Article

American Behavioral Scientist
55(8) 976–986
© 2011 SAGE Publications
Reprints and permission: http://www.sagepub.com/journalsPermissions.nav
DOI: 10.1177/0002764211409190
http://abs.sagepub.com

Theoretical Advances Spurred by "Stubborn Facts": A Defense of the Behavioral Economics Paradigm

William S. Neilson[1]

Abstract

Economists have regarded behavioral economics as contributing evidence violating accepted theories. This paper contends that a second, perhaps more important, contribution is the new theory that has arisen to address these challenges. To illustrate, the paper examines how a single observed fact, cooperation in the finitely-repeated prisoners' dilemma, has led to extensions of the folk theorem as well as to new concepts including sequential equilibrium, quantal response equilibrium, psychological game theory, and pregame and post-game perfection. Thus, the observed departures from previously-standard theories have led to not only new understandings of behavior, but also to additions to the economics toolkit.

Keywords

game theory, prisoners' dilemma, Nash equilibrium

The rise of experimental methods in economic research have led to what one might consider two basic tenets of behavioral economics: (a) For every theory, there exists an experimentalist clever enough to generate evidence violating it. (b) For every pattern uncovered through experiments, there exists a theorist clever enough to devise a model that can accommodate it. Of course, neither of these two truths is absolute, because there are some theoretical properties that cannot be tested in the lab (like

[1] University of Tennessee, Knoxville, TN, USA

Corresponding Author:
William S. Neilson, Department of Economics, University of Tennessee, Knoxville, TN 37996-0550, USA
Email: wneilson@utk.edu

differentiability), and there are some experimental patterns that so far have proven stubborn opponents for theorists (like double oral auctions). Nevertheless, these two tenets contain enough truth that they deserve some discussion, and this article is devoted to the second of the two.

In particular, this article concerns advances to the theory of noncooperative games that have arisen from the single "stubborn fact" that people seem to cooperate in a finitely repeated prisoners' dilemma game even though the prevailing solution concepts predict no cooperation at all. Through the years a number of new theories have been proposed. All of them share the attribute of elegance, but, more important, they all expand the way economists think about individuals and games. The contribution of behavioral economics, then, expands far beyond the collection of stubborn facts and their refuted theories; the contribution includes all of the theoretical advances fostered by the experimental evidence.

The pattern that emerges from this article shows that theorists begin the process of explaining cooperation by changing the game in important ways, sometimes subtly and sometimes not so subtly. Often these games cannot be solved with existing equilibrium concepts, leading theorists to devise new solution concepts for the new games. These new equilibrium concepts apply not just to the new game under consideration, but also to other important games arising in other areas of economics, making the contributions fundamental to the discipline. And all of this arises from a simple pattern in the experimental literature.

The Finitely Repeated Prisoners' Dilemma

The prisoners' dilemma is a two-player, non-zero-sum game of the form shown in Figure 1. Both players can choose to either "cooperate" or "defect." The payoffs are determined by the appropriate cell in the table, with the first payoff in each pair going to the row player, who chooses the row in the table, and the second payoff in each pair going to the column player, who chooses the column. Both players know their own and their opponent's payoffs, and the two players make their moves simultaneously.

If the game is played just once, then standard theory predicts that both players will defect. To see why, think about the decision of the row player. If Column cooperates, Row can cooperate and earn 12, or Row can defect and earn 16. Defecting generates a higher payoff. If Column instead chooses to defect, Row can cooperate and earn 2 or defect and earn 6, and once again defecting is better. In this game, defecting is a dominant strategy, that is, a best response to anything the other player can do. The theory of Nash equilibrium states that players' choices must be best responses to their opponents' equilibrium choices, in which case players must use their dominant strategies. Consequently, the only equilibrium of the game has both players defecting.

Theory predicts the same outcome when the game is played a fixed, finite number of times, say T. The relevant solution concept is subgame perfect equilibrium, which requires that equilibrium strategies in the entire game constitute a Nash equilibrium in every subgame, where a subgame is defined as any portion of the game beginning at a

		Column player	
		Cooperate	Defect
Row player	Cooperate	12, 12	2, 16
	Defect	16, 2	6, 6

Figure 1. Standard prisoners' dilemma game

decision node and continuing through the end of the game. The finitely repeated prisoners' dilemma has subgames corresponding to the remainder of the game from Period T, from Period $T - 1$, and so on back to the entire game beginning in Period 1. In Period T the two players face the exact game depicted above, and the only Nash equilibrium has both players defecting. In Period $T - 1$ they look forward and deduce that both players will defect in Period T, so cooperating now has no impact on play in the future. Because each player gets the same future payoff of 6 no matter what happens this period, they both choose to defect in Period $T - 1$. This reasoning extends all the way back to the beginning of the game, and the only subgame perfect equilibrium has both players defecting every period.

Experimental evidence, though, shows that subjects regularly cooperate in the finitely repeated prisoners' dilemma. In a meta-analysis covering 130 experiments conducted between 1958 and 1992, Sally (1995) finds an average cooperation rate of nearly 50%, and this is reflective of the consensus among behavioral economists regarding cooperation rates. Obviously, 50% cooperation clashes with the 0% prediction, which led theorists to devise explanations of the cooperative behavior.

Theoretical Advances Stemming From Observed Cooperation

As argued above, the finitely repeated prisoners' dilemma, when coupled with the reigning assumptions governing player preferences, precludes any cooperation. To allow for cooperation, either the game must be changed, the players' preferences must be changed, or both. This section provides an overview of how theorists made these changes and the greater contributions they made to game theory in the process.

The earliest efforts involved changing the game. In particular, they involved extending the game from a finite to an infinite number of periods (for an excellent and rigorous overview, see Fudenberg & Maskin, 1986). After all, whereas people (and firms) do not live forever, the exact time of their deaths are seldom known, so one rarely knows when the last period of interaction will be. Infinitely repeated games capture this notion of an unknown time for the last interaction. Working with an infinite number of periods invoked some challenges, although their solutions have become standard operating procedure over the years. The foremost problem was that subgame perfect equilibrium is typically found through the process of backward induction, starting at the end of the game and working back toward the beginning. Infinitely repeated games have no last period, and therefore backward induction is impossible. Instead, solving the game required classifying the entire set of subgames and then checking for the mutual best response property in each class of subgames. A second issue, which no longer seems to be one, is that although microeconomists were aware of discounting of the future, they rarely used it in models. Infinitely repeated games forced discounting front and center (as did Rubinstein's [1982] model of bargaining, also built on subgame perfection) because not only did it provide a way to aggregate an infinite number of payoffs through the discounted present value, but the level of the discount factor proved to be the key to the sustainability of cooperation.

Infinitely repeated games can occur in two ways. One is for the probability that the game continues for another period to be constant and equal to one, and this approach requires discounting the future to avoid comparisons of infinite payoffs. The other is for the probability that the game continues for another period to be constant but smaller than one, so that the game ends in finite time with probability one, but the probability of another period is always the same. The result from either of these approaches is the same. If the discount factor is sufficiently high (that is, if players are sufficiently patient) or if the continuation probability is sufficiently high, then *any* feasible payoff combination that dominates the dominant strategy combination in the prisoners' dilemma can be supported as the average payoff combination in equilibrium. More precisely, in terms of the specific game depicted in Figure 1, there exists a subgame perfect equilibrium in which both players receive 12 each period, and there exists an equilibrium in which both players average 9 each period, and there exists one in which the row player averages 14 and the column player averages 8, and so on. Not only is cooperation possible, but any level of cooperation is possible.

These cooperative outcomes are sustained through the following strategies, known as grim trigger strategies. Take the best symmetric outcome as an example, where both players earn 12 each period. The row player's strategy is to cooperate in the first period, cooperate in the second period if both players cooperated in the first period and defect otherwise, cooperate in the third period if both players cooperated in both of the first two periods and defect otherwise, and so on. Or, put differently, cooperate in the first period and continue cooperating as long as there were no prior defections by either party, and defect forever if anyone, including the row player, defected in any period. The strategy is a trigger strategy because a single defection by either player triggers punishment. It is a grim strategy because the punishment lasts forever. The math to check the conditions is straightforward. Following the proposed strategy leads to a payoff of 12 each period. If the discount factor (or continuation probability) is $\delta < 1$, then the present value of this payoff stream is $12/(1 - \delta)$. If a player deviates from the proposed strategy by defecting, that player earns 16 in that period but then earns 6 in every subsequent period, for a present value of $16 + \delta 6/(1 - \delta)$. Cooperating beats defecting if

$$\frac{12}{1-\delta} > 16 + \delta \frac{6}{1-\delta},$$

or $\delta > 0.4$. So, if interactions occur sufficiently often that the discount rate between interactions is greater than 0.4, which corresponds to an interest rate less than 67%, the cooperative outcome can be sustained in equilibrium using the grim trigger strategy. Intuitively, sufficient patience means that players care about the future enough that punishment hurts them, whereas players who are sufficiently impatient cannot be punished because they discount the future too heavily.

This result on cooperation extends beyond the prisoners' dilemma and has become one of the best-known results in game theory—the folk theorem of infinitely repeated games. It is a folk theorem because no one is sure who proved its first version, so credit cannot be attributed to a single set of authors. Nevertheless, it states that if players are sufficiently patient, or if continuation probabilities are sufficiently high, any feasible payoff combination that dominates the Nash equilibrium of a single play of the game can be supported as the average outcome of a subgame perfect equilibrium of the infinitely repeated game. Cooperation is possible not just in the prisoners' dilemma, but in all games that are repeated an infinite number of times. Furthermore, punishment need not come directly from the injured party, and group punishment can lead to social norms, as in Kandori (1992).

Benoit and Krishna (1985) cleverly showed that a trigger strategy can support cooperation even when the number of periods is finite for some games, but not the prisoners' dilemma. What is needed is more than one Nash equilibrium of the underlying game. Consider the augmented prisoners' dilemma shown in Figure 2. Each player has an additional strategy, eradicate, which leads to very low payoffs. The augmented prisoners' dilemma has two Nash equilibria, with defecting a best response to defecting and eradicating a best response to eradicating. No one wants to play eradicate, but it is still a best response to the other player choosing eradicate.

In this game, players can cooperate if there are at least two periods, but they cannot cooperate in every period. Specifically, suppose that the game is played exactly twice and they use the following strategy: cooperate in the first period, defect in the second period if both players cooperated in the first period, and eradicate in the second period if either player defected in the first period. The calculations are again straightforward. With no discounting, following the proposed strategy yields a total payoff of 12 + 6, whereas defecting in the first period yields a total payoff of 16 + 0, which is worse.

The key to the augmented prisoners' dilemma is that there are two Nash equilibria, not just one, and the new Nash equilibrium is worse than the existing one. The bad Nash equilibrium is used to punish deviations from the cooperative strategy. However,

		Column player		
		Cooperate	Defect	Eradicate
Row player	Cooperate	12, 12	2, 16	-2, -2
	Defect	16, 2	6, 6	-1, -1
	Eradicate	-2, -2	-1, -1	0, 0

Figure 2. Augmented prisoners' dilemma game

cooperation cannot be sustained in the last period because there is no future in which to punish using the bad Nash equilibrium. The Benoit and Krishna approach, then, points out the importance of multiple Nash equilibria, especially bad ones, and the inevitability of endgame effects in which cooperation erodes. These endgame effects explain, for example, why lame duck administrations accomplish little and why workers near the ends of their careers are difficult to motivate.

In spite of the importance of their contributions to theory, neither of these approaches describes the game actually tested in the laboratory, as they either require a different form of repetition or a game with a different action space. It would be nice to have a model that gives rise to cooperation without changing the underlying structure of the finitely repeated prisoners' dilemma. The first successful attempt at this came in what is now known as the Gang of Four paper (Kreps, Milgrom, Roberts, & Wilson, 1982). Their approach added a tiny bit of irrationality to the game. Specifically, there is a small probability that the player's opponent is irrational and follows a tit-for-tat strategy instead of an equilibrium strategy. Tit-for-tat dictates that a player begin by cooperating and then play whatever his or her opponent played in the previous period. Kreps et al. (1982) showed that when there is a small exogenous chance of facing an irrational opponent, rational opponents cooperate in equilibrium in order to convince their opponents that they are irrational, thereby generating cooperation early in the game.

This approach identified a gap in the theory, because subgame perfection cannot be directly applied to this game. For a strategy combination to be subgame perfect, it must constitute a Nash equilibrium on every subgame, and subgames begin when the player making the decision at that node can identify the exact node at which he or she plays. In other words, subgames begin at singleton decision nodes. When irrationality is introduced, though, many of the nodes are not singletons. Suppose, for example, that Row and Column both cooperated in Period 1, and consider Row's problem in Period 2. It could be that Column was irrational and played the tit-for-tat strategy in Period 1, or it could be that Column was rational but cooperated in order to mimic the tit-for-tat strategy. Row does not know whether she landed at a node emanating from an irrational-type Column or a rational-type Column, so the game from Period 2 does not constitute a subgame. Consequently, subgame perfection cannot govern play. To deal with this problem, two members of the Gang of Four, Kreps and Wilson (1982a, 1982b), developed a new solution concept: sequential equilibrium.

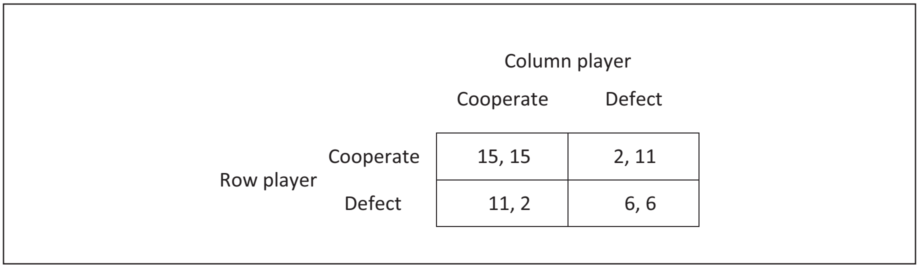
Sequential equilibrium requires that players form beliefs about their opponents' types, that their actions be best responses to their beliefs, and that their beliefs be consistent with play both in and out of equilibrium. It is similar to perfect Bayesian equilibrium except that it restricts beliefs off of the equilibrium path. Sequential equilibrium has become the standard tool for solving dynamic games of complete information, and not just the finitely repeated prisoners' dilemma with both rational and irrational players.

Sequential equilibrium has played a central role in industrial organization literature, especially regarding entry deterrence. Subgame perfection generates the result that rational entry cannot be deterred in equilibrium. Sequential equilibrium, on the other hand, provides an avenue for deterrence. If there is a possibility of irrationality, in this case the incumbent fighting the entrant no matter what the cost, then firms can mimic this irrationality and thereby deter entry. The "irrational manager" strategy has become a mainstay of business training.

McKelvey and Palfrey (1992, 1995) took the idea of rational responses to irrationality a step further. Suppose that some players might be altruistic instead of selfish, thereby making "mistakes" in the prisoners' dilemma because they value their opponents' payoffs in addition to their own. The subgame perfect equilibrium strategy may no longer be a best response in the presence of these "mistakes." McKelvey and Palfrey constructed a new equilibrium concept, quantal response equilibrium, in which players make best responses to both the intended strategies and the mistakes made by other players. The benefit of this concept is that it is parameterized and can be estimated using maximum likelihood methods, allowing for comparisons across games and across data sets.

Yet another approach comes from the idea of psychological games, introduced by Geanakoplos, Pearce, and Stacchetti (1989) and expanded upon by Rabin (1993). This idea rests on a notion of reciprocity, specifically a preference for rewarding kind behavior and punishing unkind behavior. To see how this can impact play in the prisoners' dilemma, look back at the payoffs in Figure 1. If Row regards Column's cooperation as kind and feels compelled to reward it, Row receives extra utility from playing cooperate and loses utility from playing defect. This can change the payoffs to those in Figure 3, where the extra utility from cooperating in response to perceived kindness is +3 and the loss from defecting in response to perceived kindness is –3, and the same changes are made for Column's payoffs. Now there are two Nash equilibria, one in which both players defect and one in which both players cooperate.

The difficult part of the analysis is making workable constructs out of the notions of kindness and unkindness. Psychological games allow players to form beliefs about their opponents' behavior and beliefs, and requires that in equilibrium beliefs must be consistent with reality. This extends rational expectations to games in a particular way. Rabin's primary contribution arose from formalizing the mapping from beliefs about opponents' actions and beliefs into responses to kindness and unkindness.



		Column player	
		Cooperate	Defect
Row player	Cooperate	15, 15	2, 11
	Defect	11, 2	6, 6

Figure 3. Prisoners' Dilemma Game with Reciprocity

A New Direction—Thinking Outside the Game

The papers discussed in the preceding section are 15 to 30 years old, and are all heavily cited.[1] They have had an obvious influence on economics literature, both the experimental literature and the theoretical literature. They do not provide the only explanations of cooperation, however, and a new, very recent literature models cooperation by placing the repeated prisoners' dilemma within the context of the subjects' interactions within society at large.

One approach begins with the observation that subjects come to a laboratory with a long history of social interactions, and they may have either learned some behaviors or restricted themselves from certain behaviors in ways that matter for the game. This is the approach of Conley and Neilson (2009). They assume a large population whose members will be paired off to play a prisoners' dilemma game similar to the one in Figure 1. Before the game begins, in what is referred to as the pregame, each player can choose a subset of the available actions, called a list. In the prisoners' dilemma, a player can choose a list with only cooperate in it, only defect, or both cooperate and defect. After those choices are made, players are randomly matched, observe each other's lists, and then decide whether they want to play against each other. If they both elect to play, they choose actions from their lists, receive their payoffs, and exit the game. If one or both elect not to play, they both pay a delay cost and then are randomly rematched with the population. Conley and Neilson refer to this larger game as an endogenous game, with the name reflecting the fact that both the players in the prisoners' dilemma and their action sets are endogenous within the larger game.

The endogenous game format reflects two attributes that are consistent with the outside world. One is that people have some say over whom they interact with, and refusing to interact with one individual is hardly tantamount to refusing to interact with anyone. The other is that individuals can make decisions regarding who they are, and in an endogenous game an individual is defined by his or her list. The list both restricts the strategies they can play and serves as the individual's public face on the basis of which others decide to interact with them. For the former reason, the list is a commitment device, but for the latter it is also a way to fit in with or appeal to society at large.

Equilibria in this game tend to be social norm equilibria where everyone does the same thing; specifically, they all have the same list, and they all play the same element from that list. Cooperation can occur in equilibrium if the cost of refusing to play is not too large. Enforcement of the equilibrium comes through ostracism rather than punishment, though, because when individuals choose lists containing the action "defect," no one will interact with them. In the prisoners' dilemma, cooperation means that both players choose the dominated strategy, and the ensuing payoff combination dominates the Nash equilibrium payoff combination. This is also true in infinitely repeated games, because the folk theorem states that payoff combinations that dominate the Nash equilibrium payoff combination can be supported. In endogenous games, however, bad dominated strategies can also be supported in equilibrium, which means that this form of ostracism can generate harmful social norms that show up, for example, as risky teen behavior.

As with the Gang of Four model, endogenous games require a new solution concept, and for similar reasons. In the Gang of Four model there were no singleton decision nodes, and therefore no subgames, because one player could not observe the other player's type. In an endogenous game, each player observes his or her matched partner's list, and so players are fully informed about their opponents. They are not, however, fully informed about the rest of the population. This matters off the equilibrium path. To see why, suppose that the proposed equilibrium in the endogenous prisoners' dilemma has everyone choosing only "cooperate" in their lists. Player 1 does this and then in the first period is matched with someone who has "defect" in his or her list. This was not supposed to happen. But if Player 1 refuses to play in the first period, he or she is randomly rematched in Period 2, and cannot know what the rest of the population looks like given that one person deviated from the candidate equilibrium strategy. In Kreps and Wilson (1982a, 1982b), sequential equilibrium was developed to tie down beliefs about the type of a single opponent who deviates from the equilibrium path. In Conley and Neilson (2009), pregame perfect equilibrium was developed to tie down beliefs about the lists of everyone else when one player deviates from the equilibrium path.

Adding play before the prisoners' dilemma is one way to get cooperation. Adding play afterward is another. This route is explored by List, Neilson, and Price (2009), who allow for the existence of a postgame whose details are unknown to the modeler. In particular, participants in the finitely repeated prisoners' dilemma may or may not have further interactions, unobserved by the modeler or experimenter, after the last period of the prisoners' dilemma. If they have further interactions, the unobserved future provides an opportunity for punishment for any deviations in the observed prisoners' dilemma. List et al. provide an equilibrium concept, termed post-game perfection, which uses beliefs about future, unobserved punishment opportunities.

Post-game perfection implies the existence of a testable hypothesis regarding behavior in a finitely repeated prisoners' dilemma. If participants face future interactions, they can sustain cooperation all the way through the last period of the finitely repeated game. If they do not face future interactions, they cannot. List et al. (2009)

run a field experiment in which some groups are known to have future interactions and some groups have outsiders who cannot be punished after the game ends. The insider-only groups sustain cooperation throughout the experiment, but cooperation unravels in the outsider groups. This evidence suggests that postgame considerations matter and that models, and equilibrium concepts, should include them. Future research can look at ways to impart structure on the postgame beliefs.

Conclusions

The goal of this article was to show that the contribution of behavioral economics is more than just the questioning of the standard maximization models and the call for a new paradigm. Instead, behavioral economics has spun off many critical contributions to economic theory. This article illustrated that point by discussing one stubborn fact, that people tend to cooperate in the finitely repeated prisoners' dilemma, a stubborn fact that led to new ways to think about games and new ways to solve them. Yet this is just one example, and behavioral economics has led to important theoretical advances in the modeling of behavior toward risk, contingent valuation, and many other areas of economics. The advances of behavioral economics, and the importance of the existing paradigm, do not just arise from the experimentalists clever enough to generate evidence violating existing theories. Instead, a crucial contribution of behavioral economics comes from those theorists clever enough to devise models that accommodate the new evidence, especially when those new theories provide spillovers to areas beyond the simple games tested in the laboratory.

Declaration of Conflicting Interests

The author(s) declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research and/or authorship of this article.

Note

[1] The ISI Web of Knowledge lists the following citation numbers: Benoit and Krishna (1985)—118; Fudenberg and Maskin (1986)—559; Geanakoplos et al. (1989)—107; Kandori (1992)—190; Kreps et al. (1982)—541; Kreps and Wilson (1982a)—689; Kreps and Wilson (1982b)—743; McKelvey and Palfrey (1992)—156; McKelvey and Palfrey (1995)—238; Rabin (1993)—600. This falls just shy of 4,000 total citations for the 10 papers.

References

Benoit, J. P., & Krishna, V. (1985). Finitely repeated games. *Econometrica*, *53*, 905-922.

Conley, J. P., & Neilson, W. (2009). Endogenous games and equilibrium adoption of social norms and ethical constraints. *Games and Economic Behavior*, *66*, 761-774.

Fudenberg, D., & Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, *54*, 533-554.

Geanakoplos, J., Pearce, D., & Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, *1*, 60-79.

Kandori, M. (1992). Social norms and community enforcement. *Review of Economic Studies*, *59*, 63-80.

Kreps, D. M., Milgrom, P., Roberts, J., & Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, *27*, 245-252.

Kreps, D. M., & Wilson, R. (1982a). Reputation and imperfect information. *Journal of Economic Theory*, *27*, 253-279.

Kreps, D. M., & Wilson, R. (1982b). Sequential equilibria. *Econometrica*, *50*, 863-894.

List, J., Neilson, W., & Price, M. (2009). *The effects of group membership in a strategic environment: Evidence from the field*. Manuscript, University of Tennessee, Knoxville.

McKelvey, R. D., & Palfrey, T. R. (1992). An experimental study of the centipede game. *Econometrica*, *60*, 803-836.

McKelvey, R. D., & Palfrey, T. R. (1995). Quantal response equilibria for normal-form games. *Games and Economic Behavior*, *1*, 6-38.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, *83*, 1281-1302.

Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica*, *50*, 97-109.

Sally, D. (1995). Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society*, *7*, 58-92.

Bio

William S. Neilson holds the J. Fred Holly Chair of Excellence and is a professor of economics at the University of Tennessee, Knoxville. He is currently editor-in-chief of the *Journal of Economic Behavior & Organization*.