many decisions in real societies--including some quite important ones--are made by groups. Legislators, of course, make the laws. The Supreme Court

is a committee, as are all juries. Some business decisions, e.g., in partnerships or management committees, are made collectively, rather than dictatorially. And monetary policy decisions in many countries are made by committee rather than by a single individual. The latter is, in fact, the application that motivated this research. While one of us served as Vice Chairman of the Federal Reserve Board, he came to believe that economic models might be missing something important by treating monetary policy decisions as if they were made by a single individual maximizing a well-defined loss function. As Blinder (1998, p. 20) subsequently wrote:

> While serving on the FOMC, I was vividly reminded of a few things all of us probably know about committees: that they laboriously aggregate individual preferences; that they need to be led; that they tend to adopt compromise positions on difficult questions; and-- perhaps because of all of the above--that they tend to be inertial.

This sentiment reflects what is probably a widely-held view: that groups make decisions more slowly than individuals. One major question for this paper is: Is it true?

Why are so many important decisions entrusted to groups? Presumably because of some belief in collective wisdom. In a complicated world, where no one knows the "true" model or even all the facts, where data may be hard to process or interpret, and where value judgments may influence decisions, it may be beneficial to bring more than one mind to bear on a question. While it has been said that nothing good was ever written by a committee,[1] could it be that committees sometimes make better decisions than individuals?

---

[1] The Bible is often offered as an exception.

So these are the two central questions for this paper: Do groups reach decisions more slowly than individuals do? (We have never heard it suggested that groups decide faster.) And are group decisions, on average, better or worse than individual decisions?

Our approach is experimental. We created two laboratory experiments in which literally everything was held equal except the nature of the decisionmaking body--an individual or a group. Even the identities of the individuals were the same, since each experimental group consisted of five people who also participated as individuals. We therefore had automatic, experimental controls for what are normally called "individual effects." The laboratory setting also allowed us to define an objective function--known to the experimental subjects-- that distinguished better decisions from worse ones with a clarity that is normally unattainable in the real world. That is a huge advantage of the laboratory approach. The artificiality is, of course, its principal drawback.

The experiments themselves, which will be described in detail below, were very different.[2] The first setup, which is described in detail in Section 2 below, posed a purely statistical problem devoid of any economic content: Subjects were asked to guess the composition of an (electronic) urn "filled" with blue balls and red balls. The second, discussed in Section 3, mimicked the problem faced by central bankers: Subjects were asked to steer an (electronic model of an)

---

[2] The data and program code for both experiments are available on request.

economy by manipulating the interest rate. Participants in this experiment were required to have some elementary knowledge of macroeconomics.[3]

The results were both striking and strikingly consistent. Neither experiment supported the commonly-held belief that groups reach decisions more slowly than individuals. That certainly came as a big surprise to us; our priors were like seemingly everyone else's. And both experiments found that groups, on average, made better decisions than individuals. (Here our priors were much more diffuse.) Most stunningly, groups outperformed individuals by almost exactly the same margin in each experiment. In addition, the experiments unearthed another surprising finding: There were practically no differences between group decisions made by majority rule and group decisions made under a unanimity requirement. This, too, conflicted with our priors.

The remainder of the paper is organized into four sections. Section 2 describes the urn experiment and what we found. Section 3 does the same for the monetary policy experiment. Section 4 reports briefly on some mainly-unsuccessful attempts to model the group decisionmaking process, and Section 5 draws some tentative conclusions.

2. The Purely Statistical Experiment

2.1 Description of the Urn Experiment

Our first experiment placed subjects in a probabilistic environment devoid of any economic content, but structured to capture salient features of monetary

---

[3] Most of our subjects were Princeton University undergraduates. The requirement

policy decisions where possible. While such content-free problem solving may be of limited practical relevance, our motive was to create an experimental setting into which students would carry little or no prior intellectual baggage. While artificial in the extreme, this austere setup has an important virtue: It allows us to isolate the pure effect of individual versus group decisionmaking.

Specifically, the problem--which was identical for individuals and groups--was a variant of the classic "urn problem" in which subjects sample from an urn and then are asked to estimate its composition. In our application, groups of five students were placed in front of computers which were programmed with electronic "urns" consisting, initially, of 50% "blue balls" and 50% "red balls."

They were told that the composition of the urn would, at some randomly-selected point in the experiment, change to either 70% blue balls and 30% red balls, or to 70% red and 30% blue. Subjects were not told when the change would take place, nor in which direction--in fact, the latter is what they were asked to guess. But we did inform them of the probability law that governed the timing of the color change: The change was equally likely to occur just prior to any of the first 10 draws and would definitely occur no later than the 10th.[4]

We provided subjects with a clear objective function so that we could unambiguously distinguish better decisions from worse ones. This objective function weighted the two criteria on which the quality of decisionmaking would be judged--speed and accuracy--as follows. Subjects began each round with 40

---

was that they had taken Economics 101 or the equivalent.
[4] Random number generators determined both the direction of the change and its timing. Sampling was with replacement.

points "in the bank" and could earn another 60 points by correctly guessing the direction in which the urn's composition changed.[5] Subjects were allowed to draw as many "balls" as they wished before making their guess--up to an upper limit of 40, which was rarely reached.[6] However, they paid a penalty of one point for each draw they made *after* the urn changed composition, but *before* they guessed the majority color. (Call this the decision lag, L.) For example, if the composition changed on the 8th draw, and the subject guessed correctly after the 15th draw, L=7 and the score for that round would be 40+60-7=93. If the guess was incorrect, the score would be 40-7=33. A similar penalty was assessed if the subject guessed the composition *before* the change took place (a negative decision lag). Thus, if the composition was programmed to change on the 8th draw, but the guess came after the 4th, the subject would be penalized 4 points for guessing too soon. In sum, the objective function was:

(1) $S = 40 + 60C - |L|$,

where:

S = score (0-100 scale

C = a dummy variable   = 1 if guess is correct

= 0 if guess is incorrect

L = decision lag = T – N
T = the draw on which the composition changed (a random integer drawn from a uniform distribution on [1,10])
N = the draw after which the subject guessed the composition of the urn.

Before going further, a few remarks on the structure of the experiment are in order.

First, while the entire setup was devoid of substantive content, it was designed to evoke the nature of monetary policy decisionmaking. For example, policymakers never know for sure when macroeconomic conditions (analogous to the urn's composition) call for a change in monetary policy (a declaration that the composition has changed). Instead, they gradually receive more and more information (more drawings from the urn) suggesting that a change in policy may make sense. Eventually, enough such data accumulate, and policy is changed. Nor does anyone tell the central bank whether policy should be tightened or eased. (Is the urn now 70% red or 70% blue?) In principle, after the arrival of each new piece of data (after each drawing), policymakers ask themselves whether to adjust policy now or wait for more information--which is precisely what our student subjects had to do.

Second, changes from 50%-50% to 70%-30% color ratios are pretty easy to detect, but not "too easy."[7] Again, this aspect of the experimental design was meant to evoke the problem faced by monetary policymakers. Rarely are central bankers in a quandary over whether they should tighten or ease. The policy debate is usually over whether to tighten or do nothing, or over whether to ease or do nothing.

---

[7] This is a probabilistic statement. It is certainly possible to draw, say, equal numbers of blue and red balls when the urn is, say, 70% red. Indeed, we saw this happen during the experiment.

Third, the ratio 60:1 in the objective function determines the relative values of being accurate (60 points for getting the composition right) versus being fast (each additional draw costs 1 point). This ratio was set so high for two reasons. One is that it seems to us that accuracy--that is, getting the direction right--is vastly more important than speed in the monetary policy context. The other reason was that experimentation with this parameter taught us that quite a high ratio was needed to dissuade subjects from jumping the gun by guessing the color too soon. Students seemed extremely eager to decide, even on the basis of scant information. Despite the 60:1 ratio, we still believe that, on average, they made decisions too quickly.[8]

Fourth, 40 "free points" were provided on each round in order to make negative scores impossible. The lowest possible score on any round--1 point-- would be obtained by guessing incorrectly after 40 drawings when the change in composition occurred on the 1st draw.

The game was played as follows. Each session had five subjects, mostly Princeton undergraduates. Subjects were read detailed instructions (which they also had in front of them in writing) and then allowed to practice with the computer apparatus for about five minutes--during which time they could ask any questions they wished. Scores during those practice rounds were displayed for feedback, but not recorded. At the end of the practice period, all the machines

---

[8] However, the combinatorics of this problem are so complicated that we cannot prove that our hunch is correct--because we cannot solve analytically for the optimal strategy. We can, however, place a theoretical upper bound of 89.25 on the average score attainable using the optimal strategy. This upper bound is derived from employing the optimal strategy in the urn experiment when there is no uncertainty about the period that the urn changes composition.

were reinitialized, and each student was instructed to play 10 rounds of the game *alone*--without communicating in any way with the other students. Subjects were allowed to proceed at their own pace; clock time was irrelevant. When all five subjects had completed 10 rounds, the experimenter called a halt to Part One of the experiment.[9]

In Part Two, the five students gathered around a single computer to play the same game 30 times *as a group*. The rules were exactly the same, except that students were now permitted to communicate freely with one another--as much as they pleased. During group play, all five students received the group's common scores. There were 20 sessions in all, involving 100 subjects. In half of the sessions, decisions in Part Two were made by *majority rule*: The experimenter told the group that he would do nothing until he had instructions from at least three of the five students. In the other half of the experiments, decisions were made *unanimously*: The experimenter told the subjects that he would do nothing unless all five agreed.

After 30 rounds of group play, the subjects returned to their individual machines for Part Three, in which they played the game another 10 times alone. Following that, they returned to the group computer for Part Four, in which decisions were now made *unanimously* if they had been by majority rule in Part Two, or by *majority rule* if they had previously been under unanimity. Finally,

---

[9] The experimenters were Blinder and Morgan for the first few sessions, and then a graduate student, Felix Vardy, for the rest. In the urn experiment, we found that while qualitative results were unaffected by the identity of the experimenter, there was a significant level effect in scores: subjects on average did worse in the first two sessions than in subsequent sessions – both

Part Five concluded the experiment with 10 additional individual plays. Table 1 summarizes the flow of each session.

Table 1

The Flow of the Urn Experiment

Instructions

Practice Rounds (no scores recorded)

Part One: 10 rounds played as individuals

Part Two: 30 rounds played as a group under majority rule
(alternatively, under unanimity)

Part Three: 10 rounds played as individuals

Part Four: 30 rounds played as a group under unanimity
(alternatively, under majority rule)

Part Five: 10 rounds played as individuals

Students are paid in cash, fill out a short questionnaire, and leave.

Thus each session consisted of 90 rounds--30 played individually and 60 played as a group. Since we ran 20 sessions in all, we have data on 1200 group rounds (20 x 60) and 3000 individual rounds (20 x 30 x 5).[10] Sessions normally lasted a bit under an hour, and subjects typically earned around $15--compared to a theoretical maximum of $18 for a perfect score.

---

in groups and as individuals. There were no experimenter effects in the monetary policy experiment.
[10] This is not quite true. Due to a computer glitch that we have been unable to figure out, we lost a total of 37 observations--all from individual play in Part Five.

2.2 The Three Main Hypotheses

While several subsidiary questions will be considered below, our interest focused on the three main hypotheses mentioned in the introduction, especially the first two:

H_1: Groups make decisions more slowly than individuals.

As noted earlier, the decision lag, L, can be positive (as was true 92.3% of the time) or negative. The main idea that motivated this study was the widely-believed notion that groups take longer to make decisions than individuals do. Note that we measure the decision lag in number of draws--that is, the amount of information required before a decision is reached--not in elapsed clock time, which we deemed to be irrelevant and did not measure.

Specifically, let L_i be the average lag for the i-th individual in the group (i=1,..., 5) when he or she plays the game alone, and let L_G be the average lag for those same five people when making decisions as a group. Under the null hypothesis of no group interaction, which we expected to reject, the group's mean lag would equal the average of the five individual mean lags:

$$L_G = (L_1 + L_2 + L_3 + L_4 + L_5)/5.$$

Furthermore, under this null, a simple t-test for difference in means is the appropriate test.[11]

Surprisingly, the hypothesis of equality could not be rejected. The overall mean lag was indeed slightly longer for groups than for individuals (6.60 draws

---

[11] We thank Alan Krueger for reminding us of this simple consequence of the Neymann-Pearson lemma.

versus 6.40), but that small difference is not significant at conventional levels (t=1.1), even with thousands of observations. Histograms for the variable L for individuals and groups look strikingly alike. (See Figure 1.)

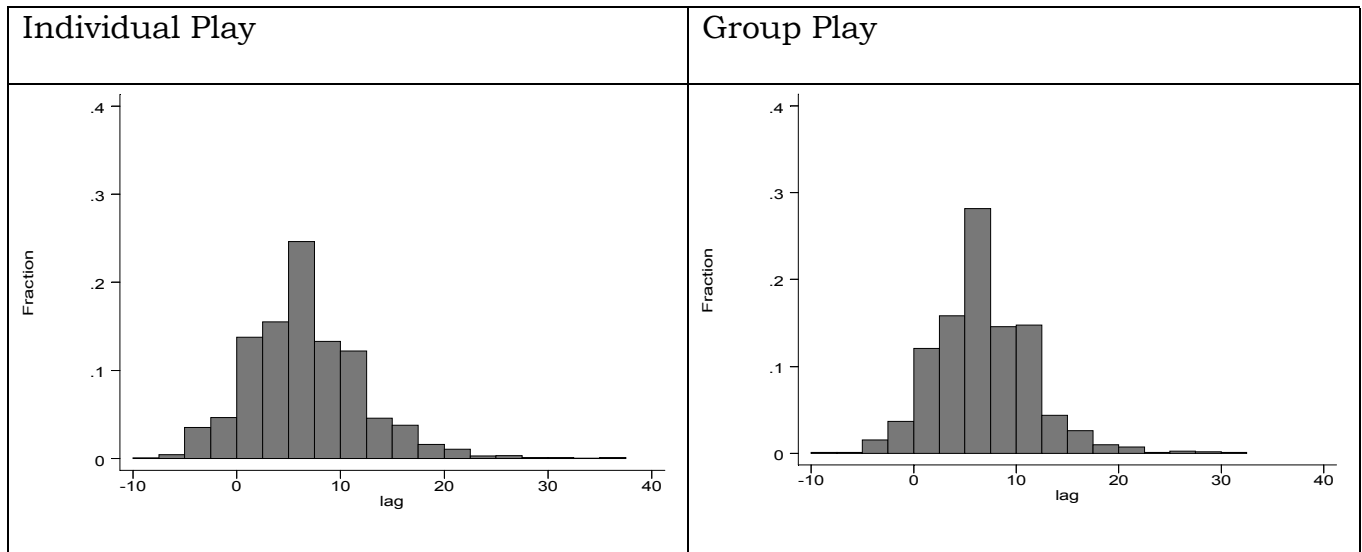| Individual Play	Group Play

Figure 1: Histograms of Lag in Urn Experiment

The individual distribution gives the impression of a mean preserving spread on the group distribution. If we eliminate the first ten rounds of each treatment (individual, majority rule, and unanimity) as a crude adjustment for learning,[12] the result is even more surprising: Group decisions were actually made slightly faster than individual decisions (6.89 draws versus 6.99), although the difference is far from significant (t-statistic = 0.5).

Looking at the test statistic $z = L_G - (L_1+L_2+L_3+L_4+L_5)/5$ for each of the sessions separately, we find that we can reject the null hypothesis of equality

___
[12] We will have much more to say about learning later, including some evidence that virtually all the learning is finished by roughly the 10th round of play.

13

(z=0) at the 5% level seven times out of 20. That sounds like a lot of evidence against the null. But the group made decisions faster than the individuals comprising the group in three of these cases and slower in four. So there is really no systematic difference in speed.

The overall conclusion, then, is a surprising one: Something that "everyone knows"--that group decisionmaking takes longer--is not supported by these experimental data.[13]

H_2: Groups make better decisions than individuals.

A quite different hypothesis concerns the *quality* of decisionmaking, rather than the speed. Do groups make better decisions than individuals? This idea may not come naturally to economists, since our discipline glorifies individual decisionmaking. Furthermore, in this particular experimental setup, every subject has the same objective function and receives the same information. So, were they all to behave like *homo economicus*, they would make identical decisions.

In reality, different people placed in identical situations do not always reach the same decisions. Furthermore, as we observed in Section 1, many important economic and social decisions in the real world are assigned to groups rather than to individuals. Presumably, there is a reason.

In any case, the hypothesis that groups outperform individuals is strongly supported by the experimental data. Remember, we designed the experiment to

---

[13] As noted earlier, we define "taking longer" in this context as requiring more drawings before reaching a decision, not as taking more clock time. While we

yield an unambiguous measure of the quality of the decision: S ("score"), as defined in equation (1). In the overall sample, the average score attained by groups was 86.8 (on a 1-100 scale), versus only 83.7 for individuals. The difference is highly significant statistically (t = 4.3). More important, it seems to be economically meaningful: Groups did 3.7% better, on average.[14] Interestingly, this gap narrows to 2.3%, but remains significant (t=2.4), when we drop the first 10 observations of each treatment to allow for learning.

Obviously, since the mean lags are statistically indistinguishable, the groups must have acquired their overall edge through *accuracy* rather than through *speed*. Specifically, the groups guessed the urn's composition correctly 89.3% of the time whereas the individuals got the color right only 84.3% of the time. Considering that the experimental apparatus was set up to make guessing the correct composition relatively easy, this gap of 5 percentage points is sizable. Look at it this way: The error rate (frequency of guessing the wrong color) was 15.7% for individuals, but only 10.7% for groups. The difference in performance is also highly significant statistically (t=4.2). However, the gap drops to only 2.9 percentage points (with t=2.2) when the first 10 observations of each treatment are excluded.

In brief, we find that group decisions are more accurate without being slower. Maybe two heads (or, in this case, five) really are better than one.

---

did not keep systematic data on this, we are quite certain that group decisions took longer on the clock.

[14] That difference is about 72% of the standard deviation across individual mean scores.

H₃: Decisions by majority rule are made faster than under a unanimity requirement.

Before we ran the experiment, it seemed obvious that requiring unanimous agreement would slow down the group decisionmaking process, relative to using majority rule. But observing the subjects interacting face-to-face in real time showed something quite different. If you observed the game without having heard the instructions, it was hard to tell whether the game was being played under the unanimity principle or under majority rule. Perhaps it was peer group pressure, or perhaps it was simply a desire to be cooperative.[15] But, for whatever reason, majority decisions quickly evolved into unanimous decisions. In almost all cases, once three or four subjects agreed on a course of action, the remaining one or two fell in line immediately.[16]

In fact, and quite surprisingly, decisions under the unanimity requirement were actually made faster, on average, than decisions under majority rule (mean L=6.34 versus 6.85). The difference is significant at the 5% level in a one-tail test. However, there was no significant difference between the two group treatments in either decisionmaking accuracy (C) or quality (S). The composition of the urn was guessed correctly 89.2% of the time under majority rule and 89.5% of the time under unanimity.

Thus, in most of what follows, we will pool data from the majority-rule and unanimity treatments. The data support such pooling.

---

[15] Students typically did not know one another prior to the experiment, though in some cases, purely by chance, they did.

2.3 Other Results

Learning

Having mentioned the issue of learning several times, we now turn to it explicitly. The game is rather cumbersome to describe in words, but is extremely easy to play "once you get the hang of it." So we suspected that there would be learning effects, at least in the early rounds: Students would get better at the game as they played it more (up to a point). This is why we began each experimental session with a "practice period" in which subjects could familiarize themselves with the apparatus. But it was clear to us that many students were still not fully comfortable with the game when play started "for real."

While we performed a variety of simple statistical tests for learning, Figure 2 probably displays the results better than any regressions or t-tests. To construct this graph, we partitioned the data by round, reflecting the chronological order of play. There are 90 rounds in each session--30 played as individuals, and 60 played as groups (see Table 1). So, for example, we have 100 observations (20 sessions times five individuals in each) on each of the first 10 rounds, 20 observations on each of rounds 11-40 (the 20 groups), and so on. Figure 2 displays the mean score by round; vertical lines indicate the points where subjects switched from individual to group decisionmaking, or vice-versa.

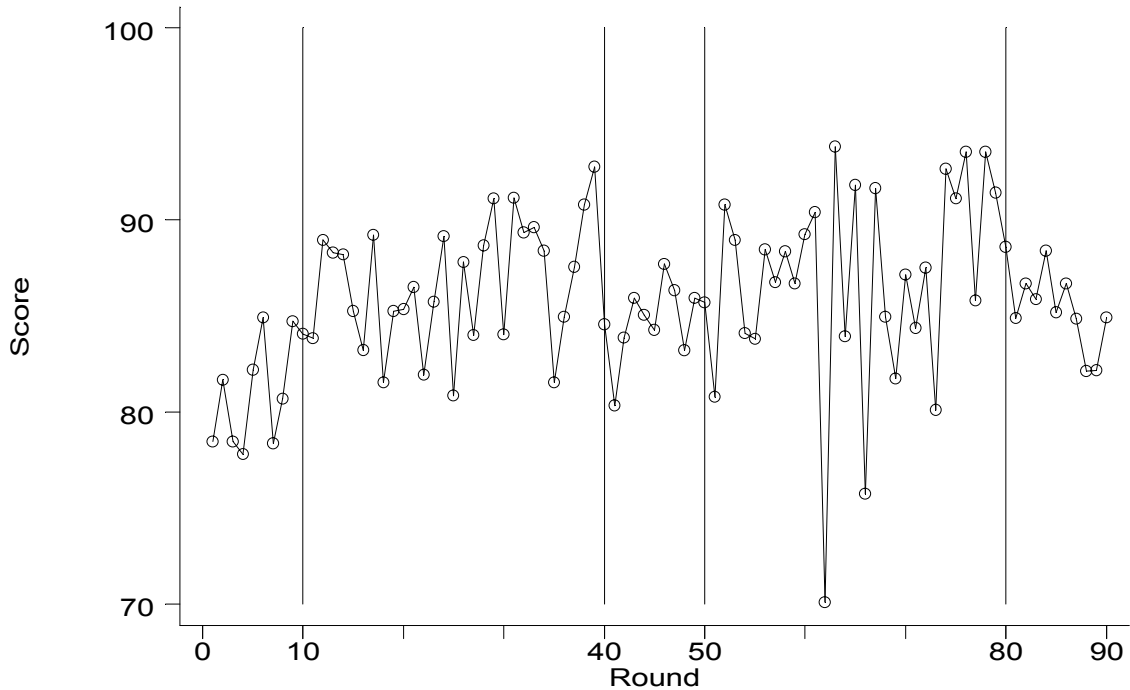[16] One student noted that her group unanimously agreed to decide by majority vote.

Figure 2: Mean Score by Round in Urn Experiment

If there are systematic learning effects, scores should improve as we progress through the rounds. The figure shows clear evidence of learning over the first 10-12 rounds, but none thereafter. In addition, it is evident that average performance jumps upward when we switch from individual to group play (the vertical lines at 10 and 50), and jumps downward when we switch from group to individual play (the vertical lines at 40 and 80). All four of these changes are statistically significant. In sum, the figure (and related statistical tests) suggest that learning occurred, but was limited to the early rounds, and was dwarfed by the difference in quality between individual and group decisions.

It is natural to wonder whether learning mostly affects speed (the decisionmaking lag, L) or accuracy (whether the urn's composition is guessed correctly, C). The answer is both, though in different ways--as Figures 3 and 4

show. Interestingly, Figure 3, which displays the mean decision lag, suggests the presence of learning throughout the experiment; there is a clear trend toward waiting longer before guessing the dominant color.[17] But Figure 4, which shows the percentage of correct guesses, looks a lot like Figure 2--learning ends after the first 10-12 rounds. The reason is clear from equation (1): In computing the score, C (correct) gets 60 times the weight of L (lag). Had we weighted L more heavily, a clearer indication of learning throughout each session would have emerged.
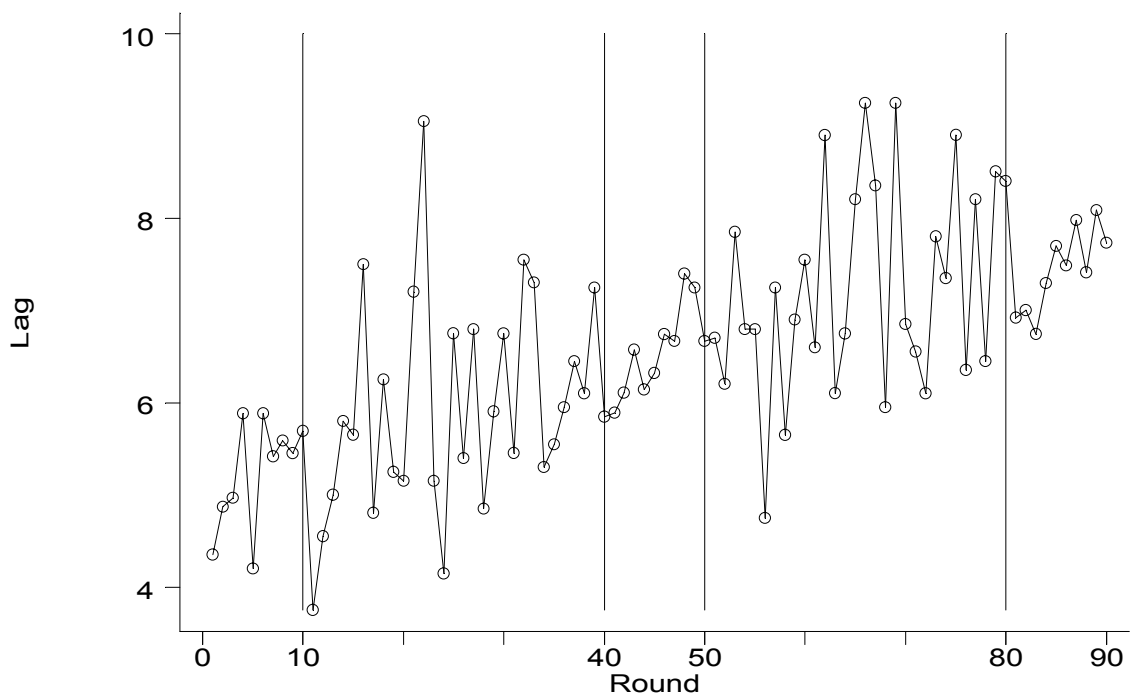
Figure 3: Mean Lag by Round in Urn Experiment

---

[17] Remember, we strongly believe that subjects tended to "jump the gun." So longer average lags are presumptively better. Indeed, several students observed that they learned to wait longer after playing as a group.
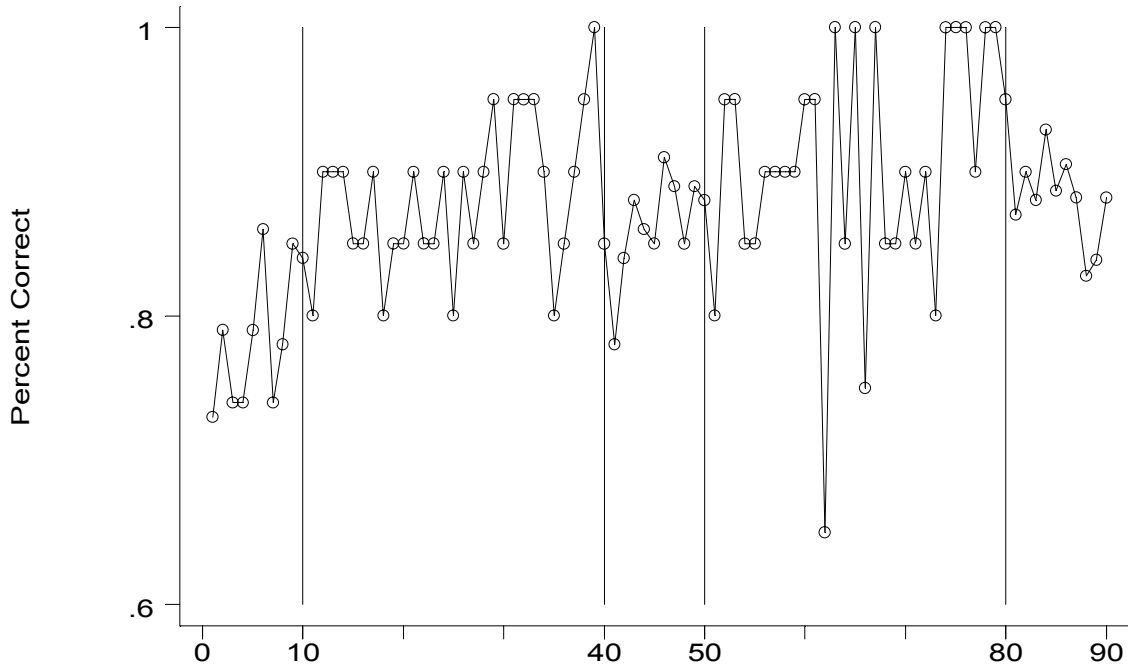
Figure 4: Mean Percent Correct in Urn Experiment

Experimental Order

In any experimental design, there is always a danger that results may be affected by the ordering of parts of the experiment. That is precisely why we arranged the parts of the experiment as we did: to have group play both precede and follow individual play, and to have unanimity both precede and follow majority rule. Nonetheless, the question remains: Does ordering matter?

Unfortunately, there is a little evidence that it does. Consider the scores obtained in the second 30 rounds of group play (600 observations from rounds 51-80). If the groups played first under the unanimity rule and then under majority rule (300 observations, the mean score was 88.7. If the order was

reversed, the mean score fell to 85.2. The difference is significant by conventional standards (t=2.4, p = 0.018), and we have no explanation for it.[18]

Fortunately, this puzzling finding was not replicated in the individual data, so we are inclined to treat it as a fluke. Rounds 41-50 and 81-90 of individual play took place *after* the subjects' first experience with group play. If their initial group experience was under unanimity, the *individual* scores in subsequent rounds averaged 84.2; but if that initial group experience was under majority rule, subsequent individual scores averaged 85.8. That difference, while not quite significant (t=1.8, p= .074), goes in the opposite direction from what we found in group play. So, on balance, experimental order does not appear to have much of an effect on the results.

Differences between men and women

Our student volunteers were 51.6% female and 48.4% male--which is a few percentage points more female than the Princeton student body. The sex composition of each group was a matter of chance: Students signed up for the time periods they preferred, and we made no effort to control the composition of any group. The resulting distribution of the 20 groups by sex turned out to correspond roughly to what would be expected from random selection: For example, ten groups had a male majority and 10 had a female majority.

Did men and women play the game differently? When playing as individuals, females on average made decisions slightly faster than males (mean

---

[18] Remember that, on average, there was no significant difference in scores between unanimity and majority rule.

lag = 6.22 draws versus 6.58 for males; t=1.8), but noticeably less accurately (percentage correct = 82.7 versus 86.0 for males; t=2.5). On balance, men's' scores were better by 2.1%--a difference that is significant at conventional levels (t=2.2). While we did not formulate any priors on differences by sex, we probably would have guessed that men were *more* inclined to "shoot from the hip" than women. But the data indicate otherwise.

We can also compare the performances of groups with a majority of men versus groups with a majority of women. Here no statistically significant differences emerged in overall score, lag length, or percent correct. Thus males do better alone, but male-dominated groups do not outperform female-dominated groups. A suggestive inference, which is tantalizing but highly tentative, is that women gain more from group interactions than do men.

3. The Monetary Policy Experiment

As noted in the introduction, we designed our urn experiment to capture many of the features of monetary-policy decisionmaking--except that it made no reference whatsoever to monetary policy (nor to any other real-world context). Our second experiment put the context back into the problem by asking subjects to assume the role of monetary policymaker explicitly. For this reason we added a prerequisite in recruiting subjects: They had to have taken at least one course in macroeconomics. Otherwise, we tried to make the mechanics of the monetary-policy experiment resemble the urn experiment as closely as possible.

3.1 Description of the Monetary Policy Experiment

Just as before, we brought students into the laboratory in groups of five, and we ran twenty sessions. But since each round of play took much longer, the groups only played the monetary policy game 20 times as individuals (versus 30 in the urn experiment), 10 times under majority rule, and 10 times under unanimity (versus 30 each in the urn experiment). Despite the much smaller number of plays, sessions in the monetary policy experiment typically lasted considerably longer: about 90 minutes.

The setup was as follows. We programmed each computer with a simple two-equation macroeconomic model that approximates a canonical model popular in the recent theoretical literature on monetary policy,[19] choosing (not estimating) parameter values that resembled the U.S. economy:

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

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

Equation (2) can be thought of as a reduced form combining an IS curve with Okun's Law. Specifically, U is the unemployment rate, and the assumed "natural rate" is 5%. Since i is the nominal interest rate and π is the rate of inflation, the term $i_t - \pi_t - 5$ connotes the deviation of the *real* interest rate from its equilibrium or "neutral" value, which is also set at 5%.[20] Higher (lower) real interest rates will push unemployment up (down), but only gradually. Our

[19] See, for example, Ball (1997) and Rudebusch and Svensson (1999).
[20] The neutral real interest rate is defined as the real rate at which inflation is neither rising nor falling. See Blinder (1998, pages 31-33).

experimental subjects, playing the role of the Federal Reserve, controlled only the *nominal* interest rate, not the *real* interest rate.

The G_t term connotes the affect of fiscal actions on unemployment and is the random event, analogous to the urn changing composition, that our experimental monetary policymakers are supposed to recognize and react to. G starts at zero and randomly changes to either +0.3 or -0.3 sometime within the first 10 periods. When this happens, it changes unemployment by that amount, but in the opposite direction (see equation (2)). Prior to the shock, the model's steady-state equilibrium is (U=5, i-π=5). Because the long-run Phillips curve is vertical, any constant inflation rate can be a steady state. But we always began the experiment with inflation at 2%--which is the inflation target. The shock changes the "neutral" real interest rate to either 6% or 4%, as is apparent from the coefficients in equation (2). Our subjects were supposed to react to this event, presumably with a lag, by raising or lowering the nominal interest rate.

Equation (3) is a standard accelerationist Phillips curve. Inflation depends on the lagged unemployment rate and on its own four lagged values, with weights summing to one. The weighted average of past inflation rates can be thought of as representing expected inflation, but the model does not demand this interpretation. The coefficient on the unemployment rate was chosen to match empirically estimated Phillips curves for the United States.

Finally, the two stochastic shocks, e_t and w_t, were drawn from uniform distributions on the interval [-.25, +.25].[21] Their standard deviations are approximately 0.14, or about half the size of the G shock. This choice, like the 70-30 composition of the urn, controls the "signal to noise" ratio in the experiment. (More on this below.)

Monetary policy affects inflation only indirectly in this model, and with a distributed lag that begins two periods later. A change in i_t affects U starting in period t+1 (see equation (2)), and that in turn affects π with a distributed lag that begins in period t+2 (see equation (3)). All of our subjects understood that higher interest rates reduce inflation and raise unemployment with a lag, and that lower interest rates do just the reverse.[22] But they did not know any details of the model's specification, coefficients, or lag structure.

Stabilizing such a system can be rather tricky. Because equation (3) builds in a unit root, the model will diverge from equilibrium when perturbed by a G shock--unless it is stabilized by monetary policy. But the lags make the divergence pretty gradual. One useful way to think about this dynamic instability is as follows. Start the system at equilibrium with U=5, π=2, and i=7, as we did. Now suppose G rises to 0.3. By (2), the neutral real rate of interest increases to 6%; so the initial real rate (5%) is now lower than neutral--and hence expansionary. With a lag, inflation begins to rise. If the central bank fails to raise

---

[21] The distributions were uniform, rather than normal, for programming convenience.
[22] Remember, all of our subjects had at least some exposure to basic macroeconomics. Lest they had forgotten, the instructions reminded them that raising the rate of interest would lower inflation and raise unemployment, while lowering the rate of interest would have the opposite effects.

the nominal interest rate, the real rate falls further--stimulating the economy even more.

Each play of the game proceeded as follows. We started the system in steady state equilibrium with $G_t=0$, current and lagged nominal interest rates at 7% (reflecting a 5% real rate and a 2% inflation target), lagged U at 5%, and all lags of π at 2%. The computer selected values for the two random shocks and displayed the first-period values, U_1 and π_1, on the screen for the subjects to see. (Normally, these were quite close to the optimal values of U=5% and π=2%.) For each subsequent period, new random values of e_t and w_t were drawn, thereby creating statistical noise, and the lagged variables that appear in equations (2) and (3) were inherited from the past. The computer would calculate U_t and π_t and display them on the screen, along with all past values. Subjects were then asked to choose an interest rate for the next period, and the game continued.

No time pressure was applied; subjects were permitted to take as much clock time as they pleased to make decisions. At some period chosen at random from a uniform distribution between t=1 and t=10, G_t was either raised to +0.3 or lowered to -0.3. (Whether G rose or fell was also decided randomly.) Students were not told when G changed, nor in which direction. But they were told the probability laws that governed the changes. All this is just as it was in the urn problem.

Even though our primary interest was in the decision lag--the number of periods it took for subjects to react to the change in G, we did not stop the game when the interest rate was first changed because this seemed unnatural in the

monetary-policy context. Instead, each play of the game continued for 20 periods. (Subjects were told to think of each period as a quarter.)

To evaluate the quality of the decisions, we needed a loss function. While quadratic loss functions are the rule in the academic literature, they are rather too difficult for subjects to calculate in their heads. So we used an absolute-value function instead. Specifically, subjects were told that their score for each quarter would be:

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

and the score for the entire game (henceforth, S) would be the (unweighted) average of s_t over the 20 quarters. The coefficients in (4) scale the scores into percentages--giving them a ready, intuitive interpretation. Equal weights on unemployment deviations and inflation deviations were chosen to facilitate mental calculations: Every miss of 0.1 cost one point. Thus, for example, missing the unemployment target by 0.5 (in either direction) and the inflation target by 0.7 would result in a score of 100 - 12 = 88 for that period. At the end of the entire session, scores were converted into money at the exchange rate of 25 cents for each percentage point. Subjects typically earned about $21-$22 out of a theoretical maximum of $25.

Finally, we "charged" subjects a fixed cost of 10 points each time they changed the rate of interest, regardless of the size of the change.[23] The reason is as follows. The random shocks, e_t and w_t, were an essential part of the experimental design because, without them, the changes in G_t would have been

---

[23] To keep things simple, only integer interest rates were allowed.

trivial to observe: No variable would ever change until G did. After some experimentation, we decided that random shocks with standard deviations about half the size of the G shock made it neither too easy nor too difficult to discern the G_t "news" amidst the e_t and w_t "noise."

But this decision created an inference problem: Our subjects might receive several false signals before G actually changed. For example, a two-standard-deviation e shock appears just like a negative G shock, except that the latter is permanent while the former is transitory. (The random shocks were iid.) Furthermore, subjects knew neither the size of the G shock nor the standard deviations of e and w; so they had no way of knowing that a two-standard-deviation disturbance would look (at first) like a G shock.

In some early trials designed to test the apparatus, we observed students moving the interest rate up and down frequently--sometimes almost every period. Such behavior would make it virtually impossible to measure (or even to define) the decision lag in monetary policy. So we instituted a small, 10-point charge for each interest rate change. Ten points is not much of a penalty--averaged over a 20-period game, it amounts to just 0.5%. But we found it was large enough to deter most of the excessive fiddling with interest rates. It also had the collateral benefit of making behavior a bit more realistic.[24] The Fed does not jigger the interest rate around every quarter, presumably because it perceives some cost in doing so that is not captured in equation (4).

---

[24] With one exception: Since the game terminated after 20 periods, students generally concluded that it was not worth paying 10 points to change the rate of interest in one of the last few periods.

The sequencing of the monetary policy game closely followed the sequencing of the urn experiment, and is shown in Table 2:

Table 2

The Flow of the Monetary Policy Experiment

Instructions

Practice Rounds (no scores recorded)

Part One: 10 rounds played as individuals

Part Two: 10 rounds played as a group under majority rule
    (alternatively, under unanimity)

Part Three: 10 rounds played as individuals

Part Four: 10 rounds played as a group under unanimity
    (alternatively, under majority rule)

Students are paid in cash, fill out a short questionnaire, and leave.

The ground rules were the same as in the urn experiment: Students could communicate freely, as much as they wished, during group play, but could not communicate with one another during individual play.

Comparing Tables 1 and 2 reveals two differences. First, there is no "Part Five" in which students finish by playing the game as individuals yet again. Hence we obtained only 20 individual observations per subject, or 2,000 in all. Second, we have many fewer group observations--just 20 per session, or 400 in all. Both changes were dictated by time constraints. Because the monetary policy game requires a great deal more thought than the urn problem, each round takes longer. Furthermore, each play of the monetary policy game always lasted 20

periods, whereas the urn problem often terminated after fewer than 10 draws. It was unrealistic to ask subjects to commit more than two hours of their time,[25] and 40 plays of the game were about all we could count on finishing within that time frame.

3.2 The Three Main Hypotheses

We were gratified to find that the monetary policy experiment--which was what originally motivated this research--produced exactly the same answers to our three main questions as the urn experiment. Remember, the urn problem was specifically designed to strip away any relevant background knowledge or institutional baggage in order to focus squarely on the decisionmaking process *per se*. But real-world decisions are not like that. Actual decisionmakers always carry into the room a wealth of experience, knowledge, prejudices, etc. Certainly, that is true of monetary policymakers. To find precisely the same results in these two very different contexts gives us some confidence that we have discovered something real.

Now to the specifics. Remember, our first and most crucial hypothesis was:

H₁: Groups make decisions more slowly than individuals.

The lags in the monetary policy game were actually quite short, averaging just over 2.4 "quarters" over the 2400 observations. In fact, many subjects "jumped the gun" by moving interest rates before G had changed. (This happened in 15.2% of all cases.) However, the group decisions were made slightly *faster*,

---

[25] Although sessions normally took closer to 1 1/2 hours, we insisted that subjects agree to commit two hours, since the premature departure of even one subject would ruin an entire session.

with a mean lag of just 2.30 periods (with standard deviation 2.75) versus 2.45 periods (with standard deviation 3.50) for the individual decisions. This scant 0.15 difference, even though it goes in a direction opposite the null hypothesis, is not close to being statistically significant (t=0.78, p=.22 in a one-tailed test).

Figure 5 displays the histograms of the variable L (the decision lag) for individual and group play. As in the urn problem, the former looks like a mean-preserving spread on the latter. Hence, once again, we find no support whatsoever for the seemingly-obvious hypothesis that groups decide more slowly than individuals.
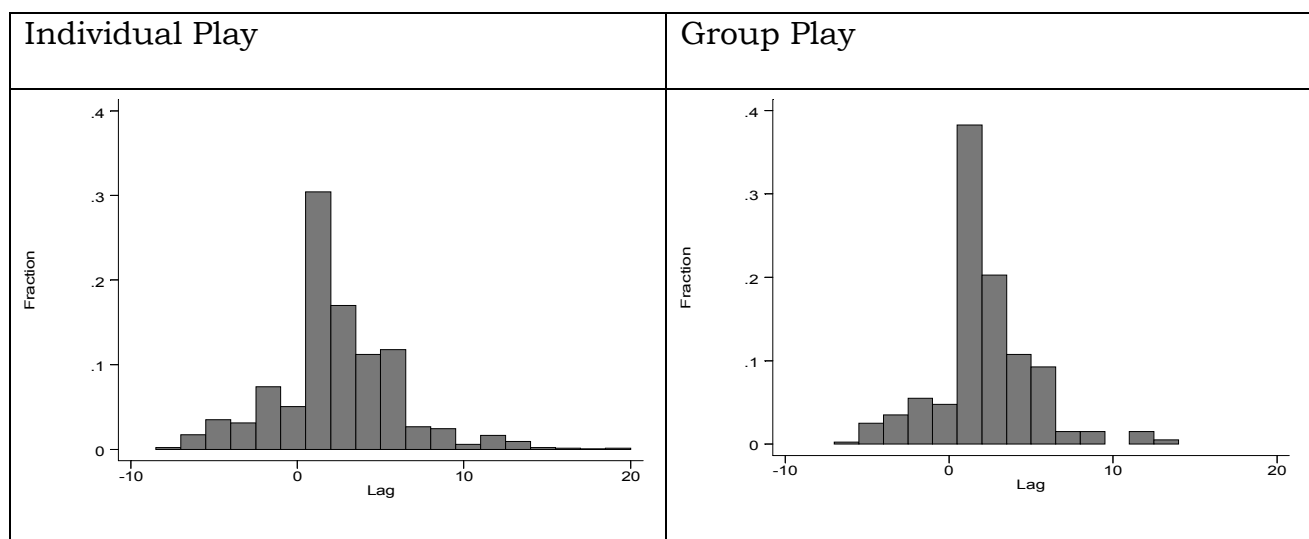
Individual Play	Group Play

Figure 5: Histograms of Lag in Monetary Experiment

H_2: *Groups make better decisions than individuals.*

Remember, we scored (and paid) our *faux* monetary policymakers according to how well they kept unemployment near 5% and inflation near 2% over the entire 20-quarter game. As in the urn problem, scores were quite high--

almost 86% on average. (We designed the experiment this way.) But, also as in the urn experiment, the groups did better than the individuals. The mean score over the 400 group observations was 88.3% (with standard deviation 4.7%), versus only 85.3% (standard deviation 10.1%) over the 2000 individual observations. The difference is economically meaningful and highly significant statistically (t=5.9). Perhaps most strikingly, the 3.5% performance gap between groups and individuals almost exactly matches what we found in the purely statistical urn experiment (a 3.7% gap). We were, frankly, stunned to find essentially the same average performance improvement in two such different experimental settings. Even if we had been determined to "rig the deck" to make the two performance gaps come out the same, we would have had no idea how to do so.

So, once again, we found that group decisions were superior to individual decisions without being slower--which suggests that group decisions dominate individual decisions.

We can also construct a variable analogous to the dummy variable C--for whether the color was guessed correctly--in the urn experiment. Specifically, when G rose, subjects were supposed to *increase* interest rates; and when G fell, subjects were supposed to *decrease* interest rates. So define the variable C ("correct") for the monetary policy experiment as 1 if the first interest rate change is in the same direction as G changes, and 0 if it is not.[26] Unlike in the urn experiment, the variable C does not enter the loss function *directly*. But we

certainly expect subjects to attain higher scores when their first move is in the right direction. In fact, the simple correlation between moving in the correct direction initially and final score is 0.37.

Here, once again, groups outperformed individuals by a notable margin. The average value of C was .843 for individuals but .905 for groups. This difference is highly significant statistically (t=3.6, p=0.003). Economically, it is even more noteworthy. When playing as individuals, our ersatz monetary policymakers moved interest rates in the wrong direction 15.7% of the time. When acting as a group, however, these same people got the direction wrong only 9.5% of the time. Finally, the margin of superiority of groups over individuals (6.2 percentage points) is again strikingly similar to what we found in the urn experiment (5.0 percentage points).

H_3: Decisions by majority rule are made faster than under a unanimity requirement.

As noted earlier, we were surprised to find almost no differences between groups operating under majority rule and groups operating under the unanimity principle in the urn experiment. In fact, contrary to our priors, decisions were made slightly faster under the unanimity requirement. By the time we got to the monetary policy experiment, we expected no differences--which is just what we found.

---

[26] For this purpose, we look only at the *first* interest-rate change. In most plays of the game, rates were changed several times.

Observationally, it was hard to tell whether groups were using majority voting or unanimous agreement to make decisions. Statistically, the mean lag under unanimity was indeed slightly longer than under majority rule--2.4 periods versus 2.2 periods--in conformity with H_3, but in contrast to what we found in the urn problem. However, the difference did not come close to statistical significance (t=0.9). When it came to average scores, the two decision rules finished in what was essentially a dead heat (just as they had in the earlier experiment): 88.0% under majority rule, and 88.6% under unanimity. Hence, we are again comfortable with pooling the majority-rule and unanimity results.

3.3 Other findings

Learning

In the urn problem, we detected sizable learning effects in the early rounds that were, however, swamped by the effect of changing from individual to group play. So scores rose whenever we moved from individual to group play and fell when we moved from groups back to individuals. That is essentially--but not quite--what we found in the monetary policy experiment as well.

Partitioning the 2400 observations by round (which now runs from 1 to 40), Figure 6 suggests a trend toward longer decision lags for about the first 30 rounds. Looked at more carefully, however, the data show an upward trend within the individual rounds (1-10 and 21-30) but no trend whatsoever within the group rounds (11-20 and 31-40).
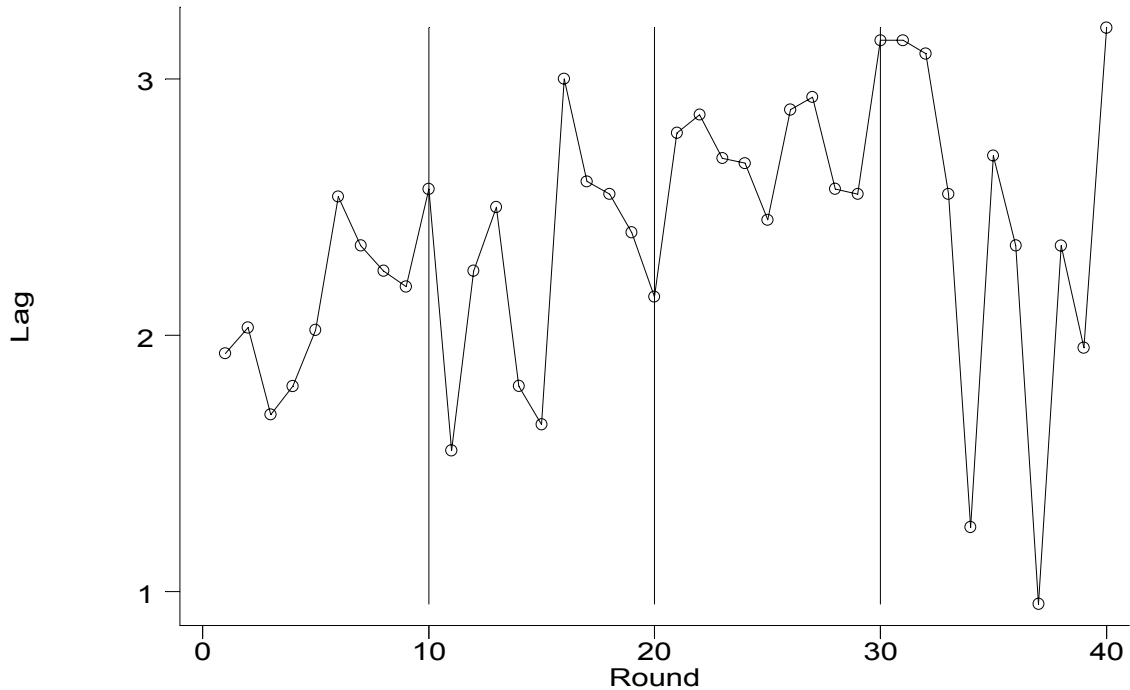
Figure 6: Mean Lag by Round in Monetary Experiment

Presuming that longer lags imply that people have "learned" assumes that decisions are typically made too quickly. We do hold this view. As we have said several times, we believe that subjects tended "jumped the gun." But a more relevant test is surely to inspect the average scores by round--as is done in Figure 7. Here we see a rather different, and quite striking, pattern. There is no indication whatsoever of any learning within the first 10 rounds of individual play. However, the first experience with group play (in rounds 11-20) clearly makes the individuals better monetary policymakers when they go back to playing the game alone (in rounds 21-30). Within that second batch of 10 rounds of individual play, there is again no evidence of learning. So our conclusion is that there is little evidence of learning, but overwhelming evidence for the superiority of groups over individuals.
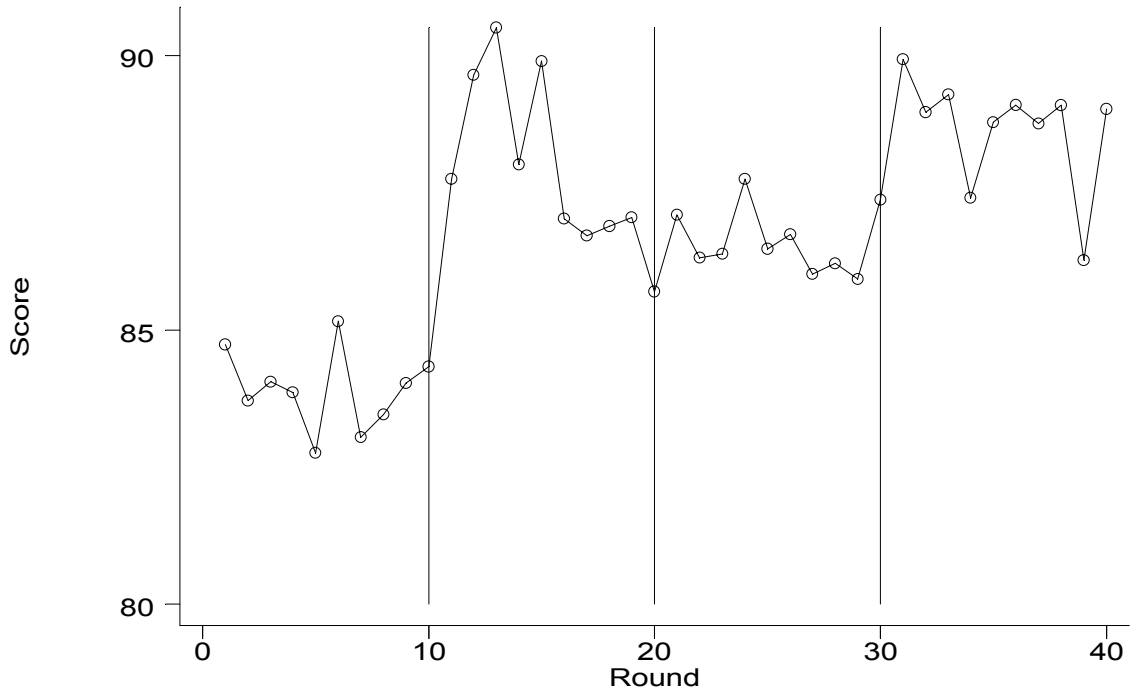
Figure 7: Mean Score by Round in Monetary Experiment

T-tests verify these graphical impressions. Looking first at individual play, the increase in mean score from rounds 1-10 to rounds 21-30 is notable (3.2%) and extremely significant (t=6.1). The standard deviation also drops markedly. All this suggests substantial learning. Learning effects were minor across the two rounds of group play--the mean score in rounds 31-40 was just 0.9% higher than the mean score in rounds 11-20. This improvement is not quite statistically significant (t=1.6, p=.12).

Experimental order

In the urn experiment, we were dismayed to find that the order of group play seemed to matter. In particular, subjects performed significantly better in subsequent group play if their *initial* exposure to group decisionmaking was under unanimity, rather than under majority rule. For individuals, however, the

performance gap went in just the opposite direction--but it was not significant. So we were inclined to write these results off as a fluke.

Results from the monetary policy experiment suggest that was the right decision. Neither the scores from group play in rounds 31-40 nor the scores from individual play in rounds 21-30 appear to be affected by whether the subjects' first participation in group decisionmaking (in rounds 11-20) was under majority rule or a unanimity requirement.

Differences between men and women

The monetary policy experiment imposed a screen for subject selection that the urn experiment did not: Students had to have taken a course in macroeconomics. Not surprisingly, that tilted the gender ratio a bit: to 54% male, 46% female. In consequence, we wound up with more groups with male majorities (11 out of 20).

As in the urn experiment, we found some differences between the ways men and women played the game--but here they were more minor. When playing as individuals, female subjects generally reacted slightly more quickly to changes in aggregate demand (mean lag = 2.31 periods versus 2.56 periods for males; t = 1.7). But the scores they earned were not significantly lower than those earned by the men.

However, groups with a female majority had a longer average decision lag (2.49 periods) than groups with a male majority (2.14 periods), rather than a shorter one. The difference, however, was not significant (t=1.3). Similarly, while

male-dominated groups outscored female-dominated groups by a slight margin (0.7%), the difference was insignificant (t=1.4).

4. Can We Model Group Decisionmaking?

It is possible to formulate and test several simple models of how groups aggregate individual views into group decisions. None of these are strictly "economic" models, however, because every *homo economicus* should make the same decision. (After all, the objective function and the information are identical for all participants.) As will be clear shortly, none of these simple, intuitive models of group decisionmaking gets us very far.

Model 1: The whole is equal to the sum of its parts

The simplest model is that there are no group interactions at all: The group's decision is simply the average of the five individual decisions. This, of course, come closest to the pure economic model (which says that everyone agrees). However, this model has, essentially, already been tested and rejected in Sections 2 and 3. Let X denote any one of our three decision variables (L, S, or C), and let X_G be the average value attained by the group and X_A be the average values attained by the five people in the group *while they played as individuals*. As noted earlier, we consistently reject $X_G = X_A$ in favor of the alternative that groups do better.

Now let us ask a slightly different question: Looking across the 20 groups, does the average performance of the five people who comprise a particular group (X_A) take us very far in explaining--in a regression sense--how well the group

does on that same criterion (X_G)? Since we have three different choices of X (L, S, and C) and data from two different experiments, we can pose six versions of this question. Rather than display the (rather unsuccessful) regression equations, Figure 8 shows the corresponding scatter diagrams. Each is based on 20 observations, one for each session. What message do these six charts convey?
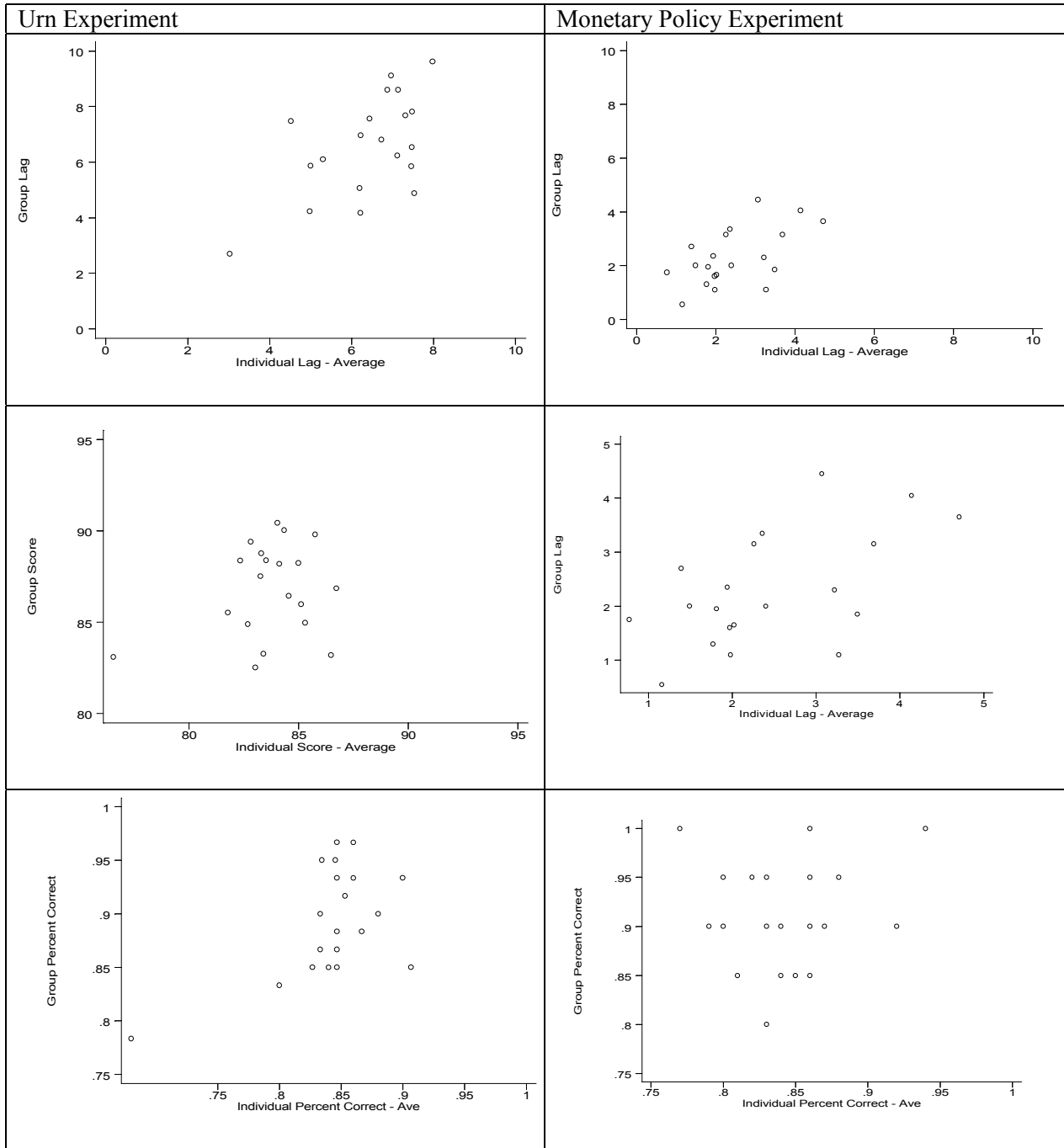
Figure 8: Group Compared to Average Individual Play

In general, they give the impression that a linear model of the form $X_G = a + bX_A + u$ does not fit the data at all well.[27] In one case, the correlation is even negative--which is really quite astounding. Looking across the three variables, L_A does by far the best job of explaining L_G, although even here the simple correlations are just 0.58 in the urn experiment and 0.57 in the monetary policy experiment--corresponding to R^2s of about 0.33. (The regression coefficients are 0.84 and 0.90, respectively.) In the monetary policy experiment, the correlations for the other two variables, S and C, are nearly zero.

In a word, the average performance of the five individuals who comprise each group carries almost no explanatory power for how well the group performed. The Yankees and the Lakers would be surprised--and would be spending too much on payroll--if this were true in professional sports.

Model 2: The median voter theory

A different concept of "average" plays a time-honored role in one of the few instances of group decisionmaking that economists have modeled extensively: voting. Where preferences are single-peaked, as they must be in these applications, a highly-pedigreed tradition in public finance holds that the views of the median voter should prevail. It seems natural, then, to ask whether the performance of the median player can explain the performances of our 5-person groups? Remember, we literally used either a majority vote or a unanimous vote to determine the group's decisions in our experiments.

---

[27] It is apparent from the diagrams that linearity is not the issue. No obvious nonlinear model does much better.

Figure 9, which follows the same format as Figure 8, shows that the median voter model generally (but not always) is a better predictor of group outcomes than simple averaging. In one case, the R^2 gets as high as .54. But, in general, these six scatters once again show that even the median-voter model has only modest success (and, in some cases, no success at all) in explaining the performance of the group. As before, the groups' L decisions are explained best; the R^2's of the two regressions are .054 for the urn data and .42 for the monetary policy data. In two cases (variables S and C in the monetary policy experiment), the correlation is actually negative.
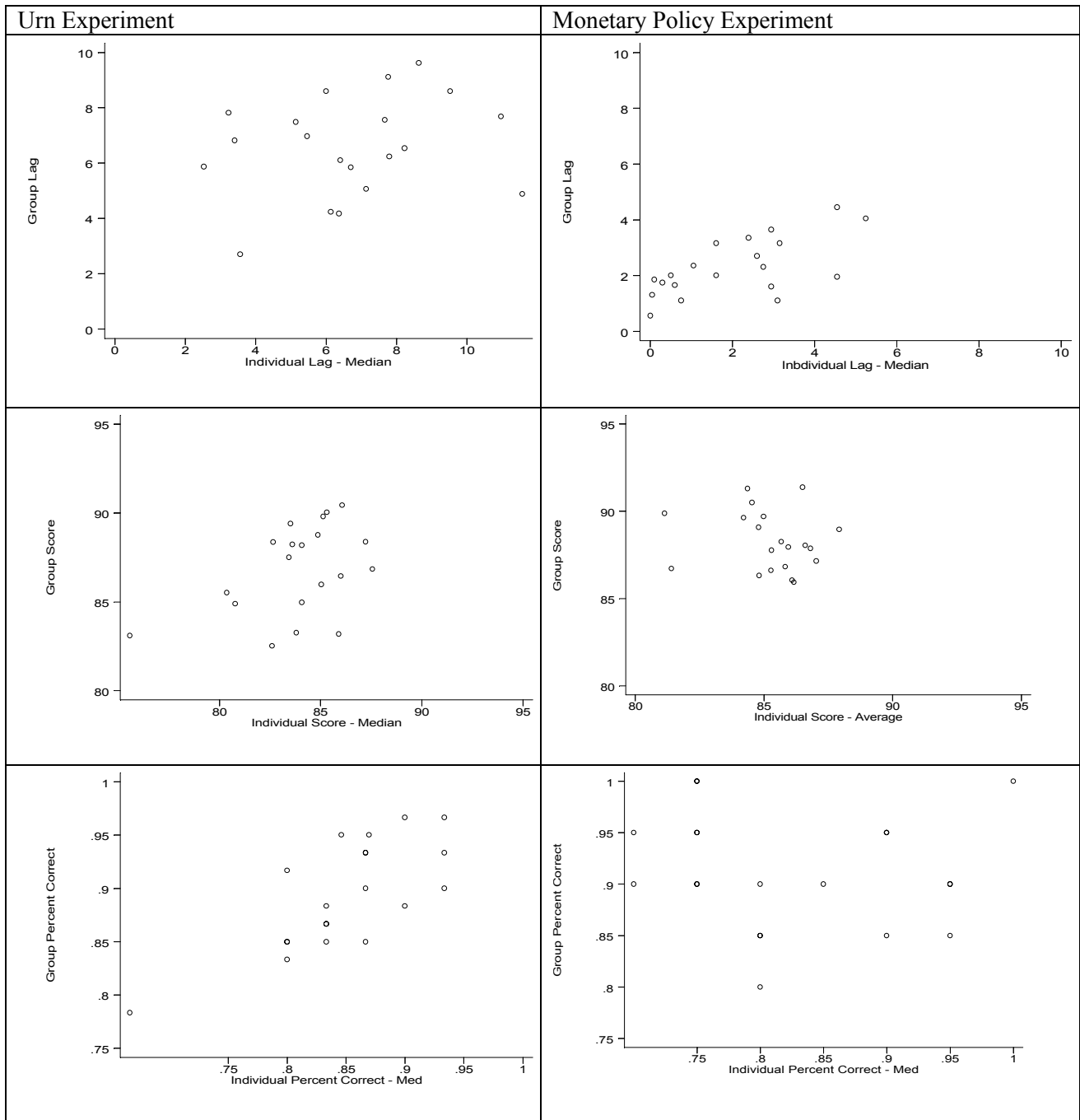
Urn Experiment	Monetary Policy Experiment

Figure 9: Group Compared to Median Individual Play

Model 3: May the best man (or woman) win

In discussing our experiment with other economists, several suggested that the group's decisions would be dominated by the best player in the group--as indicated, presumably, by his or her scores while playing alone. This hypothesis struck us as plausible, even after watching the games being played many times. So we tested models of the form $X_G = a + bX^* + u$, where X^* is the average outcome (on variable S, C, or L) of the individual who achieved the *highest* average score while playing alone.

There is, however, a logically prior question: Are there statistically significant individual fixed effects that can be used to identify "better" and "worse" players? To answer this question, we ran a series of regressions, one for each experimental session, explaining individual scores by five dummy variables, one for each player.[28] Perhaps surprisingly, this preliminary test of the idea that there is a "best player" turned up absolutely no evidence of reliable individual fixed effects in the urn experiment: Only four of the 100 individual dummy variables were significant at the 5% level. In the monetary policy experiment, however, there was some weak evidence that some players are better and others worse: 15 of the 100 individual dummies were significant at the 5% level.

With this in mind, we can now look at Figure 10, which displays the six scatter diagrams. In general, the fits appears to be quite modest. (The highest R^2 among the six scatters is .28.) In only one of the six cases (explaining C_G in the

---

[28] Thus each regression was based on 150 observations in the urn experiment and 100 observations in the monetary policy experiment.

monetary policy experiment), is this the best-fitting model; in three cases, it is the worse. Once again, the variable L is explained best.
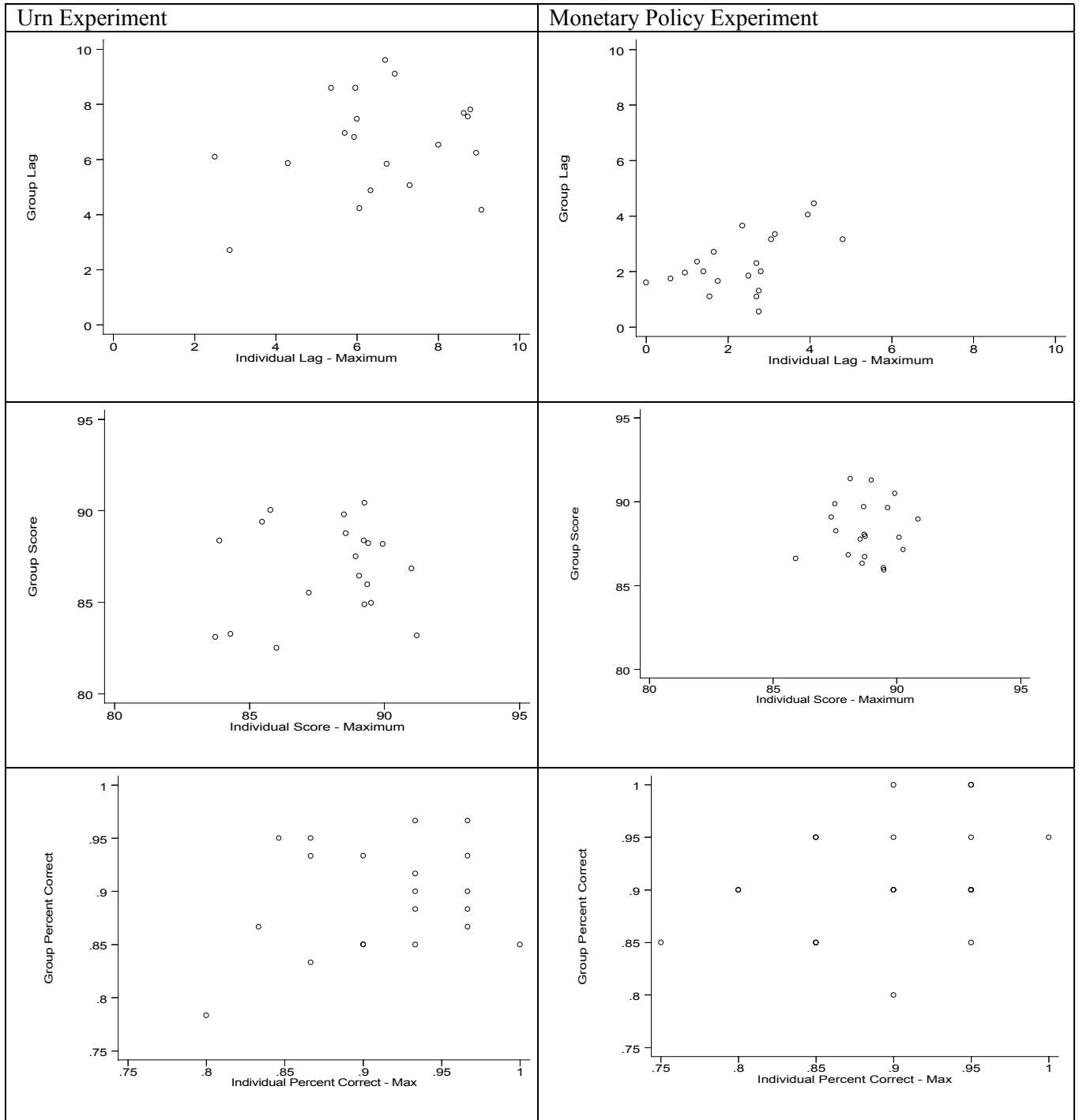

Figure 10: Group Compared to Maximum Individual Play

Finally, we note that various multiple regressions using, say, both X_A and X^* do not appreciably improve the fit. In the end, we are left to conclude that neither the average player, nor the median player, nor the best player determine the decisions of the group. The whole, we repeat, does indeed seem to be something different from--and generally better than--the sum of its parts.

5. Conclusions

Perhaps the best way to illustrate the striking similarly in findings from these two very different experiments is to rack them up, side by side, as we do in Table 3:

Table 3

	Urn experiment	Monetary Policy Experiment
1.	Groups no slower	Groups no slower
2.	Groups better by 3.7%	Groups better by 3.5%
3.	Majority rule approx. the same as unanimity	Majority rule approx. the same as unanimity
4.	Early learning improves scores	Early learning does not improve scores
5.	Women decide faster	Women decide faster
6.	Men outperform women as individuals	Men do not outperform women as individuals
7.	Male majorities do not outperform female majorities	Male majorities do not outperform female majorities

8.	Subsequent group scores higher if unanimity comes first	Subsequent group scores not higher if unanimity comes first
9.	Simple models of group behavior fit poorly	Simple models of group behavior fit poorly
10.	No significant individual effects	Significant individual effects

While there are some minor differences (noted above) between the results of the urn experiment and those of the monetary policy experiment, the correspondence is little short of amazing.

From the start, our interest centered on the first two findings:

* *Do groups reach decisions more slowly than individuals?* According to these experimental results, what seemingly everyone believes (including the authors, prior to this study) is simply not true: Groups appear to be no slower in reaching decisions than individuals are.

* *Do groups make better decisions than individuals?* The experimental answer seems to be yes. And the margin of superiority of group over individual decisions is astonishingly similar in the two experiments--about 3 1/2%.

If groups make better decisions and require no more information to do so, then two heads--or, in this case, five--are indeed better than one. Society is, in that case, wise to assign many important decisions to committees.

REFERENCES

Kenneth J. Arrow, *Social Choice and Individual Values* (New York: Wiley), 1963.

Laurence Ball, "Efficient Rules for Monetary Policy," NBER Working Paper No. 5952, March 1997.

Alan S. Blinder, *Central Banking in Theory and Practice* (Cambridge: MIT Press), 1998.

Glenn D. Rudebusch and Lars E.O. Svensson, "Policy Rules for Inflation Targeting," in J. B. Taylor (ed.), *Monetary Policy Rules* (Chicago: University of Chicago Press), 1999, pp. 203-246.