GAMES AND ECONOMIC BEHAVIOR 6, 347–369 (1994)

Fairness in Simple Bargaining Experiments*

ROBERT FORSYTHE, JOEL L. HOROWITZ, N. E. SAVIN, AND MARTIN SEFTON

Department of Economics, University of Iowa, Iowa City, Iowa 52242

Received July 1, 1991

We present an experiment to test whether fairness alone can explain proposers' willingness to make nontrivial offers in simple bargaining games. We examine two treatments: game (ultimatum or dictator) and pay (pay or no pay). The outcomes of the ultimatum and dictator games with pay are significantly different, implying that fairness, by itself, cannot explain the observed behavior. Doubling the amount of money available in games with pay does not affect these results. The outcomes of both games are replicable when players are paid, but the outcome of the ultimatum game is not replicable when players are not paid. *Journal of Economic Literature* Classification Numbers: 026, 215. © 1994 Academic Press, Inc.

1. INTRODUCTION

Extensive literature has emerged analyzing bargaining processes as finite-horizon, two-person, alternate offer games in which players take turns making offers. In these games, which were first analyzed by Stahl (1972), player 1 makes an offer on how to divide the pie. In response player 2 can either accept and end the game or reject and make a counter-offer. In response to a counter-offer player 1 can either accept or reject, etc. The last round of these games is particularly interesting. It consists of two stages in which player 1 makes a single take-it-or-leave-it offer (stage 1) that player 2 must either accept or reject without making further counter-offers (stage 2). This two-stage game is known as an ultimatum game; it is important because it provides the basis for the analysis of more complicated bargaining games.

The complete information version of an ultimatum game is simple to

* Research supported in part by NSF Grant SES-8922460.

0899-8256/94 $6.00

analyze. Suppose there are \$5 on the table and player 1 can offer to divide them by giving X to player 2 and keeping \$5 − X. In the subgame perfect Nash equilibrium, player 1 offers player 2 nothing and player 2 accepts this offer.[1]

Because of the apparent simplicity of pie-splitting games, game-theoretic predictions for them have been tested in several experiments. The experimental evidence does not strongly support subgame perfection. Güth *et al.* (1982), and Roth *et al.* (1991), carried out experiments with the ultimatum game and obtained the result that most players 1 give away a nontrivial amount of money. Other experimental work has examined more complicated games with two or more rounds of offers and counter-offers. This work (by Binmore *et al.*, 1985; Neelin *et al.*, 1988, Ochs and Roth, 1989, Spiegel *et al.*, 1990 and Bolton, 1991) offers only very limited support for the subgame perfect Nash equilibrium prediction.

In this paper we also examine dictator games, which are even simpler than ultimatum games. Here, player 1 dictates how to divide the pie and player 2 does not have the opportunity to reject this division. As in the ultimatum game, a rational self-interested player should not offer anything to player 2.[2]

Our major objective in this paper is to determine whether nontrivial offers in ultimatum and dictator games can be explained by proposers' concerns with fairness. If nontrivial offers are due solely to proposers' concerns with fairness, the distributions of offers will be the same in the two games. Alternatively, if the distributions differ, other factors must influence offer distributions. The main focus of this paper is on testing the "fairness hypothesis" that the distributions of offers are the same in the ultimatum and dictator games.

As a secondary hypothesis, we also test whether paying subjects makes a difference in the outcomes. In all games, players were asked to divide a \$5 pie. With pay, the players receive their shares of the pie; without pay they receive nothing. The pay hypothesis is that the distributions in games with and without pay are identical. Economic theory suggests that incentives should matter while, on the other hand, Thaler (1986), argues that there is little evidence of this.

To test these hypotheses we conducted two sessions of both games at two different times: April and September, 1988. We test whether the distributions of proposals in a particular game are the same at each date.

[1] Depending on whether the receiver is assumed to accept or reject an offer when he is indifferent towards the amount he receives, an offer of one cent can also be the subgame perfect equilibrum in an ultimatum game.

[2] We take license of calling the dictator game a "game" although it is a single person decision problem.

We find that we can reject the hypothesis that the distributions of proposals are the same at different times only in the ultimatum games without pay.

We also find that

(1) we can reject the fairness hypothesis in games with pay, implying that a proposer's taste for fairness, by itself, does not explain the distributions of proposals in the ultimatum game.

(2) the results on the fairness hypothesis are less clear in the games without pay due to lack of replicability in the ultimatum games. If the September ultimatum game is used, the fairness hypothesis is not rejected, but the test results are ambiguous if the April ultimatum game is used.

(3) we reject the pay hypothesis for the dictator games.

(4) we find some evidence in favor of the pay hypothesis in the ultimatum game but again, the lack of replicability in the ultimatum games without pay makes these results inconclusive.

An additional question related to the pay hypothesis is whether the size of the pie matters. To investigate this, ultimatum and dictator games with pay were conducted using pies of $10. The hypothesis that the distributions of the proportions of the pie offered in the $5 and $10 games are identical was tested and accepted. This suggests that the size of the pie does not matter in the range considered.

The remainder of this paper is organized follows. We describe our experimental design in the next section and the design of our analysis in Section 3. The results of our experiment are given in Section 4 and an interpretation of the results and concluding remarks are presented in Section 5.

2. Design of the Experiments

In this paper, we investigate the effects of two treatments: game and pay. The game is either the ultimatum game or the dictator game. The pay treatment is either pay or no pay. Thus, with an exception discussed below, our analysis is based on a 2×2 experimental design.

Each experimental session consisted of a one-round game played by students who were randomly chosen from a subject pool recruited from undergraduate accounting and economics classes and MBA economics classes at the University of Iowa. Individuals had joined the subject pool voluntarily by completing a form indicating their interest in participating in experiments. On this form they were told that "earnings will vary, but previous subjects in experiments lasting two hours have earned between $15 and $30." When called and asked to participate, students were told that they "would earn between $3 and $8 for participating in an experiment that would last about 20 minutes." Each participant played a single game

against an anonymous opponent. To minimize any possible repeated game influences on the outcomes of the experiments, the participants were given no opportunity to meet or see their opponents before, during or after the experiment.

We conducted two sets of experiments. The first set was conducted in April 1988. The second was conducted in September 1988 to test the replicability of the results obtained in April. The procedures used in the second set of experiments were identical to those used in the first set. We planned to have 24 observations per cell in each set, but we did not achieve this goal because of absenteeism among subjects. We obtained 87 observations out of a planned 96 in April and 95 observations out of a planned 96 in September.

We used two connecting rooms, each of which could accommodate eight individuals. Subjects were assigned to rooms randomly when they were recruited. At the beginning of a session, subjects were given instruction sets (reproduced in Appendix A). The experimenter read these instructions aloud and answered questions. All subjects in the same session faced the same experimental treatment, and all subjects in the same room faced the same task (that is, they were either all senders or all receivers of a proposal).

Each person was randomly paired with someone in the other room, and each pair was given a $5 pie to divide. Communication between members of a pair was by written proposal forms that were carried between rooms by the experimenter. The rules of the game and the actual payoffs varied across four cells. Two different sets of rules described the ultimatum and dictator games. For each set of rules, payoffs were determined under the two pay treatments, pay and no pay.

The ultimatum games were take-it-or-leave-it games. Each subject in Room A proposed a division of the $5, and the subject he was paired with in Room B could either accept or reject the proposal. Acceptance meant that the proposal was enforced, and rejection meant that neither subject received anything. The dictator games were "take-it" games. Each subject in Room A chose how to allocate the $5; the subjects in Room B could benefit from the allocation but had no accept or reject decision to make.

Each subject received $3 for participating in the experiment. Subjects participating in experiments with pay received additional amounts according to the rules of the game they played and the decisions that they made. Each session was completed in less than 20 minutes.

In a further set of experiments, we increased the size of the pie with pay to $10 to obtain evidence on the effects of changing the amount of money available. These experiments were carried out in November 1988

using procedures identical to those of our previous experiments except for the size of the pie. There were 24 participants in the $10 ultimatum game and 24 in the $10 dictator game. (The data from all experimental sessions may be found in Appendix B.)

3. Design of the Analysis

In the subgame perfect equilibrium, all proposals made in the ultimatum game have the same value ($0.01), and all proposals made in the dictator game have the same value (zero). However, in all reported experiments with these games, including ours, the proposals made are distributed over a range of values. Therefore, assuming that the proposals in a particular game are a random sample from some underlying probability distribution, a complete description of the outcomes of the ultimatum and dictator games must give the distributions of the values of the proposals. Moreover, the statement that the outcomes of games with different groups of players, rules or pay schedules are the same (different) means that the probability distributions of proposals are the same (different).

In this research, we are interested in testing whether the outcomes of the ultimatum and dictator games are replicable, whether the outcomes of the ultimatum and dictator games are replicable, whether the outcomes of the dictator and ultimatum games are the same (the fairness hypothesis), and whether the pay treatment influences the outcomes of the games (the pay hypothesis). We do this by testing the hypotheses that the probability distributions of proposals in different games (or in repetitions of the same game with different players) are identical. Specifically, we test the hypotheses (1) that the distributions of proposals in the April and September ultimatum games are identical and similarly for the dictator games, (2) that the distributions of proposals in ultimatum and dictator games with the same pay treatment are identical, and (3) that the distributions of proposals in each game are identical under different pay treatments. When the pie sizes are different, the proposals are expressed as fractions of the pie, rather than in dollars, when testing hypothesis (3).

The hypotheses are all stated in terms of testing the invariance of the distribution of proposals rather than particular characteristics of the distribution such as the mean and variance. This is done because conventional theory predicts that proposals will be concentrated at a single point as was discussed in the introduction to this paper. Since theory does not predict a distribution of proposals, it provides no guidance about which functionals of the distribution should be tested. Invariance of the entire distribution has the appealing property of implying that all functionals are invariant.

A. *Test Statistics*

To formalize the hypothesis tests, assume that the proposals obtained in two experiments, 1 and 2, are random samples from probability distributions with cumulative distribution functions $F_1(\cdot)$ and $F_2(\cdot)$, respectively. We test the null hypothesis H_0: $F_1 = F_2$ against the alternative H_1: $F_1 \neq F_2$.

One way of testing H_0 is to compute the difference between consistent estimates of F_1 and F_2, using a suitable metric, and reject H_0 if the difference is too large. The Cramer–von Mises (CM), Anderson–Darling (AD), Kolmogorov–Smirnov (KS), and Wilcoxon rank–sum (RS) tests are well-known examples of tests based on this idea. To formulate the corresponding test statistics, let $\hat{F}_1$ and $\hat{F}_2$ denote the empirical distributions of the proposals in experiments 1 and 2, respectively, and let $\hat{F}_{12}$ be the empirical distribution of the pooled proposals. Let N_1 and N_2 be the sizes of the samples obtained from experiments 1 and 2. The CM test statistic is

$$\mathrm{CM} = N_1 N_2 (N_1 + N_2)^{-1} \int_{-\infty}^{\infty} [\hat{F}_1(x) - \hat{F}_2(x)]^2 d\hat{F}_{12}(x).$$

The AD statistic is

$$\mathrm{AD} = N_1 N_2 (N_1 + N_2)^{-1} \int_{-\infty}^{\infty} w(x)[\hat{F}_1(x) - \hat{F}_2(x)]^2 d\hat{F}_{12}(x),$$

where

$$w(x) = \{\hat{F}_{12}(x)[1 - \hat{F}_{12}(x)]\}^{-1}.$$

The KS statistic is

$$\mathrm{KS} = \operatorname*{Max}_x |\hat{F}_1(x) - \hat{F}_2(x)|.$$

The RS statistic is

$$\mathrm{RS} = \int_{-\infty}^{\infty} \hat{F}_1(x)\, d\hat{F}_2(x).$$

The CM, AD, and KS tests reject H_0 if their respective test statistics are too large. The RS test rejects H_0 if $|\mathrm{RS} - \frac{1}{2}|$ is too large. Tables of asymptotic critical values of the CM and AD tests are given in Shorack

and Wellner (1986).[3] Exact critical values of the KS test are given by Kim and Jennrich (1973), and exact critical values of the RS test are given by Wilcoxon *et al.* (1973).

Another test of H_0 can be developed by observing that H_0 is true if and only if the characteristic functions of F_1 and F_2 are equal. Epps and Singleton (1986), have proposed a test in which H_0 is rejected if the difference between the empirical characteristic functions corresponding to samples on F_1 and F_2 is too large. The test statistic is

$$\mathrm{CF} = (N_1 + N_2)(\hat{\phi}_1 - \hat{\phi}_2)'\hat{\Omega}^{-1}(\hat{\phi}_1 - \hat{\phi}_2),$$

where for $i = 1$ or 2, $\hat{\phi}_i \equiv [\mathrm{Re}\ \hat{\phi}_i(0.4), \mathrm{Im}\ \hat{\phi}_i(0.4, \mathrm{Re}\ \hat{\phi}_i(0.8), \mathrm{Im}\ \hat{\phi}_i(0.8)]'$, $\hat{\phi}_i(\cdot)$ is the empirical characteristic function of the sample from F_i, and $\hat{\Omega}$ is a consistent estimator of the covariance matrix of $(N_1 + N_2)^{1/2}(\hat{\phi}_1 - \hat{\phi}_2)$. Under H_0, CF is asymptotically distributed as chi-square with 4 degrees of freedom. Epps and Singleton (1986), give a small-sample correction that improves the quality of the asymptotic approximation when N_1 and N_2 are small.

The CM, AD, KS, and RS tests assume that F_1 and F_2 are continuous. To make the tests applicable to the ultimatum and dictator games, where tied proposals occur frequently, we have added to each proposal a number sampled randomly from the uniform distribution on (0, 0.001). This procedure insures that the proposals follow continuous distributions but does not significantly distort the outcomes of the games or affect the truth or falsity of H_0. Although the CF test does not require F_1 and F_2 to be continuous, we have found that its small-sample performance can be poor when tied proposals are frequent. Therefore, we have also used the foregoing tie-breaking technique in the CF test.

B. Power of the Tests

To be useful, a test must have sufficient power to discriminate among interesting alternative hypotheses. In Appendix C we report the results of an extensive Monte Carlo investigation of the powers of the 5 tests just described. The distributions F_1 and F_2 used in the investigation span a range of alternatives that was suggested by the results of previous experiments. The distributions consist of mixtures of the gamesman pro-

[3] Shorack and Wellner (1986), give critical values for the AD statistic for significance levels of 0.15, 0.10, 0.05, 0.025, and 001. We computed critical values of this statistic for intermediate significance levels by carrying out a Monte Carlo simulation with $N_1 = N_2 = 24$ and 10,000 replications. The intermediate significance levels reported in Section 5 are based on this simulation.

posal to offer nothing and proposals that are distributed around an equal division of the pie. We find that with $N_1 = N_2 = 25$ (approximately the same sizes used in the experiments), the CF and AD tests are noticeably more powerful than the others. With each pair of alternatives, at least one of the tests AD and CF has power that equals or exceeds the power of all the other tests. At the 0.05 level, the CF test has power exceeding 0.76 for 27 of the 28 pairs of alternatives we considered, the exception being one in which F_1 and F_2 are particularly close (alternatives 6 and 7 in Appendix C). The AD test has power exceeding 0.84 for all but 4 of the 28 pairs of alternatives. We conclude that the CF and AD tests have good power for discriminating among economically interesting alternatives with the sample sizes used in the experiments. Since these tests are more powerful than the others, we report only the results of using CF and AD to carry out formal tests of hypotheses about the outcomes of the ultimatum and dictator games.

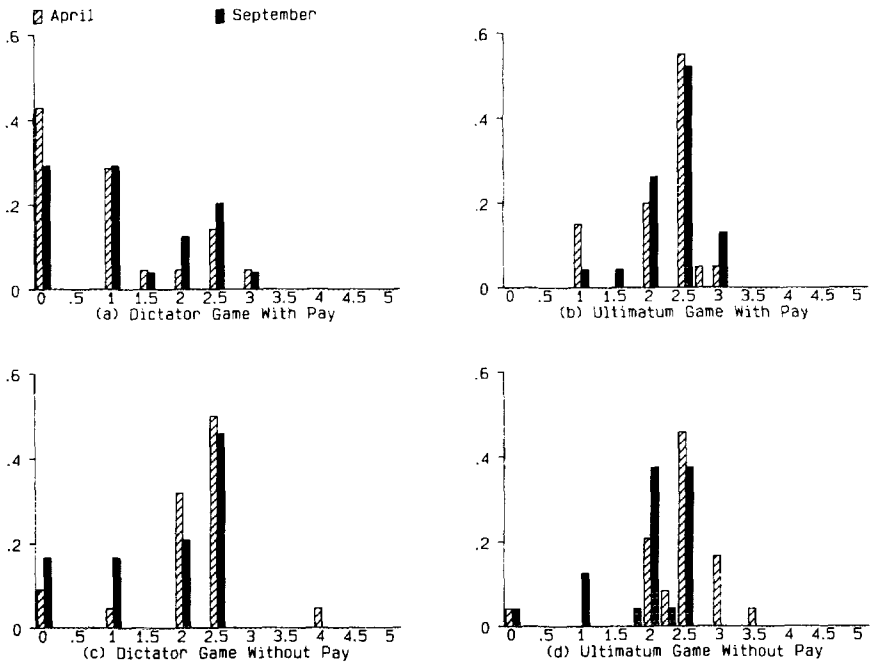

FIG. 1. Histograms of April and September proposals. Each histogram measures the amount of the proposal in dollars on the horizontal axis and the fraction of proposals of this amount on the vertical axis.

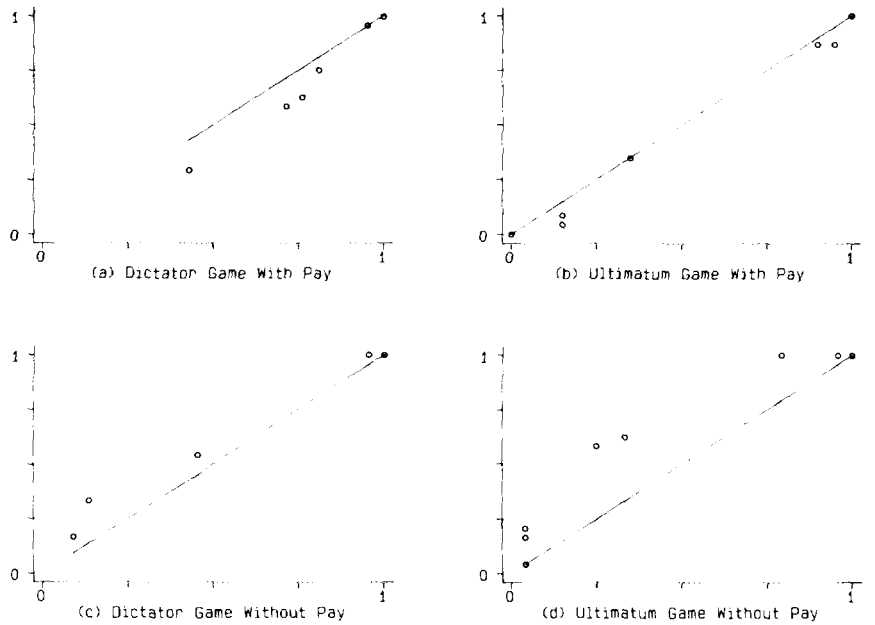

FIG. 2. Quantile–quantile plots of April and September proposals.

4. RESULTS OF THE EXPERIMENTS

The results of the experiments are shown in Figs. 1 and 2. Figures 1a–1d present histograms of the offers under each set of treatments. The four panels show the distributions of offers in each game in April and September and are organized in April–September pairs according to treatment. Figures 2a–2d show quantile–quantile plots comparing the April and September results of each treatment. These plots provide another way to compare the distributions of outcomes visually. In a quantile–quantile plot, the horizontal axis gives the proportion of observations less than or equal to a given value in one sample, and the vertical axis gives the proportion of observations less than or equal to the same value in the other sample. For example, if the point (0.4, 0.6) appears in a quantile–quantile plot, 60% of the observations in one sample are less than or equal to a particular value, and 40% of the observations in the other sample are less than or equal to the same value. If the two samples are drawn from the same distribution, the points in a quantile–quantile plot scatter around a 45° line. Systematic departures from the 45° line indicate that the two samples are drawn from different distributions.

TABLE I

TESTS OF REPLICABILITY[a]

	Sample size		Test statistic (p-value)	
	April	September	AD	CF
1. Dictator: Pay	21	24	0.58 (0.66)	1.14 (0.89)
2. Ultimatum: Pay	20	23	0.74 (0.52)	2.10 (0.72)
3. Dictator: No pay	22	24	0.83 (0.45)	3.17 (0.53)
4. Ultimatum: No pay	24	24	5.09 (0.00)	9.46 (0.05)

[a] The null hypothesis in each row is that the distributions of proposals in April and September are the same. The p-value is the probability that the test statistic exceeds the given value when the null hypothesis is true.

We first test to see whether the results from the April sessions are replicated in the September sessions. The tests of replicability are carried out for each game using a $5 pie with and without pay. The results of the formal tests are shown in Table I. The April–September replicability of the distribution of outcomes of each game was tested using the CF and AD tests. Using conventional significance levels, neither of the formal tests rejects replicability of the two dictator games and the ultimatum game with pay. Both tests reject replicability of the ultimatum game with no pay at the 0.05 level, and the AD test rejects at the 0.01 level. Thus, it appears that the ultimatum game with no pay is not replicable. Further evidence against the replicability of this game is presented in Sections 4A and 4B.

Replicability of the dictator games and ultimatum game with pay implies that the results of the April and September experiments with each game can be pooled, and we have done this for purposes of comparing the outcomes of different games and (in the case of the dictator game) pay treatments. We also report the results of comparing the games using the April and September results separately. With two exceptions, the two sets of comparisons lead to the same conclusions. One important and unsurprising exception occurs in comparisons involving the ultimatum game with no pay, whose results are not replicable. The other exception concerns the pay hypothesis and is discussed in Section 4B.

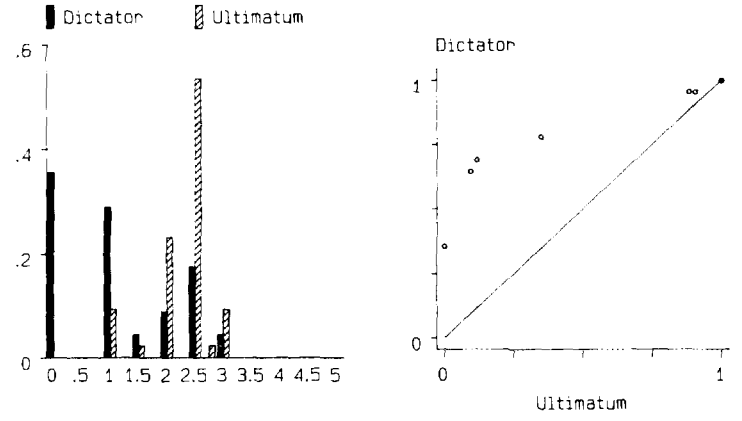

FIG. 3. Dictator with pay (pooled) vs ultimatum with pay (pooled).

A. A Comparison of the Ultimatum and Dictator Games: The Fairness Hypothesis

As we discussed in Section 1, the fairness hypothesis states that the distribution of proposals in the ultimatum and dictator games are identical. The results of formal tests of this hypothesis are presented here. Row 1 of Table II shows the results of formal tests of the hypothesis that the distributions in the ultimatum and dictator games with pay are identical. Both tests reject this hypothesis at the 0.01 level. The histograms and quantile–quantile plot in Fig. 3 reveal the cause of this result. Players are more generous in the ultimatum game than in the dictator game. In the dictator game 36 percent of the players are gamesmen (that is, they offer nothing), and the histogram of proposals is bimodal, with peaks at the outcome predicted by conventional theory (an offer of zero) and at the equal-shares outcome. In the ultimatum game, no sender tries to keep the entire pie; the outcomes are distributed around the equal-shares proposal.

Rows 2 and 3 of Table II give the results of formal tests of the fairness hypothesis based on comparing the ultimatum and the dictator games with pay separately for April and September. The hypotheses tested are that the two games have identical distributions in each experimental period. As with the pooled data, the tests show clearly that the distributions in the ultimatum and dictator games with pay differ from each other in both April and September.

Expectations for the games with no pay are ambiguous. There are no strong reasons for believing that the outcomes of the ultimatum and dicta-

TABLE II

TESTS OF THE FAIRNESS HYPOTHESIS[a]

	Sample size		Test statistic (p-value)	
	Dictator	Ultimatum	AD	CF
	Games with pay			
1. Pooled	45	43	13.00	43.58
			(0.00)	(0.00)
2. April	21	20	8.24	22.32
			(0.00)	(0.00)
3. September	24	23	6.54	20.05
			(0.00)	(0.00)
	Games with no pay			
4. Pooled dictator vs April ultimatum	46	24	2.95	11.48
			(0.03)	(0.02)
5. Pooled dictator vs September ultimatum	46	24	0.52	3.53
			(0.72)	(0.47)
6. April	22	24	1.51	6.28
			(0.18)	(0.18)
7. September	24	24	1.30	3.63
			(0.23)	(0.46)

[a] The null hypothesis in each row is that the distributions of proposals in the ultimatum and dictator games are the same. The p-value is the probability that the test statistic exceeds the given value when the null hypothesis is true.

tor games with no pay should be different, but the lack of financial incentives makes any prediction of outcomes problematic. Rows 4 and 5 of Table II show the results of testing the pooled April–September outcomes of the dictator game with no pay against the separate April–September outcomes of the dictator game with no pay. The outcomes of the April and September outcomes of the ultimatum game are not pooled since it is unlikely that they are sampled from the same distribution. The hypothesis that the distributions in the pooled dictator and September ultimatum games are the same cannot be rejected. The hypothesis that the distributions in the pooled dictator and April ultimatum games are the same is rejected. Examination of the data shows that this is because there are substantially fewer outcomes below $2.00 in the April ultimatum game than in the other games with no pay. Since the tests of the fairness hypothesis for the games with no pay yield different results depending on whether the April or September ultimatum game is used, we are unable to reach a firm conclusion as to whether the fairness hypothesis holds in the games with no pay.

This ambiguity is not resolved when the outcomes of the April and September games are compared separately. Rows 6 and 7 of Table II show the results of tests in which the outcomes of the April and September games with no pay are compared separately. The tests accept the hypothesis that the April distributions are identical and that the September distributions are identical. In contrast, recall that the tests reject equality of outcomes of the pooled April–September dictator games and the April ultimatum game. Thus, the tests of the outcomes of the no-pay ultimatum and dictator games against each other yield different results, depending on whether the April and September dictator results are pooled. This finding is not necessarily surprising. If the April and September dictator outcomes are sampled from the same distribution, as the evidence presented above suggests, whereas the April and September ultimatum outcomes are sampled from different distributions, the difference will be easier to detect using the larger sample obtained from the pooled dictator games than using only the April dictator results.

B. The Effect of Pay: The Pay Hypothesis

The pay hypothesis (that the distributions of proposals are the same with and without pay) is rejected for the $5 dictator game on the basis of the results of the pooled April and September experiments. The formal test results are reported in row 1 of Table III. The pay hypothesis also is rejected on the basis of the April experiments alone (see row 2 of Table III) but not on the basis of the September experiments alone (row 3 of Table III). Since our other evidence suggests that the outcomes of the dictator games are replicable, we suspect that failure to reject equality of the distributions in the September games is a reflection of the power of the tests.

We now turn to the ultimatum game. The pay hypothesis is accepted when the distribution in the pooled April–September experiments with pay is compared with the distribution in either the April or the September experiments with no pay. The pay hypothesis also is accepted when the distributions in the April experiments with and without pay are compared and when the distributions of the September experiments with and without pay are compared. The formal test results are given in rows 4–7 of Table III.

These results, by themselves, suggest that the pay hypothesis holds in the ultimatum game. However, concluding this would imply that the distribution of proposals in the games with pay is the same as those in the April ultimatum game without pay *and* the September ultimatum game without pay. In fact, the distributions in the April and September ultimatum games without pay are not identical and, therefore, cannot both equal the distribution in the pooled ultimatum games with pay. Such

TABLE III

TESTS OF THE PAY HYPOTHESIS[a]

	Sample size		Test statistic (p-value)	
	Pay	No pay	AD	CF
	Dictator games			
1. Pooled	45	46	6.72	20.28
			(0.00)	(0.00)
2. April	21	22	6.13	23.47
			(0.00)	(0.00)
3. September	24	24	1.46	5.26
			(0.19)	(0.26)
	Ultimatum games			
4. Pooled with pay vs April no pay	43	24	0.53	6.49
			(0.71)	(0.17)
5. Pooled with pay vs September no pay	43	24	1.90	7.27
			(0.11)	(0.12)
6. April	20	24	0.57	5.86
			(0.67)	(0.21)
7. September	23	24	2.09	6.11
			(0.08)	(0.19)

[a] The null hypothesis in each row is that the distributions of proposals are the same with and without pay. The p-value is the probability that the test statistic exceeds the given value when the null hypothesis is true.

inconsistency among test results occurs frequently when multiple hypotheses are tested. One possible solution to this problem is to carry out a joint test of replication and pay hypotheses for the ultimatum game. A critical region for the test can be obtained with the Bonferroni bounding procedure. In this procedure a test based on k separate test statistics rejects at the α level if any of the k statistics is significant at the α/k level (Savin, 1984). When applied to the results of our ultimatum games, however, this procedure yields different results depending on the number of hypotheses tested and the test statistics used. For example, a test based on the AD statistic rejects the joint replication and pay hypotheses, whereas a test based on the CF statistic does not. Of course, conflicting results are familiar consequences of formulating hypotheses and choosing tests after having examined the data. Accordingly, our results concerning the pay hypothesis for the ultimatum game remain inconclusive.

The last question we investigate is whether increasing the size of the pie beyond $5 affects the distributions of the proposals in the games with

TABLE IV

TESTS OF PIE-SIZE EFFECTS IN GAMES WITH PAY[a]

	Sample size		Test statistic (p-value)	
	\$5 pie	\$10 pie	AD	CF
1. Pooled \$4 dictator vs \$10 dictator	45	24	0.46 (0.76)	2.82 (0.59)
2. April \$5 dictator vs \$10 dictator	21	24	0.86 (0.43)	3.26 (0.52)
3. September \$5 dictator vs \$10 dictator	24	24	0.31 (0.92)	1.45 (0.84)
4. Pooled \$5 ultimatum vs \$10 ultimatum	43	24	0.65 (0.60)	1.94 (0.75)
5. April \$5 ultimatum vs \$10 ultimatum	20	24	0.65 (0.60)	1.72 (0.79)
6. September \$5 ultimatum vs \$10 ultimatum	23	24	0.89 (0.41)	1.84 (0.77)
7. \$10 dictator vs \$10 ultimatum		24[b]	10.02 (0.00)	31.07 (0.00)

[a] The \$10 games were conducted in November and the proposals halved before comparing them with the proposals from the \$5 games. The null hypothesis in rows 1–6 is that the distributions of proportions of the pie offered are the same with pies of \$5 and \$10. The null hypothesis in row 7 is that the distributions of proposals in the ultimatum and dictator games with \$10 pies are the same. The p-value is the probability that the test statistic exceeds the given value when the null hypothesis is true.

[b] The sample size is 24 in both games.

pay. We tested the hypothesis that the distributions of the proportions of the pie offered are equal in the \$5 and \$10 versions of the same game with pay. The hypothesis is accepted at conventional significance levels for both the ultimatum and dictator games. The test results are given in rows 1–6 of Table IV.

The tests also reject at the 0.01 level the hypothesis that the distributions of the \$10 dictator and \$10 ultimatum games are identical (the fairness hypothesis). See row 7 of Table IV. This result is consistent with the findings in Section 4A for the \$5 dictator and ultimatum games.

5. DISCUSSION

We have found in all of our experiments that most players give away nontrivial proportions of the money available to them, contrary to the

subgame perfect Nash equilibrium prediction. This finding is consistent with the results of Güth *et al.* (1982), and Spiegel *et al.* (1990). The results of our tests of the fairness hypothesis show that this behavior cannot be fully explained by a taste for fairness among proposers. If players give away money only because of a desire to be fair, the distributions of proposals in dictator and ultimatum games with equal pies would be identical. This clearly did not happen in our games with pay, where the fairness hypothesis is strongly rejected. In the pooled $5 dictator games with pay 36% of the players are pure gamesmen, and only 22% give their opponents an equal share or better. But in the pooled $5 ultimatum games with pay there are no pure gamesmen, and 65% offer at least an equal share of the pie. Similarly, in the $10 dictator game 21% of the players are pure gamesmen and 21% give away an equal share (none give more than an equal square), whereas in the $10 ultimatum game there are no pure gamesmen and 75% offer at least an equal share.

The results of testing the fairness hypothesis in the games with no pay are ambiguous due, at least in part, to the nonreplicability of the outcome of the no-pay ultimatum game. This lack of replicability makes drawing conclusions about fairness in the absence of pay risky. Nonetheless, the fact that the fairness hypothesis is rejected in a comparison of the pooled dictator and April ultimatum games with no pay provides further evidence that the distributions of proposals in the ultimatum games cannot be fully explained by a taste for fairness among proposers.

If a taste for fairness, by itself, cannot explain the outcomes of ultimatum games, how can they be explained? One possibility is to treat the ultimatum game as one in which there are different types of players, rather than one of complete information. In this incomplete information ultimatum game, some proposers are pure gamesmen, and others are concerned (to varying degrees) with fairness.

Similarly, some receivers are pure gamesmen, whereas others have "spite" components in their utility functions and reject proposals that offer them too little.[4] In this situation, a proposer who is a pure gamesman may find it optimal to offer his opponent a non-trivial share of the pie. Some evidence consistent with this view is reported in Roth *et al.* (1991), which reports a set of repeated ultimatum games, where subjects participated in a sequence of ten games against different anonymous opponents. Given the receivers' observed acceptance rates for different proposals, the authors report that by the tenth game, proposers are choosing to make offers which maximize their expected earnings. In this sense, proposers are acting in a manner consistent with expected income maximization.

[4] See Binmore *et al.*, (1984) for more on this point.

However, the same cannot be said of receivers who continue to reject positive offers.

Spitefulness on the part of some receivers can explain why some reject positive offers that are less than an equal share of the pie. In our pooled $5 ultimatum games with pay, 3 out of 15 such offers were rejected, and in the $10 ultimatum game 1 out of 6 was rejected. No offers equal to or greater than an equal share were rejected. Similar results have been obtained by Güth *et al.* (1982).

Descriptive theories of the behavior of players in ultimatum and dictator games remain to be developed and tested. It is usually assumed that the dictator game measures players' taste for fairness, but other interpretations are possible. For example, players may believe there is a risk that anonymity will not be preserved or that failure to act fairly will have adverse consequences for them. We have suggested that the behavior of players in the ultimatum game might be explained by incomplete information, but this idea remains to be developed fully and tested empirically. Of course, a finding that players in the dictator game are motivated by considerations other than fairness would reduce the attractiveness of our suggestion for the ultimatum game, so the explanation of behavior in the ultimatum game depends on having a satisfactory explanation of the dictator game.

Appendix A: Instructions to the Players

In the instructions presented below, differences in language corrsponding to different treatments are indicated in brackets. The treatment is shown in boldface. The instructions are for games with $5 pies. The obvious changes were made in the instructions for the experiments with $10 pies.

Instructions

You have been asked to participate in an economics experiment. For your participation today we will pay you $3 in cash at the end of the experiment.

[**Pay:** You may earn an additional amount of money, which will also be paid to you in cash at the end of the experiment.]

In this experiment each of you will be paired with a different person who is in another room. You will not be told who these people are either during or after the experiment, and they will not be told who you are either during or after the experiment.

You will notice that there are other people in the same room with you who are also participating in the experiment. You will not be paired with any of these people. The decisions that they make will have absolutely no effect on you nor will any of your decisions affect them.

The experiment is conducted as follows: A sum of $5 has been provisionally allocated to each pair and the person in Room A can propose how much of this each person is to receive. To do this, the person in Room A must fill out a form titled "Proposal Form".

[**Ultimatum:** You will find a copy of this form on the desk in front of you. The first line of

this form tells you your identification number (if you are in Room A) or the identification number of the person you are paired with (if you are in Room B). The identification number of the person you are paired with (if you are in Room A) or your identification number (if you are in Room B) is on line {2}. The amount to be divided is on line {3}.]

[**Dictator:** If you are in Room A, you will find two copies of this form on the desk in front of you. The first line of this form tells you your identification number. The identification number of the person you are paired with is on line {2}. The amount to be divded is on line {3}.]

The person in room A makes the proposal. The proposal consists of an amount the person in Room B is to receive (entered on line {4}) and the amount the person in room A is to receive (entered on line {5}). The amount the person in Room A is to receive is simply the total amount to be divided, $5, minus the amount the person in Room B is to receive.

If you are in Room A you will have five minutes to come to a decision about your proposal. At the end of five minutes, a buzzer will sound. Do not talk to the other people in your room until your session is completed. Do not be concerned if other people make their decisions before you, we will not collect the forms until the buzzer sounds.

[**Dictator:** You should record the same amounts on the other copy of your proposal form. One copy is for your records and the other copy will be sent to the person in Room B with whom you are paired.]

[**Ultimatum:** Your proposal form will then be sent to the person in Room B with whom you are paired.]

[**Ultimatum:** The person in Room B will then be given a chance to accept or reject the proposal. If the person in Room B accepts the proposal, then the amount of money will be divided as specified in the proposal. If the person in Room B rejects the proposal, then both people in the pair receive zero. If the person in Room B wishes to accept the proposal he or she should check "Accept" on line {6} of the proposal form. If the person in Room B does not wish to accept the proposal he or she should check "Reject" on line {6} of the proposal form.]

[**Ultimatum:** If you are in Room B you will have five minutes to come to a decision about whether to accept or reject. At the end of five minutes, a buzzer will sound. Do not talk to the other people in your room until your sessions is completed. Do not be concerned if other people complete their proposal forms before you, we will not collect them until the buzzer sounds. You should record the same amounts on the other copy of the proposal form. One copy is for your records and the other copy will be sent to the person in Room A with whom you are paired.]

[**Pay:** After the proposal forms have been returned to Room A each person will be paid. Each person will receive $3 for participating. Each person in Room A will also receive the amount shown on line {5} of his or her proposal form if the proposal was accepted. Each person in Room B will also receive the amount shown on line {4} of the form if the proposal was accepted.]

[**No Pay:** After the proposal forms have been returned to Room A each person will be paid. The experimenter will pay you the $3 for your participation in this experiment.]

Are there any questions?

PROPOSAL FORM

{1} Identification Number ___A
{2} Paired With ___B
{3} Amount to divide _____
{4} Person in Room B receives _____
{5} Person in Room A receives {3} − {4} _____

[**Ultimatum:** {6} Accept ___ Reject ___]

APPENDIX B: DATA

The tables below list the dollar values of the proposals made in each experiment. An asterisk (*) next to a proposal in an ultimatum game indicates that the proposal was rejected.

EXPERIMENTS WITH $5 PIES

April				September			
Dictator		Ultimatum		Dictator		Ultimatum	
Pay	No pay	Pay	No pay	Pay	No pay	Pay	No pay
0	2	2	2.5	0	2	2.5	2.5
1	2	2.5	3	2	2.5	2	2
1	1	2.5	2.5	0	1	1.5	2
2.5	0	1	2.5	1	2	3	0*
0	2.5	2.5	0*	0	2.5	2	2.5
2.5	2.5	1*	2*	0	2.5	2	2.5
0	2	3	3	3	0	2.5	1*
0	2.5	2.5	3	0	2	1	2.5
0	2.5	2.5	2.5	1	2.5	2.5	2*
1	2.5	2*	3.5	2.5	0	2	2.5
3	2	2.5	2.5	2.5	0	2.5	2.5
1	2.5	2.5	2.25	1	0	2.5	2.5
2	2.5	2.5	2	2.5	2.5	2.5	2
1.5	2.5	2.5	2.5	2.5	1	3	2*
2.5	2	2.5	2.5	0	2.5	2.5	2.25
0	2.5	2	2.5	0	1	2	2
0	2	2.75	2.25	2	1	2.5	1.75*
0	2.5	2	2.5	2	2.5	2	2
1	4	1*	2	1	2.5	2.5	2
0	2.5	2.5	2	1	2.5	2.5	2
1	0		2.5	1	2.5	2.5	1
	2		2	2.5	2	2.5	2.5
			3	1.5	2.5	3	1*
			2.5	1	2		2.5

EXPERIMENTS WITH $10 PIES AND PAY

Dictator	Ultimatum
5	4
3	2
1	5
0	5
2	6
3	5
3	5
0	5
2	5
3	5
5	5
5	5
3	5
1	5
5	5
1	4*
3	5
3	5
0	4
0	5
0	4
1	5
5	3
2	5

APPENDIX C: THE POWERS OF THE TESTS

The powers of the CM, AD, KS, RS, and CF tests were investigated using a series of Monte Carlo simulations in which samples were drawn from pairs of alternative distributions that spanned a range suggested by the outcomes of previous experiments with simple bargaining games. Sample sizes of $N_1 = N_2 = 25$ and $N_1 = N_2 = 50$ were used. There were 5000 replications in each simulation. The distributions were based on a $5 pie. Letting X represent the proposal, the distributions used are

Distribution 1: $\Pr(X = 0) = 1$
Distribution 2: $X \sim U[1.5,3.5]$
Distribution 3: $X \sim U[2,3]$
Distribution 4: $\Pr(X = 2.5) = 1$
Distribution 5: $X \sim U[0,5]$
Distribution 6: A 50–50 mixture of 1 and 2
Distribution 7: A 50–50 mixture of 1 and 3
Distribution 8: A 50–50 mixture of 1 and 4

TABLE V

POWERS OF TESTS WHEN NOMINAL SIZE = 0.05

Distributions compared[a]	2 samples of 25 observations					2 samples of 50 observations				
	CF	CM	AD	KS	RS	CF	CM	AD	KS	RS
1 and 1	0.04					0.04				
2 and 2	0.04					0.04				
3 and 3	0.05					0.04				
4 and 4	0.04					0.04				
5 and 5	0.04	0.05	0.05	0.03	0.05	0.04	0.06	0.05	0.03	0.05
6 and 6	0.03					0.04				
7 and 7	0.04					0.03				
8 and 8	0.02					0.02				
1 and 2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1 and 3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1 and 4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1 and 5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1 and 6	0.77	0.94	0.96	0.94	0.89	1.00	1.00	1.00	1.00	0.99
1 and 7	0.77	0.93	0.96	0.95	0.89	1.00	1.00	1.00	1.00	0.99
1 and 8	0.87	0.93	0.96	0.94	0.89	0.93	1.00	1.00	1.00	0.99
2 and 3	0.87	0.35	0.61	0.21	0.07	1.00	0.82	0.98	0.67	0.06
2 and 4	1.00	1.00	1.00	1.00	0.11	1.00	1.00	1.00	1.00	0.12
2 and 5	0.98	0.58	0.84	0.43	0.07	1.00	0.98	1.00	0.93	0.07
2 and 6	0.99	0.93	0.96	0.88	0.88	1.00	1.00	1.00	1.00	1.00
2 and 7	1.00	0.90	0.96	0.85	0.90	1.00	1.00	1.00	1.00	1.00
2 and 8	1.00	1.00	1.00	0.98	0.90	1.00	1.00	1.00	1.00	1.00
3 and 4	1.00	1.00	1.00	1.00	0.11	1.00	1.00	1.00	1.00	0.11
3 and 5	1.00	0.97	1.00	0.90	0.08	1.00	1.00	1.00	1.00	0.08
3 and 6	1.00	0.99	1.00	0.98	0.87	1.00	1.00	1.00	1.00	0.99
3 and 7	0.99	0.93	0.96	0.89	0.89	1.00	1.00	1.00	1.00	0.99
3 and 8	1.00	1.00	1.00	0.98	0.90	1.00	1.00	1.00	1.00	0.99
4 and 5	1.00	1.00	1.00	1.00	0.11	1.00	1.00	1.00	1.00	0.12
4 and 6	1.00	1.00	1.00	1.00	0.85	1.00	1.00	1.00	1.00	0.99
4 and 7	1.00	1.00	1.00	1.00	0.86	1.00	1.00	1.00	1.00	0.99
4 and 8	0.89	0.94	0.96	0.88	0.88	0.94	1.00	1.00	1.00	0.99
5 and 6	1.00	0.91	0.97	0.86	0.89	1.00	1.00	1.00	1.00	0.99
5 and 7	1.00	0.95	0.99	0.92	0.89	1.00	1.00	1.00	1.00	0.99
5 and 8	1.00	1.00	1.00	0.99	0.90	1.00	1.00	1.00	1.00	1.00
6 and 7	0.37	0.07	0.07	0.06	0.05	0.86	0.09	0.12	0.11	0.05
6 and 8	0.86	0.13	0.17	0.24	0.06	1.00	0.32	0.58	0.65	0.06
7 and 8	0.83	0.13	0.17	0.24	0.06	1.00	0.30	0.59	0.65	0.06

[a] Comparisons of identical distributions give the empirical size of the test.

In the mixture distributions, X was sampled from distribution 1 with probability 0.5 and from the other member of the mixture with probability 0.5. The tie-breaking procedure described in the text was used in sampling from all distributions.

The simulations were conducted using a nominal size of 0.05. The exact distribution of the KS statistic is discrete and has no mass at the 0.05 point. The critical value corresponding

to the nearest lower size was used in the simulations. Critical values for the AD, CM, KS, and RS tests were obtained from the tables described in Section 4A. The critical value for the CF test is based on the chi-square distribution with 4 degrees of freedom. The results of the simulations are shown in Table V. Confidence intervals for the true powers of the tests can be computed using the normal approximation to the binomial distribution. When the distributions compared in Table V are identical, the powers of the tests equal their sizes (probabilities of Type I errors). The sizes of the AD, CM, KS, and RS tests were computed for only one of the 8 distributions because the distributions of their test statistics under the null hypothesis are independent of the distribution from which the data are sampled. The finite-sample distribution of the CF test statistic depends on the distribution from which the data are sampled, so the size of the CF test was computed for all nondegenerate distributions of the data.

REFERENCES

BINMORE, K., SHAKED, A., AND SUTTON, J. (1984). "Fairness or Gamesmanship in Bargaining: An Experimental Study," ICERD Discussion Paper, London School of Economics.

BINMORE, K., SHAKED, A., AND SUTTON, J. (1985). "Testing Noncooperative Bargaining Theory: A Preliminary Study," *Amer. Econom. Rev.* **75,** 1178–1180.

BOLTON, G. E. (1991). "A Comparative Model of Bargaining: Theory and Evidence," *Amer. Econ. Rev.* **81,** 1096–1136.

EPPS, T. W., AND SINGLETON, K. J. (1986). "An Omnibus Test for the Two-Sample Problem Using the Empirical Characteristic Function," *J. Statist. Comput. Simul.* **26,** 177–203.

GÜTH, W., SCHMITTBERGER, R., AND SCHWARZE, B. (1982). "An Experimental Analysis of Ultimatum Bargaining," *J. Econ. Behav. Organ.* **3,** 367–388.

KIM, P. J., AND JENNRICH, R. I. (1974). "Tables of the Exact Sampling Distribution of the Two-sample Kolmogorov–Smirnov Criterion, D_{mn}, $m \leq n$," in *Selected Tables in Mathematical Statistics I,* (H. L. Harter and D. B. Owen, Eds.), pp. 79–170. Providence, RI: Amer. Math. Soc.

NEELIN, J., SONNENSCHEIN, H., AND SPIEGEL, M. (1988). "A Further Test of Noncooperative Bargaining Theory," *Amer. Econ. Rev.* **78,** 824–836.

OCHS, J., AND ROTH, A. E. (1989). "An Experimental Study of Sequential Bargaining," *Amer. Econ. Rev.* **79,** 355–384.

ROTH, A. E., PRASNIKAR, V., OKUNO-FUJIWARA, M. AND ZAMIR, S. (1991). "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," *Amer. Econ. Rev.* **81,** 1068–1095.

SAVIN, N. E. (1984). "Multiple Hypothesis Testing," in *Handbook of Econometrics,* Vol. 2. (Z. Griliches and M. D. Intrilligator, Eds.), Amsterdam: North Holland.

SHORACK, G. R., AND WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics.* New York: John Wiley.

SPIEGEL, M., CURRIE, J., SONNENSCHEIN, H. AND SEN, A. (1990). "Fairness and Strategic Behavior in Two-Person, Alternating-Offer Games: Results from Bargaining Experiments," paper presented to the Fifth International Conference on the Foundation and Applications of Utility, Risk, and Decision Theories, Duke University, Durham, N. C.

STAHL, I. (1972). *Bargaining Theory,* Stockholm: Economic Research Institute.

THALER, R. (1986). "The Psychology and Economics Handbook: Comments on Simon, on Einhorn and Hogarth, and on Tversky and Kahnemann," *J. Bus.* **59,** S95–S100.

WILCOXON, F., KATTI, S. K., AND WILCOX, R. A. (1973). "Crtical Values and Probability Levels for the Wilcoxon Rank Sum Test and the Wilcoxon Signed Rank Test," in *Selected Tables in Mathematical Statistics I,* (H. L. Harter and D. B. Owen, Eds.), pp. 171–260. Providence, RI: Amer. Math. Soc.