Rationality in International Relations: A Game-Theoretic and Empirical Study of the U.S.-China Case

Author(s): Catherine C. Langlois and Jean-Pierre P. Langlois

Source: *World Politics*, Apr., 1996, Vol. 48, No. 3 (Apr., 1996), pp. 358-390

Published by: Cambridge University Press

Stable URL: https://www.jstor.org/stable/25053970

RATIONALITY IN INTERNATIONAL RELATIONS
A Game-Theoretic and Empirical Study of the U.S.-China Case

By CATHERINE C. LANGLOIS and
JEAN-PIERRE P. LANGLOIS*

I. INTRODUCTION

THE premise that states behave as rational actors in the anarchy of international relations has led to a proliferation of game models.[1] To advance the debate about their relevance, we reexamine a central game-theoretic concept of rationality, the subgame-perfect equilibrium,[2] focusing on a "countervailing" effect that is found to be involved in any equilibrium strategy. Countervailing behavior broadens our understanding of what is rational and thus holds out the promise of compatibility with observed behavior in international relations. And indeed when the data on U.S.-China relations from 1972 to 1988 are anlyzed in light of this countervailing model, we find that each country's behavior toward the other can be interpreted as rational.

In most game-theoretic studies of international relations, the behavior of the players is inferred from the a priori stipulation of the structure of the game and from various considerations on the nature of rationality and the incompleteness of information. However, there is much debate over whether the game structures introduced in the models have any relevance to the problems being addressed, and such basic doubt puts the rational strategies that are derived from them in danger of

* The authors, considering their contributions to be equal, have listed their names alphabetically.

[1] An extensive survey of that literature is given by Barry O'Neill, "A Survey of Game Models of Peace and War," in Robert Aumann and Sergiu Hart, eds., *Handbook of Game Theory,* vol. 2 (New York: Springer-Verlag, 1992).

[2] Strategies are in subgame-perfect equilibrium (SGPE) if each is best against the other(s) in any contingency. The concept of SGPE comes from Reinhart Selten, "Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory* 4 (1975).

World Politics 48 (April 1996), 358–90

being consigned to irrelevance. As Jervis puts it, "How do we know that a situation resembles a Prisoner's Dilemma?"[3] In fact, it has been argued that simple observation of the interaction between states speaks against the usefulness of a game-theoretic approach, whatever the structure of the game, because it reveals allegedly irrational behavior such as inertia, cultural bias, or bureaucratic delay. So even if the relevance of a particular game structure can be argued convincingly, the rationality of real-world players' responses to it can be questioned. And such doubt is only reinforced by the technical requirements often associated with the solution concepts. For instance, equilibrium play is believed to require that players share a "common conjecture" about each other's choice of strategy,[4] a premise that is hardly credible when states search for a tacit modus vivendi and seek to adapt to an ever-changing environment.

Because our interest is in long-term relationships, we focus on the so-called repeated games and show that, in order to participate in a subgame-perfect equilibrium (SGPE), a strategy must exhibit a combination of two effects on the opponent. On the one hand, it has an *expectation* effect by which a player expresses his belief about future developments and may punish his opponent for not meeting that expectation. On the other hand, it has a *countervailing* effect by which a player uses the shadow of the future to manipulate his opponent's long-term outcome, punishing or rewarding as he sees fit. Although players are likely to entertain expectations, they need not actively punish their opponent for not meeting them, especially if they are unsure about that opponent's intentions. In fact, the countervailing effect satisfies the need for retaliation while avoiding any need for coordination.[5] We therefore focus on the countervailing strategies that rely solely on that effect and find that they considerably broaden our understanding of what is rational. Thus, bureaucratic inertia, delayed responses, reciprocity, aggressivity, and tacit bargaining can all find expression in a countervailing strategy. Moreover, when we examine the data on U.S.-China relations in light of these ideas, we find empirical support for countervailing behavior, but none for the punishment that the expectation effect could involve.

We chose to examine China-U.S. relations from 1972 to 1988, the

[3] Robert Jervis, "Realism, Game Theory, and Cooperation," *World Politics* 40 (April 1988).

[4] Loosely speaking, this means that players share the same expectation of future interaction, thus determining the path of future play. See James D. Morrow, *Game Theory for Political Science* (Princeton: Princeton University Press, 1994).

[5] If each side punishes its opponent for not meeting its expectation, those expectations must be identical or they will remain hopelessly out of equilibrium. This requires a *coordination* of expectations.

years during which they were emerging and taking on official stature. In such a context, neither country would be expected to have a preconceived idea of the particular path the relationship would take. This is precisely the situation in which rational strategies are countervailing. And, indeed, we do find, in conducting a statistical analysis of U.S.-China interaction as coded in McClelland's World Events Interaction Survey (WEIS), that the data are consistent with their adoption.[6] When examining the data in light of our game-theoretic model, we do not need to impose a narrowly defined game structure a priori. Even more importantly, we are able to infer the parameters of the game itself from an analysis of behavior. As a result, this game-theoretic approach provides the tools for detecting possible changes in the underlying game structure itself, and the development of China-U.S. relations from 1972 to 1988 yields a case study of particular interest. Indeed, a cursory overview of this history suggests that, if a game-theoretic approach can capture the evolution of this interaction, we should expect a bifurcation of behavior and priorities when China opens its doors to economic relations with the West in 1979. We therefore break our sample into two subperiods: the period from 1972 to 1978 and that from 1979 to 1988. This partitioning is of course motivated by historical events (President Nixon's visit in 1972 and the beginning of China's open-door policy in 1979), but it also turns out to be appropriate when conducting statistical tests for break points in the data. Interestingly, two distinct game structures emerge from the analysis. During the first sample period, the U.S. appears to have Prisoner's Dilemma–like payoffs while the Chinese payoff is characteristic of a Deadlock-type game.[7] During the second period, we find a true Prisoner's Dilemma game with strong incentives on both sides to engage in cooperative behavior. Thus, while rational play in the 1980s finds expression in positive reactivity to each other's moves, prior to 1979 it takes the form of an inverse response to Chinese initiatives on the part of the U.S., a behavior that moves China to a more cooperative stance despite its Deadlock-type payoff structure.

The paper is organized as follows. In Section II we review the exist-

[6] Our characterization of SGPEs in Appendix 2 also suggests how to test whether countervailing strategies have been adopted, by searching in the data for traces of coordination. Such tests, performed on data for U.S.-China relations for the period 1972 to 1988 and reported in Appendix 4, support the hypothesis that a pure countervailing strategy rather than any other type of strategy was adopted by the two countries during the period under consideration.

[7] In 2×2 games, if C denotes cooperate, D defect, and XY the choices X by row and Y by column, a Prisoner's Dilemma structure for row means his preferences satisfy DC > CC > DD > CD. There are several Deadlock-type games. The relevant one here satisfies DD > CD > DC > CC.

ing literature both on game-theoretic and empirical approaches to international relations and on U.S.-China relations, as we introduce our modeling and empirical approach. Section III describes the model formally in three successive sections of increasing complexity. Finally, Section IV presents the empirical work and its interpretation.

II. An Overview of Related Work

Game-Theoretic Approaches

At the risk of being simplistic, one may distinguish three main classes of game-theoretic models.[8]

1. The *finite* games, usually in matrix form, were the first to be introduced in the political science literature. They range from what O'Neill has called *proto*–game theory to the more sophisticated models of Brams and Kilgour.[9]
2. The *Bayesian* games focus on how incomplete information about the preferences of one's opponent(s) affects decisions. They have been successfully applied to the modeling of international crises and nuclear deterrence.[10]
3. The *repeated* games focus on how the shadow of the future affects present decisions. They are best known for the study of Tit-for-Tat in the Prisoner's Dilemma[11] and are well suited to the study of long-term relationships.

At the core of game theory lies the concept of equilibrium: each actor adopts a strategy that is best for all contingencies, given that all others behave similarly. But behind this simple statement stands a mountain of controversy and refinements. First, what is best depends on what is known, believed, or expected. And although the various equilibrium refinements can stretch the scope of applications and weaken the need for information, they always require a highly structured and quantified view of the decision maker. Second, in the repeated game models of long-term relationships, the number of equilibria is vast.[12] And because a strategy in one equilibrium is usually not rational against an opponent's strategy in another equilibrium,

[8] We refer here to the noncooperative models that assume individualistic actors.

[9] O'Neill (fn. 1); Steven J. Brams and D. Marc Kilgour, *Game Theory and National Security* (New York: Basil Blackwell, 1988).

[10] See Robert Powell, *Nuclear Deterrence Theory: The Search for Credibility* (New York: Cambridge University Press, 1990).

[11] Robert Axelrod, *The Evolution of Cooperation* (New York: Basic Books, 1984).

[12] The folk theorem asserts that any "individually rational" outcome (i.e., no worse than the minimax) can be supported by a trigger strategy equilibrium in a repeated game. See Drew Fudenberg and Eric Maskin, "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information," *Econometrica* 54 (1986).

there is a need for coordination.[13] But if, as argued by Downs and Rocke,[14] states engage in "tacit" bargaining to manage their relationship, how could they *agree* on how they will play in order to influence each other later on through such agreed-upon play? Third, the reality of a state's decision-making process involves constraints, habits, or cultural biases that seem difficult to reconcile with the assumption of a rational actor on the international scene. Jervis adduces such commonly observed phenomena as bureaucratic delays and inertia as evidence that state decisions are not entirely rational.[15]

Some of the above criticism may, however, be based on a misperception of the resources of game theory. As we demonstrate in Appendix 3, equilibrium can be achieved without a highly restrictive coordination requirement because the strategies we call *countervailing* allow for flexible responses to actual opponent moves as they come to pass, without requiring players to entertain a specific expectation of what these moves will be. In fact, all equilibrium strategies can be broken down into two components: a countervailing effect and the *possible* punishment of an opponent who deviates from a particular expectation. This general characterization, formally spelled out in the theorem of Appendix 2, highlights the fact that a countervailing strategy is, so to speak, at the core of any rational strategy. Indeed, the addition of a punishment for unexpected play can even seem to be an unreasonable feature for the rational decision maker who recognizes the futility of trying to impose his own expectation on an opponent who is likely to think differently.[16]

A countervailing player focuses on his opponent's discounted payoff in an attempt to manipulate it. Because of the shadow of the future, what an opponent may gain or lose today can be reversed in a discounted sense beginning tomorrow, using appropriate retaliation or reciprocation. Future moves can thus be tailored to compensate in discounted terms for present disturbances. The player adopting a countervailing strategy does not need to see his expectations fulfilled, since an

[13] For example, it is not enough to recognize that one's opponent has chosen to play a trigger strategy (which involves collaborating as long as the other players do and reverting to noncooperative play for a fixed number of turns to punish the opponent for deviating from cooperative play). Each player must understand which *particular* trigger strategy has been chosen (how many punishment turns and which precise outcome is considered as the cooperative point) and chose to adopt the same one, since, if not, the strategies do not form an equilibrium.

[14] George W. Downs and David M. Rocke, *Tacit Bargaining, Arms Races, and Arms Control* (Ann Arbor: University of Michigan Press, 1990).

[15] Jervis (fn. 3).

[16] While countervailing is inherently reciprocative, the additional punishment of unexpected play could lead to retaliations for *cooperative initiatives*! This paradox is inherent to all trigger strategy schemes in continuous games.

unexpected move by his opponent, whether a cooperative initiative or an aggressive move, will be compensated for by retargeting the opponent's discounted payoff at the next turn. Moreover, the countervailing player can expect his behavior to be entirely optimal if he simply believes that his opponent will behave similarly but has no knowledge of which countervailing strategy his opponent will use among an extremely vast set of possibilities.

Since the focus of countervailing strategies is a determination of the *opponent's* discounted payoff, reaching some modus vivendi will build in inertia, while adjustments to unexpected moves will show up as reactivity. Thus a countervailing strategy can be highly reactive or involve inertia, it can be quasi-instantaneous or arbitrarily delayed, and it can be reciprocal, retaliative, or even counterreciprocal. Such behavioral components of strategy that have been viewed as unreconcilable with rationality thus find justification within the highly demanding SGPE concept. Moreover, a countervailing strategy is an ideal vehicle for Downs and Rocke's concept of tacit bargaining between states, since it focuses on how each side can influence the other side's payoff without presuming how he will behave.

Empirical Approaches

The relevance of mixed-motive games to the modeling of great power preferences has been examined by a number of authors.[17] Some argue that states are motivated by relative rather than absolute gains, which may create zero-sum conditions; others, such as Beer, claim that it is not possible to tell whether the game has a Prisoner's Dilemma structure.[18]

Although one might hope to resolve such questions empirically, Axelrod claims that only behaviors can be discerned and that payoffs cannot be inferred.[19] And indeed it is behavior that has been widely examined empirically and probed for evidence of reactivity and reciprocity. Since rational play must involve reactivity, such feedback is the preferred target of empirical search. One strand of literature examines arms race data for traces of reactive behavior. It is well known that empirical tests based on the Richardson model rarely reveal significant

[17] See, among others, George W. Downs, David M. Rocke and Randolph M. Siverson, "Arms Races and Cooperation," in Kenneth A. Oye, ed., *Cooperation under Anarchy* (Princeton: Princeton University Press, 1986); and Kenneth A. Oye, "Explaining Cooperation under Anarchy: Hypotheses and Strategies," in Oye, ed., *Cooperation under Anarchy*.

[18] Francis A. Beer, "Games and Metaphors," *Journal of Conflict Resolution* 30 (1986).

[19] Axelrod (fn. 11).

reactive behavior.[20] Another large body of literature examines foreign policy interaction as it manifests itself through events data.[21] However, what is tested in all these cases is a model of behavior that is posited to be cooperative *at the outset.* The issue is to fit "equations to data sets and then to assess the degree of fit and the values of the parameters."[22]

One way of circumventing this problem is to examine the data from an atheoretical perspective, searching for causal links without modeling a priori the form those links might take. The standard methodology in this case is vector autoregression (VAR).[23] Vector autoregression is associated with causality tests (Granger causality) that determine whether the set of lagged values of a given variable adds explanatory value to the VAR. Then, instead of searching for a significantly positive coefficient on rival behavior when explaining one's own behavior, Granger causality tests for significance of the *whole set* of lagged rival behavior variables in a given VAR in explaining one's own current behavior. The use of Granger causality tests to detect reactivity has proved successful using events data, although such tests have failed to predict causal reaction in both directions when applied to arms race data.[24]

It is a return to specific modeling of behavior that bolstered the case for reactivity in arms races. Williams and McGinnis show that when expectations of rival behavior are taken into account, the data provide evidence of reactivity.[25] A rational expectations model would predict that a state's military expenditures are determined by the previous period's decision and by any new information that cropped up subsequently. Such new information would appear as the error made in

[20] See Charles W. Ostrom, "Evaluating Alternative Foreign Policy Decision-Making Models: An Empirical Test between Arms Race Models and an Organizational Politics Model," *Journal of Conflict Resolution* 21 (1977); Steven J. Majeski and David Jones, "Arms Race Modeling: Causality Analysis and Model Specification," *Journal of Conflict Resolution* 25 (1981); and Kendall D. Moll and Gregory Luebbert, "Arms Race and Military Expenditure Models: A Review," *Journal of Conflict Resolution* 24 (1980).

[21] See William J. Dixon, "Reciprocity in United States-Soviet Relations: Multiple Symmetry or Issue Linkage?" *American Journal of Political Science* 30 (1986); and Michael D. Ward, "Cooperation and Conflict in Foreign Policy Behavior," *International Studies Quarterly* 26 (1982). See also Dina Zinnes, "Three Puzzles in Search of a Researcher," *International Studies Quarterly* 24 (1980); Bruce Russett, *The Prisoner's Insecurity: Nuclear Deterrence, the Arms Race and Arms Control* (San Francisco: W. H. Freeman, 1983); and Walter Isard and Charles Anderton, "Arms Race Models: A Survey and Synthesis," *Conflict Management and Peace Science* 8, no. 2 (1985).

[22] Zinnes (fn. 21).

[23] See John R. Freeman, John T. Williams, and Tse-Min Lin, "Vector Autoregression and the Study of Politics," *American Journal of Political Science* 33 (November 1989).

[24] See Joshua S. Goldstein and John Freeman, "U.S.-Soviet-Chinese Relations: Routine Reciprocity, or Rational Expectations?" *American Political Science Review* 85 (March 1991); and idem, *Three-Way Street: Strategic Reciprocity in World Politics* (Chicago: University of Chicago Press, 1990).

[25] See John Williams and Michael McGinnis, "Sophisticated Reaction in the U.S.-Soviet Arms Race: Evidence of Rational Expectations," *American Journal of Political Science* 32 (November 1988).

predicting current expenditures with past values. Thus the rational expectations arms race model predicts the lack of significance of past rival behavior in a state's current decision (since if the rival has behaved as expected, that behavior is already accounted for in its own past decisions and adding past rival behavior to its own past behavior to explain current expenditures would therefore be redundant) and embeds possible reactivity in a correlation of the errors made in predicting current expenditures on own past expenditures (since the error term contains the *unexpected* changes in rival behavior that warrant direct response). Williams and McGinnis do indeed find evidence of reactivity when testing such a model on U.S.-Soviet arms race data and on a wider set of data including economic indicators and events data.[26] But here again, it is behavior that is modeled and tested for; payoffs are not inferred.

The analysis of U.S.-China relations that we present here is most closely related to the work of Goldstein and Freeman and Goldstein.[27] We also use McClelland's WEIS data, and our empirical tests involve restricted VAR. Our approach differs from that of Goldstein and Freeman, however, in that we move away from the atheoretical VAR to model state interaction specifically. Like Goldstein, we analyze the U.S.-China dyad in isolation, relying on Goldstein and Freeman's finding that U.S.-China relations are relatively independent empirically from U.S.-Soviet and Soviet-Chinese relations. Goldstein begins with the description of U.S. and Chinese payoffs in a mixed-motive game and describes a method of play that would fractionally reciprocate cooperation (a kind of incremental Tit-for-Tat). He provides empirical evidence that suggests that such partial reciprocation has characterized U.S.-China relations since the 1950s. However, Goldstein does not define play in a game-theoretic sense; nor does he seek to determine empirically the payoff structure that underlies the behavior he describes. This is precisely what we endeavor to do here.

U.S.-China Relations

China officially opened its doors to the development of economic transactions with the Western world in 1978. Before then Chinese foreign trade was minimal and balanced. With the open-door policy, China's international trade soared, increasing by more than 500 percent

[26] See Michael McGinnis and John Williams, "Change and Stability in Superpower Rivalry," *American Political Science Review* 83 (December 1989).

[27] Goldstein and Freeman (fn. 24, 1990 and 1991); and Joshua S. Goldstein, "Great Power Cooperation under Conditions of Limited Reciprocity: From Empirical to Formal Analysis," *International Studies Quarterly* 39 (December 1995).

between 1978 and 1988.[28] By 1992 China's exports represented 18 percent of its gross national product with one-third of those exports absorbed by the United States. Moreover, the U.S. is the second largest single source of imports for China after Japan. In 1990 U.S. exports to China represented 23 percent of total Chinese imports. The United States has therefore played a crucial role in the development of China's economic overtures. After Hong Kong and Japan, the U.S. is the largest investor in China, and while the sums involved are still modest by international standards, U.S. interests are represented in critical ventures because the U.S., contrary to Japan, has been willing to transfer technology to China.[29]

International trade and investment create a situation of mutual economic dependency and provide the foundation of a common interest in economic growth. But trade relations can be interrupted or disadvantaged, as was made abundantly clear by the U.S. in the wake of the Tiananmen Square massacre. Such trade disruptions are costly to both parties. "I can state unequivocally that withdrawing or conditioning MFN (most favored nation) status (to China) would be a recipe for disaster for U.S. workers, consumers and employers," declared the president of the United States-China Business Council to the U.S. Congress in 1993, as he proceeded to comment on sectors of particular importance to the U.S.: "One need look no further than the airline industry to understand the critical importance the China market represents to U.S. industry. China represents one of the world's only growing major markets for aircraft and U.S. manufacturers account for 76 percent of all Chinese purchases." But just as surely, the Chinese depend on the United States for the success of their programs of economic reform. In their discussion of whether China can be influenced, Milholin and White point to the fact that most favored nation status is "a benefit worth billions to Chinese exporters."[30]

Trade clearly enhances the possible gain from mutual cooperation, and the drastic change in trade regime that the open-door policy meant for China suggests that the structure of the relationship between the U.S. and China is likely to have changed after 1979. As pointed out by O'Leary, diplomatic considerations that exclude reference to major economic developments cannot adequately explain changes in China's in-

[28] See James E. Shapiro et al., *Direct Investment and Joint Ventures in China: A Handbook for Corporate Negotiators* (New York: Quorum Books, 1991).

[29] Ibid.

[30] Gary Milholin and Gerard White, *Bombs from Beijing: A Report on China's Nuclear and Missile Exports*, report prepared for the House Foreign Affairs Committee, U.S. Congress, May 1991.

ternational relations.[31] In particular, one would expect to see stronger incentives to cooperate once trade relations intensify, although the circumstances that make defection attractive can still come into play: the Chinese can still gain from the sale of sensitive weaponry while the U.S. will still want to benefit from cordial relations with Taiwan.

Such circumstances suggest that a mixed-motive game reflects the reality of U.S.-Chinese relations.[32] But then, is it legitimate to represent the relative interests of these great powers identically before and after 1979? A glance at Sino-U.S. relations after Nixon's visit to China in March of 1972 leaves the observer groping for the [sound] foundations of the common interest. While Nixon's visit was hailed by some as marking "the emergence of a new positive era internationally," others feared that Nixon "would lose more in Japan and Russia than he would gain in Peking."[33] In the absence of formal diplomatic relations, liaison offices were established in 1973, but relations between the U.S. and China then slackened in subsequent years. In the words of Choudhury, many Asian observers simply regarded the Nixon-Kissinger policy as a trade-off of Vietnam for Taiwan.[34] The development of economic relations in the 1970s was also timid. Despite U.S. enthusiasm in the wake of the Canton Trade Fair of 1972, the two countries remained relatively insignificant trading partners in the pre-open-door era. Miller reports that by 1977 U.S. exports to the People's Republic of China represented a mere 0.1 percent of all exports, while the U.S. overall represented only 2.5 percent of China's imports and exports.[35]

On the face of it, it does not seem likely that U.S.-China interests would fit the mixed-motive game model in the decade of the 1970s. Casual observations of the facts are more suggestive of a zero-sum interaction or perhaps a game structure such as Deadlock, where defection by both parties actually provides higher payoff than mutual cooperation. But if this is the case, what behavior can be expected of these powers?[36]

[31] See Greg O'Leary, "China's Foreign Relations: The Reintegration of China into the World Economy," in Bill Brugger, ed., *China since the Gang of Four* (London: Crown Helm, 1980).

[32] Well-known mixed-motive 2 × 2 games include the Prisoner's Dilemma, Chicken, and the Stag Hunt.

[33] William G. Miller, *The People's Republic of China's United Front Tactics in the United States, 1972–1988* (Bakersfield, Calif.: Charles Schlacks, 1988).

[34] Golam W. Choudhury, *China in World Affairs: The Foreign Policy of the PRC since 1970* (Boulder, Colo.: Westview Press, 1982).

[35] Miller (fn. 33).

[36] Harrison Wagner, "The Theory of Games and the Problem of International Cooperation," *American Political Science Review* 77 (1983), argues that games such as Harmony and Deadlock should be considered before any mixed-motive game. And in a game such as Deadlock, the very meaning of the term *cooperate* becomes questionable: if both sides prefer the defect-defect outcome, is this not mutual accommodation?

And how can we interpret Goldstein and Freeman's finding of "limited reciprocity" in the relations between China and the U.S. in the 1970s and 1980s?[37] Goldstein and Freeman do not analyze Chinese and U.S. interaction in a game-theoretic context. By contrast our investigation of U.S-China relations begins with a game-theoretic model. But the incentive for cooperation depends on the structure of the game. Whether or not the U.S. and China can be described as playing a game of mixed motives is of paramount importance and must be addressed. If game theory is to have *empirical* relevance, the structure of payoffs must be identified if observed behavior is to be interpreted. The well-documented changes that occur in China-U.S. relations in 1972 and again in 1979 with the open-door policy provide us with an ideal test case of whether the data support a game-theoretic analysis of the facts.

III. The Formal Model

A First Formal Example

Let x_{us} represent the U.S. level of cooperation measured on a scale from 0 to 1. Here, 0 represents full cooperation while 1 represents full defection. Similarly $x_{ch} \in [0,1]$ represents China's level of cooperation. In this model, therefore, cooperation and defection are a question of degree. The [0,1] scale is, of course, arbitrary. In the empirical section, x_{us} and x_{ch} are levels of cooperation or defection as measured by coded WEIS data, using the same scale as Goldstein and Freeman.[38] *As a first example,* consider the linear payoff (for the U.S.)

$$U_{us}(x_{us}, x_{ch}) = x_{us} - 2x_{ch}. \tag{1a}$$

In this first example, we consider a typical Prisoner's Dilemma payoff for the United States. The success of countervailing strategies, as measured by their ability to promote a favorable outcome for both parties, depends on the payoff structure of the game. Games such as Harmony or Deadlock may require little strategic effort. But mixed-motive games do and often provide a fruitful environment for the pure countervailing approach.[39] Loosely speaking, success in establishing cooperation de-

[37] Goldstein and Freeman (fn. 24, 1990 and 1991).

[38] The data used by Goldstein and Freeman (fn. 24, 1990 and 1991) run on a scale from −6 to +6, where +6 represents full cooperation. In our notation, and if we did not aggregate events, this translates into $x_{us} = \frac{1}{2} - \frac{UC}{12}$, UC being the coded WEIS series representing U.S. behavior, and $x_{ch} = \frac{1}{2} - \frac{UC}{12}$, CU being the coded WEIS series representing Chinese behavior.

[39] Countervailing does not require a Prisoner's Dilemma structure in order to succeed. Indeed, our empirical estimates for the pre-open-door period (1972–78) point to a different game that, although

pends on the relative magnitudes of the incentives to cooperate or defect. If we refer to "fear" as the loss of payoff resulting from the other side's unilateral defection and to "temptation" as the gain from one's own unilateral defection, it turns out that the reciprocity inherent in a countervailing strategy is favored if both parties' fears exceed their temptations. Returning to our linear payoff function for the U.S., temptation as defined above is the rate of change of the U.S. payoff with an increase in its defection level, while fear is the absolute value of the change in payoff suffered by the U.S. when China increases its level of defection. So, Temptation = $\frac{\partial U_{us}}{\partial x_{us}} = 1$, and Fear = $\left|\frac{\partial U_{us}}{\partial x_{ch}}\right| = 2$.

We define the Chinese payoff similarly as a linear function of both its own level of cooperativeness and that of the U.S. And *for the purposes of our simple example,* we assume the Chinese payoff to have the symmetric parameters:

$$U_{ch}(x_{ch}, x_{us}) = x_{ch} - 2x_{us}. \tag{1b}$$

This game is a Prisoner's Dilemma defined on the unit square $[0,1] \times [0,1]$. In the repeated game the two sides simultaneously choose their level of cooperation/defection at each point in time t, thus determining an outcome $X^t = (x^t_{us}, x^t_{ch})$ in the square. If the players are concerned with the future consequences of their decisions, a natural representation of how they value their relationship is the discounted payoff stream V_{us} (for the U.S., with discount factor ω_{us}, at time t):

$$V_{us}(t) = \sum_{s=0}^{\infty} \omega_{us}{}^{s} \times U_{us}(x^{t+s}_{us}, x^{t+s}_{ch}). \tag{2}$$

In such a discounted repeated game, a strategy is a rule of play that determines each player's next move, given the prior history. So if the game begins at time t = 0 and is played periodically (at unit time intervals), $H^t = (X^0, X^1, \ldots, X^{t-1})$ is the history at time t, and each side bases its next decision on that history. Formally, we may write $x^t_{ch} = \psi_{ch}(H^t)$ to express China's move as a function of its strategy ψ_{ch} and, similarly, $x^t_{us} = \psi_{us}(H^t)$ for the U.S. However, assuming that a repeated game of

listed in exhaustive studies, has received little attention; see Steven J. Brams, *Theory of Moves* (Cambridge: Cambridge University Press, 1994). That game has a Nash equilibrium in which both sides defect, leaving one side (China) with its best outcome and the other (U.S.) with its next to worst. But countervailing succeeds in promoting a relatively cooperative outcome while providing an SGPE in that game.

great power relations has a well-defined starting date, t = 0 is at best unconvincing. We therefore ignore any specific starting date and focus on strategies that are *stationary* in the length of past history that are taken into account in the next decision.[40] However, we make the standard assumption that players seek to maximize their discounted payoff stream, given their expectation of future behavior. *As a first example,* let us examine the implications of the following strategy for China:[41]

$$x_{ch}^{t} = \psi_{ch}(H^{t}) = \frac{1}{2\omega_{us}} x_{us}^{t-1}. \tag{3a}$$

In this case the only relevant part of the past history in China's strategy would be the prior U.S. decision x_{us}^{t-1}. And as long as the discount factor ω_{us} is more than $\frac{1}{2}$, (3a) defines a true move x_{ch}^{t} (in the segment [0,1], for any move $x_{us}^{t-1} \in [0,1]$) and therefore describes a true strategy for China. Such a strategic choice by China consists in a dampened mirroring of the U.S. position and has remarkable implications *for the United States.* Plugging the Chinese response, as determined by (3a), into the U.S. objective (2) and using the explicit payoff (1a) yields by cancellation of matching terms

$$\begin{aligned} V_{us}(t) = {} & x_{us}^{t} - 2x_{ch}^{t} + \omega \times (x_{us}^{t+1} - 2x_{ch}^{t+1}) \\ & + \omega^{2} \times (x_{us}^{t+2} - 2x_{ch}^{t+2}) + \ldots = -2x_{ch}^{t}. \end{aligned} \tag{4}$$

This means that if China plays according to (3a), it is able to manipulate the value of the U.S. discounted payoff completely! It can do so *regardless of what the U.S. might play in response to the Chinese strategy at any given time.* It turns out that a whole class of strategies has as a consequence the full determination of the opponent's payoff. This class is made up of what we have referred to as *countervailing* strategies. The particular Chinese strategy described by (3a) is a countervailing strategy because it exactly compensates for any move by the U.S., so as to hold the U.S. to a particular discounted payoff that depends solely on the Chinese current position.[42] To see this, we will rewrite the U.S. payoff. By definition:

[40] When trying to define an initial date, the question is *how far back* one should consider past events and political regimes to be relevant to the day-to-day management of a relationship. By focusing on stationary strategies, we avoid this problem, although it is possible to reconstruct strategies that account for a starting date (see fn. 41).

[41] If an initial date t = 0 can be meaningfully defined, we need to describe x_{ch}^{0} separately. But as it turns out, an *arbitrary* x_{ch}^{0} will be optimal against the U.S. countervailing strategy we describe below.

[42] As the Chinese position x_{ch}^{t} evolves with time t, so does the U.S. discounted payoff in (4). Therefore, this Chinese countervailing strategy does not make the U.S. values *constant* in time. In fact, if the

$$V_{us}(t) = U_{us}(x^t_{us}, x^t_{ch}) + \omega_{us} V_{us}(t+1). \quad (5a)$$

Given China's strategy as expressed by (3a), and according to (4), it is therefore the case that

$$-2x^t_{ch} = U_{us}(x^t_{us}, x^t_{ch}) - 2\omega_{us} x^{t+1}_{ch} \quad (5b)$$

or

$$-2x^t_{ch} = (x^t_{us} - 2x^t_{ch}) - 2\omega_{us} x^{t+1}_{ch}. \quad (5c)$$

Equation (5b) highlights that at time t+1 China will choose x^{t+1}_{ch} in order to counterbalance whatever move x^t_{us} is chosen by the U.S. at date t: If x^t_{us} is high, reflecting an aggressive U.S. stand, the U.S. payoff $U_{us}(x^t_{us}, x^t_{ch}) = (x^t_{us} - 2x^t_{ch})$, at time t, will be correspondingly high, given the choice x^t_{ch} that China makes before it can observe x^t_{us}. Conversely, a low x^t_{us} reflecting a U.S. cooperative stand yields a lower U.S. payoff $U_{us}(x^t_{us}, x^t_{ch})$, given the same x^t_{ch}. Since the left-hand side of (5b) is not influenced by the U.S. choice, it is x^{t+1}_{ch} that must counterbalance the U.S. payoff on the right-hand side of (5b). The move x^{t+1}_{ch} will thus retaliate in proportion to the magnitude of U.S. aggression and it will similarly reciprocate U.S. cooperative initiatives. It is because such moves are tailored to the gains or losses made in payoff that we call such strategies *countervailing.*

When the U.S. payoff is made explicit in (5c), the countervailing condition becomes an equation that directly relates China's next move to the current situation. In fact, (5c) is equivalent to (3a) after canceling the x^t_{ch} terms, solving for x^{t+1}_{ch}, and shifting the time index. This is a general fact that we will exploit in the next two sections to construct a class of countervailing strategies far vaster than (3a).

If both players play countervailing strategies, the resulting pair of strategies happens to form a subgame-perfect equilibrium (SGPE).[43] To see why countervailing strategies form SGPEs, it is enough to observe the following: if China plays countervailing toward the United States, all possible responses by the U.S. are equivalent to the U.S., since China actually determines the U.S. payoff. This means that one countervailing choice by the U.S. is as good as any other, whatever the his-

Chinese are provoked into higher and higher x^t_{ch}'s by a tough U.S. stand, the discounted values in (4) *decrease* as a result.

[43] Recall that this means that play prescribed by such strategies maximizes discounted payoff *regardless of the prior developments of the game.*

tory of the game. If the U.S. also plays a countervailing strategy toward China, then the same reasoning applies to China. Countervailing strategies are therefore best responses to each other, whatever the past history of the game.

While providing an SGPE is an important check of rationality for strategies, equally important is the fact that countervailing strategies lead players to converge to a stable steady-state outcome, that is, a point in the outcome space $X^* = (x^*_{us}, x^*_{ch})$ that is stable given the strategies adopted by the players. In the terms of our example the steady-state outcome must satisfy for China:

$$x^*_{ch} = \frac{1}{2\omega_{us}} x^*_{us}. \tag{6a}$$

To find the steady-state outcome, it is, of course, necessary to specify the U.S. strategy. For the purposes of our example, assume that the U.S. plays the symmetric countervailing strategy:

$$x^t_{us} = \frac{1}{2\omega_{ch}} x^{t-1}_{ch}. \tag{3b}$$

At the steady state it must also be the case that

$$x^*_{us} = \frac{1}{2\omega_{ch}} x^*_{ch}. \tag{6b}$$

It follows that the steady-state outcome is $X^* = (x^*_{us}, x^*_{ch}) = (0,0)$, or full cooperation on both sides. That the steady state is a desirable outcome is important, but even more important is the fact that the strategies we define actually lead players to the steady state, provided both discount rates are strictly greater than $\frac{1}{2}$. Indeed, if we let $\rho = \frac{1}{2\omega_{us}} \times \frac{1}{2\omega_{ch}} < 1$, (3a) and (3b) result in (with $t = 2s$):

$$x^t_{ch} = \rho \times x^{t-2}_{ch} = \ldots = \rho^s \times x^{t-2s}_{ch} = \rho^s \times x^0_{ch}. \tag{7}$$

This shows that, starting from any $X^0 \in [0,1] \times [0,1]$, play as defined in our example will lead the players to full cooperation.

Of course, in response to (3a), the U.S. need not adopt the strategy (3b) or any other of the countervailing strategies we will study below. Indeed, since *any* U.S. strategy appears as good as (3b) at any point in time, due to the Chinese countervailing effect, why should the U.S. not adopt some other strategy? Although such a switch appears to have no

effect on the day-to-day appraisal provided by the discounted payoff (4), another U.S. choice might, for instance, lead to a steady state X^* other than full cooperation, one that could benefit the U.S. unilaterally! In fact, this is impossible: any other steady-state X^* would still satisfy (6a) according to the Chinese strategy, and the U.S. steady-state payoff would still sum up to $-2x^*_{ch}$ by the countervailing effect. So a x^*_{us} higher than zero would provoke a x^*_{ch} correspondingly higher than zero, and a U.S. discounted steady-state payoff strictly less than the zero of full cooperation. If the U.S. were to adopt some other strategy, it had better ensure that it fully reciprocates cooperation, given the Chinese countervailing choice (3a).

We will now progressively generalize our example.

A First Generalization

In the numerical example of the preceding section, we specified a strategy for China that maintained the U.S. discounted payoff at specific levels that depended solely on China's moves. In order to specify the nature and consequences of countervailing strategies in a more general context, the introduction of more general notation is in order. Still denoting by ψ_{us} and ψ_{ch} the strategies adopted by the two players, and by $\Psi = (\psi_{us}, \psi_{ch})$ the resulting pair, the U.S. discounted sum of payoffs in (2) can be legitimately written $V_{us}(\Psi|H^t)$, in order to stress its dependence upon the strategic choices and the prior history of the game H^t. With this notation, recall that according to the definition of the discounted payoff

$$V_{us}(\Psi|H^t) = U_{us}(x^t_{us}, x^t_{ch}) + \omega_{us} V_{us}(\Psi|H^{t+1}). \tag{8}$$

What is special about the class of strategies that we call countervailing is that they define a discounted payoff for each player that *depends only on the opponent's current move* and the history of the game. Therefore, if China plays countervailing it will always be the case that

$$V_{us}(\Psi|H^t) \equiv g_{us}(x^t_{ch}|H^t) \tag{9}$$

for some function g_{us} that will be further specified below. For instance, if the countervailing strategies under consideration specify moves that depend on the previous position only, excluding any history prior to that, then the U.S. discounted payoff depends exclusively on the current Chinese move and

$$V_{us}(\psi|H^t) \equiv g_{us}(x^t_{ch}) = U_{us}(x^t_{us}, x^t_{ch}) + \omega_{us} g_{us}(x^{t+1}_{ch}). \quad (10)$$

Rather than define a countervailing strategy directly, as we did in the above numerical example, we will now focus on both the particular form that can be given to the function g_{us} that results from countervailing play by China and the consequences of this choice for strategy. In our example we chose a strategy for China that determined the following g_{us} for the United States:

$$g_{us}(x^t_{ch}) = -2x^t_{ch}. \quad (11)$$

Consider the following, slightly more general g_{us}:

$$g_{us}(x^t_{ch}) = -\gamma_{ch} x^t_{ch} \quad (12)$$

where $\gamma_{ch} > 0$ is a parameter that will characterize the Chinese strategy. Keeping linear payoffs as in the example of the preceding section but generalizing these payoff structures to allow differences between the two countries we can write[44]

$$\begin{aligned} U_{us}(x_{us}, x_{ch}) &= x_{us} - b_{us} x_{ch} \\ U_{ch}(x_{ch}, x_{us}) &= x_{ch} - b_{ch} x_{us}. \end{aligned} \quad (13)$$

Applying the equality spelled out in (10) yields for the U.S.:

$$-\gamma_{us} x^t_{ch} = x^t_{us} - b_{us} x^t_{ch} - \gamma_{ch} \omega_{us} x^{t+1}_{ch} \quad (14a)$$

or

$$x^{t+1}_{ch} = \frac{1}{\gamma_{ch} \omega_{us}} x^t_{us} + \frac{\gamma_{ch} - b_{us}}{\gamma_{ch} \omega_{us}} x^t_{ch}. \quad (14b)$$

Equality (14b) makes explicit China's countervailing strategy by expressing its move at date t + 1 as a function of the previous position X^t.[45] Interestingly, China's move will depend on its own previous position x^t_{ch} and will therefore exhibit policy inertia as measured by the parameter $\frac{\gamma_{ch} - b_{us}}{\gamma_{ch} \omega_{us}}$. But China's strategy will also exhibit some degree

[44] Since we wish to discuss the Prisoner's Dilemma structure in this section, we must assume $b_{ch} > 0$ and $b_{us} > 0$. But it is possible to relax this restriction and to consider Deadlock-type games.

[45] In this approach, where the strategy is derived from the countervailing assumptions (9) and (12), we express x^{t+1}_{ch} as a function of X^t instead of x^t_{ch} as a function of X^{t-1} as in (3a). Of course, the two approaches are equivalent.

of reciprocity, since we assumed $\gamma_{ch} > 0$. Indeed, the Chinese will respond to the U.S. through a partial mirroring of the previous U.S. move. Thus a cooperative initiative on the part of the U.S. is reciprocated in part, but an aggressive move by the U.S. will elicit a lessening of cooperation or a strengthening of China's aggressiveness.

Of course, some restrictions on γ_{us} are necessary for (14b) to express a true strategy, that is, a rule that defines a *true* move in the decision space [0,1]. Here, assuming that X^t is in the unit square, x_{ch}^{t+1} is in [0,1] provided that

$$b_{us} \leq \gamma_{ch} \leq \frac{b_{us} - 1}{1 - \omega_{us}}. \tag{15}$$

For this condition on γ_{ch} to be realized, we must require $b_{us} > \frac{1}{\omega_{us}}$ so that (15) defines a nonempty range of countervailing strategies. In the payoff (13), where temptation has been normalized to 1, this means that *fear must outweigh temptation* for countervailing strategies to exist in a Prisoner's Dilemma.[46] Intuitively, when this condition fails, the temptation to defect overwhelms any attempt at stabilizing the game by a countervailing opponent. But when fear outweighs temptation, since the strategy defined in (14b) is countervailing, it forms an SGPE together with *any* similarly defined strategy by the United States. Moreover, if the inequalities in (15) are strict, one can show as in (7) that play according to this strategy pair converges to the full cooperative steady state (0,0).[47]

If China and the U.S. were to adopt such countervailing strategies and their payoffs could be safely approximated by linear functions, then we would seek in the data estimates of γ_{ch} and similarly defined γ_{us}, the parameters characteristic of each player's strategic behavior, as well as b_{us} and symmetrically defined b_{ch} to characterize the payoff structure of each of the players. But it is likely that true strategic interaction will take into account a number of past moves, by oneself and by one's opponent, in determining the appropriate next step. Thus we will introduce a final generalization that will be the model for our empirical estimates.

[46] Again we emphasize that payoff structures other than the Prisoner's Dilemma can be successfully dealt with using our countervailing approach. In other game structures the fear and temptation concepts do not necessarily apply and other relationships emerge.

[47] All other remarks made about the robustness of the countervailing strategy (3a) extend to (14b). In particular, no U.S. strategy could promote a better steady state for the U.S.

THE MODEL TO BE ESTIMATED

In the examples we have considered thus far, the strategies specify moves that depend only on the immediate past. We now want to build in a dependence on several past positions in order to capture possible decision lags that can differ from country to country, as well as a response that develops gradually over several periods. To examine such strategies we will proceed as we did in the example of the preceding section, first defining the level of discounted payoff to which each side holds its opponent, and from that inferring the particular strategic choice. Our previous example started with a function g_{us}, characteristic of China's countervailing strategy, that had the form $g_{us}(x^t_{ch}) = -\gamma_{ch}x^t_{ch}$. We now consider the more general function:

$$\begin{aligned} g_{us}(x^t_{ch}\,|H^t) &= g_{us}(x^t_{ch}, X^{t-1}, X^{t-2}, \ldots, X^{t-n}) \\ &= -\gamma_{ch}x^t_{ch}ch - \alpha^1_{ch}x^{t-1}_{us} - \beta^1_{ch}x^{t-1}_{ch} - \ldots \\ &\quad - \alpha^n_{ch}x^{t-n}_{us} - \beta^n_{ch}x^{t-n}_{ch} - k_{ch} \end{aligned} \tag{16}$$

where k_{ch}, the alphas, and the betas are parameters of arbitrary sign.[48] If China's strategy can achieve a countervailing effect as specified in (16), it must be the case that

$$\begin{aligned} &g_{us}(x^t_{ch}, X^{t-1}, X^{t-2}, \ldots, X^{t-n}) \\ &= x^t_{us} - b_{us}x^t_{ch} + \omega_{us}g_{us}(x^{t+1}_{ch}, X^t, X^{t-1}, \ldots, X^{t-n+1}). \end{aligned} \tag{17}$$

Combining (16) and (17), China's move of t+1 can be expressed as a function of past moves:[49]

$$\begin{aligned} \omega_{us}\gamma_{ch}x^{t+1}_{ch} &= (1 - \omega_{us}\alpha^1_{ch})x^t_{us} + (\gamma_{ch} - b_{us} - \omega_{us}\beta^1_{ch})x^t_{ch} \\ &\quad + (\alpha^1_{ch} - \omega_{us}\alpha^2_{ch})x^{t-1}_{us} + (\beta^1_{ch} - \omega_{us}\beta^2_{ch})x^{t-1}_{ch} \\ &\quad \cdots \\ &\quad + (\alpha^{n-1}_{ch} - \omega_{us}\alpha^n_{ch})x^{t-n+1}_{us} + (\beta^{n-1}_{ch} - \omega_{us}\beta^n_{ch})x^{t-n+1}_{ch} \\ &\quad + \alpha^n_{ch}x^{t-n}_{us} + \beta^n_{ch}x^{t-n}_{ch} + (1 - \omega_{us})k_{ch}. \end{aligned} \tag{18}$$

Before we turn to the statistical estimations, a few words must be said about the rationality of the kind of strategy described by relation-

[48] g_{us} could also be written as a sum. The particular form chosen is for notational convenience when writing out China's move at time (t + 1) as a function of the past history of play.

[49] We could have generalized the payoff function and the g_{us} and g_{ch} functions to include terms in $x_{us}x_{ch}$. Such cross terms allow for estimation of a wider variety of payoff functions. However, the inferred coefficient on the cross term for the payoff functions were not found to be significantly different from zero.

ship (18). First, it must be the case that (18) *actually defines* a strategy. In other words China's move x_{ch}^{t+1}, at time t + 1, must belong to the interval specified in the model. Although we used the unit interval [0,1] in the previous sections, it is entirely possible to use the range [−6,6] of the WEIS data. However, since +6 means the most cooperative attitude in WEIS, it is necessary to interpret x_{ch} as − CU (the score attributed to a Chinese move toward the U.S.), and x_{us} as − UC, in order to fit the payoff function (13). Then, the move x_{ch}^{t+1} specified by (18) is within the extended range [−6,6] provided that[50]

$$|1 - \omega_{us}\alpha_{ch}^{1}| + |\gamma_{ch} - b_{us} - \omega_{us}\beta_{ch}^{1}| + \sum_{i=1}^{n-1} |\alpha_{ch}^{i} - \omega_{us}\alpha_{ch}^{1+1}| + \sum_{i=1}^{n-1} |\beta_{ch}^{i} - \omega_{us}\beta_{ch}^{i+1}| + |\alpha_{ch}^{n}| + |\beta_{ch}^{t}| + (1 - \omega_{us})|k_{ch}| \div 6 \leq \omega_{us}\gamma_{ch}. \quad (19)$$

If (19) holds, (18) does define a strategy, and if the U.S. behaves similarly, then the strategy pair forms a SGPE since each side is countervailing the other. In fact, if the inequality in (19) is strict, it is also possible to infer the existence of a *dynamically stable* steady state whose position depends on the value of the constant k_{ch}.

What is remarkable about relationship (18) is that the parameters of China's strategy as well as the U.S. payoff parameter b_{us} are identifiable and can be inferred by a vector autoregression.[51] As outlined in Appendix 1, the estimated coefficients of the x_{us}^{t-s} terms allow for an inductive calculation of the α_{ch}^{s}'s and γ_{ch} (with appropriate confidence intervals). Then, the estimated coefficients of the x_{ch}^{t-s} terms yield inductively the β_{ch}^{s}'s and b_{us}. If both estimated b_{us} and b_{ch} are greater than 1, the game is a Prisoner's Dilemma; and the higher their magnitudes, the greater the players' incentives to establish tacit cooperation. If both coefficients turn out to be positive, but less than 1, the game is a Deadlock type. And when both coefficients fall into the negative, defection provides both the best individual and the best collective outcome.[52] But the case

[50] The denominator 6 in the k-term accounts for the range. The use of monthly data leads to aggregation of events and a widening of the range of the data. This would magnify the denominator 6 and would not affect the rest of our analysis. Of course, we still need to assume *some bounded* interval to guarantee the boundedness of payoffs required by the theorem in Appendix 2.

[51] An autoregression following (18) will identify the parameters of a countervailing strategy. As the main theorem in Appendix 2 demonstrates, strategies that form SGPEs are either countervailing or coordinated. If the players do not adopt a countervailing strategy but instead coordinate play, then the data should reveal punishing responses by each player. As mentioned above, we provide, in Appendix 3, a test of the hypothesis that China and the U.S. could be coordinating their moves rather than implementing countervailing strategies. The data do not support the hypothesis of coordination.

[52] In that case, countervailing would not succeed in providing an SGPE. A finding of negative b-coefficients for both sides would therefore refute the countervailing hypothesis.

where one side has a b-coefficient that is greater than one and the other has a negative b-coefficient provides an unusual and interesting situation that does not appear to have received much attention in the literature. Since that game appears as the outcome of the empirical study in the period 1972–78, we will discuss it more extensively in a concluding section.

IV. The Empirical Estimates

Data and Methodology

To estimate the parameters of relationship (18) for the U.S. and for China, we used coded monthly WEIS data, since that database extends through the 1980s, a period of critical importance for our purposes. The data were coded according to the Vincent scale, as in Goldstein and Freeman.[53] This scale has been criticized and improved,[54] but the improved scaling system did not alter the substantive results obtained with the Vincent scale to code the data. That events data are noisy is a well-known fact. We will therefore expect relatively low R-squared coefficients on our estimations. Furthermore, the ω-parameters that appear in the definition of the discounted objective (2) are not identifiable. However, it is standard practice to give them an arbitrary value such as $\omega_{us} = \omega_{ch} = 0.95$. In fact, setting ω-parameters within a realistic range (from about 0.9 to 1.0) consistently yields the same regression results.

Our empirical methodology entails estimating a full vector autoregression for each country and then trimming that regression by eliminating those right-hand side variables that do not contribute to the regression according to the Akaike information criterion. Before estimating the VAR, we needed to choose the maximum lag length. This is necessarily a somewhat arbitrary process, since, even if blocks of lags do not appear to have explanatory power when conducting a Granger causality test, it could be the case that lags from the more distant past are *individually* significant in a regression. Thus we somewhat arbitrarily considered a maximum lag of six months in all our regressions. This means that we assume China and the U.S. do not spread the response to any single move over more than six months, and that it does not take either of these states more than six months to respond to a cooperative

[53] Jack E. Vincent, *Project Theory* (Landam: University Press of America, 1979); Goldstein and Freeman (fn. 24, 1990).

[54] Joshua S. Goldstein, "A Conflict-Cooperation Scale for WEIS Events Data," *Journal of Conflict Resolution* 36, no. 2 (1992).

initiative or an act of aggression.[55] Such a maximum for the lagged variables on the right-hand side is consistent with the lag structure used in full-fledged VARs by Goldstein and Freeman.

Once our VARs are trimmed of statistically insignificant variables, the equations for each country end up having different variables on the right-hand side. It is therefore no longer appropriate to estimate these relations separately using ordinary least squares. Instead they must be estimated as a system using seemingly unrelated regression in order to ensure that the coefficients are unbiased.[56] Finally, since the time series we use are stationary, we can associate them without running the risk of spurious regression.[57] We now turn to the regression results.

The Regression Results

We divided our data set into two samples: 1972 to 1978 and 1979 to 1988. This partitioning is motivated by historical events, with 1972 marked by Nixon's visit to China and 1979 by the beginning of the open-door policy.[58] Estimation of relationship (18) for each country yielded strikingly different results across samples. The first set of equations reproduced below was estimated using data running from March 1972 to December 1978; standard errors are in parentheses:

$$\begin{aligned} x_{ch}^{t+1} &= \underset{(.49)}{2.07} - \underset{(.10)}{.26}\, x_{ch}^{t-1} + \underset{(.08)}{.12}\, x_{ch}^{t-4} + \underset{(.12)}{.33}\, x_{us}^{t-3} \\ x_{us}^{t+1} &= \underset{(.09)}{.32}\, x_{us}^{t-1} - \underset{(.08)}{.17}\, x_{ch}^{t-1} - \underset{(.07)}{.17}\, x_{ch}^{t-2} - \underset{(.07)}{.14}\, x_{ch}^{t-4} + \underset{(.066)}{.099}\, x_{ch}^{t-5} \end{aligned} \quad (20)$$

Estimates for the values of the b-coefficients (b_{us} and b_{ch}) in the payoff function (13) can be identified by induction on the estimated coefficients of system (20). Calculations are described in Appendix 1. For the U.S., we find[59]

[55] Such a restriction does not imply that the effects of an aggressive move will necessarily evaporate after six months. First, if the defector persists in an aggressive position, countervailing will continue and even deepen with time. And if the defector moves back to cooperation, the countervailing responses will only slowly reestablish cooperation.

[56] See William Greene, *Econometric Analysis* (New York: Macmillan, 1993).

[57] See Clive Granger and Paul Newbold, "Spurious Regressions in Econometrics," *Journal of Econometrics* 2 (1974).

[58] It turns out that when the equations estimated for the period 1972 to 1978 are used to explain subsequent years, a break in the data is apparent about 1979 and is confirmed by a formal Chow test.

[59] Even under the standard assumption of normal distribution for the coefficients estimated in (20), the resulting distribution of the b-coefficients is not normal since their derivation involves nonlinear transformations.

$\hat{b}_{us} = 4.23$ and $b_{us} \geq 1$ with 98 percent confidence.

Therefore, the United States payoff is characteristic of the Prisoner's Dilemma in its interaction with China during the 1970s.

Interestingly, our estimate for b_{ch} is *negative.* This means that China's payoff *increases* with U.S. defection. We find

$\hat{b}_{ch} = -2.11$ and $b_{ch} \leq -1$ with 99 percent confidence.

Nevertheless, countervailing strategies are appropriate courses of action for each player in this case, and they lead to a steady-state outcome. We will expand on this below. The steady state for the 1970s is about

$$x^*_{ch} = 1.56 \text{ and } x^*_{us} = -0.87 \tag{21}$$

showing steadier cooperation on the U.S. side than on the Chinese side.

Our estimates for the period running from January 1979 to December 1988 are as follows:

$$\begin{aligned} x_{ch}^{t+1} &= \underset{(.44)}{1.24} + \underset{(.08)}{.19}\, x_{ch}^{t-1} + \underset{(.08)}{.19}\, x_{us}^{t-3} \\ x_{us}^{t+1} &= \underset{(.45)}{-1.81} + \underset{(.09)}{.17}\, x_{ch}^{t-2}. \end{aligned} \tag{22}$$

In contrast to the pre-open-door estimates presented above, estimates for b_{us}, and b_{ch} reveal a Prisoner's Dilemma. For the U.S.

$\hat{b}_{us} = 5.42$ and $b_{us} \geq 1$ with 98 percent confidence

and for China

$\hat{b}_{ch} = 6.78$ and $b_{ch} \geq 1$ with 95 percent confidence.

The strategies adopted by each country now lead to a dynamically stable steady state

$$x^*_{ch} \simeq 0.0010 \text{ and } x^*_{us} \simeq -0.0015 \tag{23}$$

showing a softening of the Chinese attitude and a lessening of U.S. cooperation.

Interpreting the Results

The expected payoff structure found in the period 1972–78 is

$$U_{us} = x_{us} - 4.23x_{ch}$$
$$U_{ch} = x_{ch} + 2.11x_{us}. \qquad (24)$$

Such a payoff structure may reflect the internal political strife in China in the mid-1970s. As Miller points out, Chou En-Lai's declining health and death in 1976 enabled his opponents to discredit his initiatives.[60] Rhetoric from the People's Republic of China in the mid-1970s "attacked the U.S. superpower hegemony and imperialist exploitation" and such action could indeed mean political gain to signs of defection on the part of the United States. On the U.S. side, however, a more pragmatic attitude and an awareness of the economic implications of a Chinese opening lead to a positive valuation of Chinese cooperation.

In the one-shot game defined by (24), both sides benefit from unilateral defection. However, although Defect-Defect appears to be the (possibly ideologically motivated) favorite outcome for the Chinese, it is not so for the U.S. Symmetric moves toward cooperation would considerably improve the U.S. lot. In fact, on the scale [−6,6] that underlies the empirical findings, China's steady-state level of about 1.56 is not as uncooperative as one could fear. By contrast, the U.S. steady-state level of about −0.87 is rather accommodating. But a closer look at the U.S. strategy holds some explanation: in (20) the U.S. response exhibits policy inertia (with factor 0.32) and mirrors the Chinese stand with a predominantly *negative* coefficient (of about −0.38 overall). This means that the U.S. responds cooperatively to Chinese defection and uncooperatively to Chinese cooperation! The effect of this U.S. strategy on China's steady-state payoff is most interesting: given any strategy ψ_{ch} for China, whatever steady state (not necessarily unique) X^* may result must satisfy, according to the second line of (20):

$$x^*_{us} = .32\, x^*_{us} - .17\, x^*_{ch} - .17\, x^*_{ch} - .14\, x^*_{ch} + .099\, x^*_{ch}$$

or: $$x^*_{us} \simeq -0.56\, x^*_{ch} \qquad (25)$$

so that: $U_{ch}(X^*) = (1 - 2.11 \times 0.56)x^*_{ch} = -0.18\, x^*_{ch}$.

[60] Miller (fn. 33).

So China's steady-state payoff *increases* with China's cooperation in response to the U.S. strategy! Of course, it is China's own strategy choice, which, by providing a second steady-state equation similar to (25), finally settles a steady-state position. Since any two countervailing strategies form an SGPE (according to the theorem in Appendix 2), China's problem is only to find one that will promote the best steady state for itself. This would suggest looking for one that makes x^*_{ch} as small as possible, possibly equal to −6. However, a true countervailing strategy for China is incompatible with a steady state on or near the edge of its action space. Not surprisingly, the countervailing requirement restricts the region of possible steady states, just like any other SGPE concept does.

Apart from the above countervailing argument, a U.S. strategy that responds cooperatively to Chinese defection and uncooperatively to Chinese cooperation appears as a very rational choice in this game structure. To see this, consider the 2 × 2 matrix game of Figure 1 which has the same preference structure as that defined by (24).

The Nash equilibrium here is Defect-Defect. But if the U.S. could somehow *commit* to move (up) from DD to CD and (down) from CC to DC, the Chinese would find it in their best interest to move (left) from CD to CC, while attempting to remain at DD. Interestingly, this is precisely what Brams's Theory of Moves predicts in this game (called #35 in his nomenclature).[61] The outcome DC, which substantially improves the lot of the U.S., is what Brams calls a Non-Myopic Equilibrium in his theory, and it is achieved by what he would call here the U.S.'s "Moving Power." The countervailing approach is, of course, different. But it makes the above "commitment" an implicit and totally credible part of U.S. strategy.

As the 1970s drew to a close, the benefits of normalized relations with the U.S. took on more and more prominence until, finally, diplomatic relations were established, and China was given most favored nation status in early 1979. The game then turned into a typical Prisoner's Dilemma and resulted in a more symmetrical behavior. In (21), the United States now exhibits reactive behavior with a *positive* factor (0.17) and has no significant policy inertia, possibly indicating more flexibility toward China. The resulting steady state brings China to a more cooperative stance in the 1980s, while the U.S., still cooperative, slightly hardens its position relative to its posture for the 1970s. Over-

[61] Brams (fn. 39).

	China cooperates	China defects
U.S. cooperates	CC [3,1] (19.38,–18.66)	CD [1,2] (–31.38,–6.66)
U.S. defects	DC [4,3] (31.38,6.66)	DD [2,4] (–19.38,18.66)

FIGURE 1
THE U.S.-CHINA GAME[a]
(1972–78)

[a] Payoffs using (24) are in parentheses; ordinal utilities are in square brackets (from least preferred = 1 to most preferred = 4); DD is a Nash equilibrium; and DC is a Non-myopic equilibrium.

all, the outcome is mildly cooperative, a finding in line with that of Goldstein and Freeman.[62]

V. CONCLUSION

This paper develops a model of rational behavior that is found to be compatible with the evolution of China-U.S. relations from the early 1970s to the late 1980s. Our work is most closely related to that of Goldstein and Freeman and Goldstein.[63] In contrast to these authors, we work with a game-theoretic model that highlights a class of rational behavior that we call "countervailing" and that is found to be involved in any equilibrium strategy. Countervailing behavior broadens the scope of what was believed to be rational in a game-theoretic context, accommodating inertia, delays, and reactivity, and allowing for reciprocity in the face of a cooperative initiative and for measured retaliation in response to aggression. The terms of a countervailing strategy can be traced in the coefficients of a vector autoregression, and we find that the data on U.S.-China relations are compatible with rational behavior on both sides. Interestingly, the nature of these rational re-

[62] Goldstein and Freeman (fn. 24, 1990 and 1991).
[63] Goldstein and Freeman (fn. 21, 1991); and Goldstein (fn. 27, 1994).

sponses is found to change at a critical point in the history we examine. Indeed, once the terms of the strategic interaction between the U.S. and China have been identified, we are able to infer the payoffs that support the observed dynamic interaction. This is a new and critical step, since the identification of the payoff structure allows for proper interpretation of observed behavior. We find a break point in the data in 1979, when China opens its doors to the development of economic relations with the West. We also find that the payoff structure we infer for the period prior to 1979 conforms to a little-known mixed-motive game while the 1980s are characterized by the Prisoner's Dilemma. In the 1980s rational play finds expression in positive reactivity on the part of each country, while prior to 1979 rational behavior on the part of the U.S. takes the form of an inverse response to Chinese initiatives, a behavior that draws its rationality from its ability to move China to a more cooperative stance despite its Deadlock-type payoff structure.

Appendix 1: Inferring Game Parameters from Estimated Reactions

The empirical estimation in (20) can be written

$$x_{ch}^{t+1} = a_{ch}^{0}x_{us}^{t} + a_{ch}^{1}x_{us}^{t-1} + ..+ a_{ch}^{n}x_{us}^{t-n} + b_{ch}^{0}x_{ch}^{t} + b_{ch}^{1}x_{ch}^{t-1} + ..+ b_{ch}^{n}x_{ch}^{t-n} + c_{ch} \quad (26)$$

where the a_{ch}^{k} and b_{ch}^{k} (for k = 0, . . , n) are estimated statistically. But, by (18), if the countervailing model is believed, the following relationships hold:

$$\frac{\alpha_{ch}^{n}}{\omega_{us}\gamma_{ch}} = a_{ch}^{n} \text{ and, by backward induction (for k = 1, . . , n):}$$

$$\frac{\alpha_{ch}^{k}}{\omega_{us}\gamma_{ch}} = a_{ch}^{k} + \omega_{us}a_{ch}^{k+1} + ..+ \omega_{us}{}^{n-k}a_{ch}^{n} \text{ and, at k = 0:}$$

$$\frac{1}{\omega_{us}\gamma_{ch}} = a_{ch}^{0} + \omega_{us}a_{ch}^{1} + ..+ \omega_{us}{}^{n}a_{ch}^{n}. \quad (27)$$

So, if we let $A_{ch} = \sum_{k=0}^{n} \omega_{us}{}^{k+1}a_{ch}^{k}$, we may write $\gamma_{ch} = 1 \div A_{ch}$. A similar induction on the β_{ch}^{k} and b_{ch}^{k} coefficients, with the notation $B_{ch} = \sum_{k=0}^{n} \omega_{us}{}^{k+1}b_{ch}^{k}$, yields the important formula

$$b_{us} = (1 - B_{ch}) \div A_{ch} \tag{28}$$

and similarly for b_{ch}. Formula (28) can now be exploited to infer an expected value. It can also be used to obtain confidence intervals for b_{us} from the covariance matrix of the a^k_{ch} and b^k_{ch} coefficients. For instance, $b_{us} \geq 1$ is equivalent to the condition

$$\{(A_{ch} > 0) \text{ and } (A_{ch} + B_{ch} \leq 1)\} \text{ or}$$
$$\{(A_{ch} < 0) \text{ and } (A_{ch} + B_{ch} \geq 1)\}. \tag{29}$$

Furthermore, $\text{var}(A_{ch}) = J\Sigma J^T$, where Σ is the covariance matrix of the estimated coefficients and J is the (row) vector of the partial derivatives $\partial A_{chs}/\partial a^k_{ch} = \omega_{us}{}^{k+1}$. Using (29), we may write

$$\text{proba}(b_{us} \geq 1) \geq \text{proba}(A_{ch} > 0)$$
$$+ \text{proba}(A_{ch} + B_{ch} \leq 1) - 1 \tag{30}$$

and use the appropriate t-values on a one-tail test.

For instance, in our post-open-door estimates for b_{us}: $(\sigma_{A_{ch}})^2 = \text{var}(A_{ch}) = \omega_{us}{}^8(\sigma_{a^2_{ch}})^2 \simeq (0.95)^8(0.0768)^2 \simeq 0.00391$. Similarly, $(\sigma_{A_{ch}+B_{ch}})^2 = \text{var}(A_{ch} + B_{ch}) \simeq 0.0529$. Since we find means $\tilde{A}_{ch} \simeq 0.153$, and $\tilde{A}_{ch} + \tilde{B}_{ch} \simeq 0.323$, using the above standard deviations, $\tilde{A}_{ch} - k\sigma_{A_{ch}} > 0$ implies $k \simeq 2.4$, and $\tilde{A}_{ch} + \tilde{B}_{ch} + k\sigma_{A_{ch}+B_{ch}} \leq 1$ implies $k \simeq 2.94$. Using (30), probability $(b_{us} \geq 1) \geq 0.98$.

Appendix 2: A Characterization of SGPEs in Discounted Supergames

The following result provides a fairly general characterization of SGPEs in discounted repeated games. It is stated for two players and no state variable other than the players choices and prior history, but it can be generalized to finitely many players and stochastic games. For each player i, it highlights two functions g_i and π_i that will be interpreted below in terms of the countervailing and expectation effects that characterize SGPE strategies.

Theorem: In a discounted repeated game with bounded payoffs,[64] a

[64] This is the case when payoffs are continuous and defined over a compact decision space. Here, for instance, the linear payoffs (13) are continuous and the decision space $[0,1] \times [0,1]$ is compact (as closed and bounded in $\Re^2$).

strategy pair $\Psi = (\psi_i, \psi_j)$ is an SGPE if and only if there exist two real valued and bounded functions g_i and two nonnegative and bounded functions π_i such that for any decision pair X^t, and any prior history $H^t = \{X^0, X^1, .., X^{t-1}\}$ at any time $t \geq 0$, Ψ satisfies for i = 1,2:[65]

$$\begin{cases} g_i(x_j^t | H^t) - \pi_i(X^t | H^t) = U_i(X^t) + \omega_i g_i(\psi_j(H^{t+1}) | H^{t+1}) \\ \pi_i(\Psi(H^t) | H^t) = 0 \quad \text{with the notation } H^{t+1} = \{H^t, X^t\}. \end{cases} \tag{31}$$

Proof: First assume that Ψ is an SGPE. Then, for i = 1,2, any pair X^t, and any history H^t, let

$$\begin{aligned} & \gamma_i(X^t | H^t) = \sum_{s=0}^{\infty} \omega_i^s U_i(X^{t+s}) \\ \text{with} \quad & H^{t+s+1} = \{H^{t+s}, X^{t+s}\} \text{ for all } s \geq 0 \\ \text{and} \quad & X^{t+s} = \Psi(H^{t+s}) \text{ for all } s \geq 1. \end{aligned} \tag{32}$$

Clearly, γ_i is bounded since U_i is. So, we may define the bounded functions:

$$g_i(x_j^t | H^t) = \underset{x_i^t}{\text{supremum}}\ \gamma_i((x_i^t, x_j^t) | H^t)$$

$$\text{and} \quad \pi_i(X^t | H^t) = g_i(x_j^t | H^t) - \gamma_i(X^t | H^t) \geq 0. \tag{33}$$

We may rewrite, by definition of the discounted objective (32):

$$\gamma_i(X^t | H^t) = U_i(X^t) + \omega_i \gamma_i(X^{t+1} = \Psi(H^{t+1}) | H^{t+1}). \tag{34}$$

However, since Ψ is assumed to be an SGPE, $\psi_i(H^{t+1})$ is a *maximizand* with respect to x_i^{t+1} of $\gamma_i(x_i^{t+1}, \psi_j(H^{t+1}) | H^{t+1})$. Therefore $\gamma_i(\Psi(H^{t+1}) | H^{t+1}) = g_i(\psi_j(H^{t+1}) | H^{t+1})$, and $\pi_i(\Psi(H^{t+1}) | H^{t+1}) = 0$. Replacing this in (34), together with $\gamma_i(X^t | H^t)$ according to the second line of (33), yields (31).[66]

Conversely, if Ψ satisfies (31) for all X^t, H^t, and i = 1,2, with bounded g_i and π_i as described, let us consider (32) again. If we *further* assume that $X^{t+s} = \Psi(H^{t+s})$ for all $s \geq 0$, then, by assumption (31), using the notations (32), since all $\pi_i(X^{t+s} | H^{t+s}) = 0$, by the second line of (31):

[65] At time $t = 0$, $H^0 = \emptyset$. Also note that the history H^t is of length t and that no *stationarity* assumption is necessary for either the strategies or the g_i and π_i functions.

[66] Since H^t is as arbitrary as H^{t+1}, we may use either one in the equality of the second line of (31).

$$\begin{aligned} g_i(x_j^t|H^t) &= U_i(X^t) + \omega_i g_i(\psi_j(H^t) = x_j^{t+1}|H^{t+1}) \\ &= U_i(X^t) + \omega_i\{U_i(X^{t+1}) + \omega_i\{ \dots + \omega_i\{U_i(X^{t+s}) + \\ &\quad \omega_i g_i(x_j^{t+s+1}|H^{t+s+1})\}\dots\}\} \\ &= \sum_{s=0}^{\infty} \omega_i^s U_i(X^{t+s}). \end{aligned} \tag{35}$$

Now, consider an arbitrary strategy ϕ_i for Player i (instead of ψ_i), and let $\Phi = (\phi_i, \psi_j)$. We also assume play according to Φ from an arbitrary H^t at an arbitrary time $t \geq 0$. We let $G^t = H^t$, and denote for all $s \geq 0$: $Y^{t+s} = \Phi(G^{t+s})$ and $G^{t+s+1} = \{G^{t+s}, Y^{t+s}\}$. Clearly $y_j^t = x_j^t$, and since ψ_j satisfies the first line of (31) independently of the choice ϕ_i, by boundedness of g_i and π_i we have

$$\begin{aligned} g_i(x_j^t|H^t) &= g_i(y_j^t|G^t) \geq U_i(Y^t) + \omega_i g_i(y_j^{t+1}|G^{t+1}) \\ &\geq U_i(Y^t) = \omega_i\{U_i(Y^{t+1}) + \omega_i g_i(y_j^{t+2}|G^{t+2})\} \geq \dots \\ &\geq U_i(Y^t) + \omega_i\{U_i(Y^{t+1}) + \dots + \omega_i\{U_i(Y^{t+s}) \\ &\quad + \omega_i g_i(y_j^{t+s+1}|G^{t+s+1})\}\dots\} \\ &\geq \sum_{s=0}^{\infty} \omega_i^s U_i(Y^{t+s}). \end{aligned} \tag{36}$$

Moreover, at least one of these inequalities is strict if at any point $\pi_i(Y^{t+s}|G^{t+s}) > 0$. This shows that (35) is greater that (36) for any ϕ_i and that it is strictly so if, at any point, Φ fails to make $\pi_i = 0$. So, ψ_i is the best that Player i can do in response to ψ_j given the arbitrary H^t. Since this holds for i = 1,2, Ψ is an SGPE. *Q.E.D.*

APPENDIX 3: INTERPRETATION

The most interesting aspect of the first line of (31) is that, although it concerns i's payoffs, it can be viewed as a condition on i's opponent's strategy ψ_j. In order to decipher that condition, we will proceed in two parts.

First, let us assume that the two players play according to a strategy pair Ψ such that $\pi_i \equiv 0$. Player i's successive discounted payoff streams are then given by g_i, as in (35), and evolve according to

$$g_i(\psi_j(H^t)|H^t) = U_i(X^t) + \omega_i g_i(\psi_j(H^{t+1})|H^{t+1}). \tag{37}$$

This is what we call the countervailing effect of ψ_j on i's successive payoff streams: given the current history H^t, the left-hand side of (37) is

fully determined by g_i and ψ_j, and whatever i may gain or lose in the contemporary payoff $U_i(X^t)$, depending on what x_i^t he chooses, will be exactly compensated for in tomorrow's value of g_i. For instance, if g_i is a decreasing function of its argument ψ_j and is independent from its other argument H^t, an increase in U_i due to a variation of x_i^t requires a commensurate increase in $\psi_j(H^{t+1})$ so that the value of the left-hand side of (37) remains unchanged. Note that nothing specific is said here about whether i plays according to ψ_i.

However, and this is the second part, the function π_i need not satisfy the condition $\pi_i \equiv 0$. If it does not, then a deviation x_i^t from $\psi_i(H^t)$, while j plays $\psi_j(H^t)$, may entail a positive $\pi_i(x^t{}_i, \psi_j(H^t)|H^t)$. Since that quantity is subtracted from the left-hand side of (37) in the full equation (31), it means that $\psi_j(H^{t+1})$ must further compensate for that variation (by increasing if g_i is decreasing). Since we are interpreting equation (31) as a condition on ψ_j, the magnitude of $\pi_i(x_i^t, \psi_j(H^t)|H^t)$ must be viewed as a punishment that $\psi_j(H^{t+1})$ imposes on i for deviating from $\psi_i(H^t)$. Contrarily to the countervailing effect that involves only ψ_j, the values taken by π_i are inherently determined by ψ_i as well. In fact, the very equation $\pi_i(x_i^t, \psi_j(H^t)|H^t) = 0$ can be viewed as a restriction imposed by j on what x_i^t ought to be. For that reason, we call this the "expectation" effect: j is expecting what ψ_i player i will use and punishes him if he does not do so.

There is an entirely symmetric argument concerning the effects of ψ_i on j's payoffs. So i can also be seen as countervailing j and presuming of ψ_j if π_j is not identically nil. But as the theorem expresses, (31) must hold for both i = 1,2, with the *same* pair Ψ. This means that the equation $\pi_i(\Psi(H^t)|H^t) = 0$ must hold simultaneously for both i = 1,2. In other words, what i presumes must also be presumed by j and they must share what is usually referred to as a *common conjecture.*

Interestingly, as we demonstrate in the main text, this is not the case for the countervailing effect: each side may independently decide on how to countervail the other (in one of many ways) and still reach an SGPE. So the common conjecture requirement is weakened into the reciprocal assumption of countervailing behavior, meaning the choice of a countervailing strategy among a continuum of possibilities. This implies that players may shift from one countervailing strategy to another without explicitly renegotiating the terms of a common conjecture.

The trigger strategies that usually appear in the proofs of folk theorems exhibit a combination of countervailing and expectation effects. So our main theorem is neither a substitute for nor a consequence of the folk theorem. It simply focuses on another aspect of SGPE strategies

that is neither predicted nor prohibited by the folk theorem. Besides, since one of our main objectives is to *reveal* the payoff structures, the folk theorem analysis could only be conducted ex post.

Appendix 4: Testing for the Expectation Effect

Our game-theoretic interpretation of U.S.-China relations in the 1970s and 1980s relies on the assumption that each country adopts a countervailing strategy. We provide here empirical evidence to substantiate this hypothesis. If the strategy of either China or the U.S. involves an expectation effect that entails punishments (for example, $\pi_{us} > 0$), this will result in a *coordination* of their two strategies, and our main theorem predicts that the discounted payoff each player would enjoy under countervailing strategies must be adjusted from (37) to (31). In the expectation effect, each player's move of date t should be explained by the opponent's *concurrent* and past plays as well as its own past play. Assuming that such expectations are linear, the U.S. move of date t should be predicted by

$$\begin{aligned} x^t_{us} &= \delta_0 + \delta_1 x^t_{ch} + \eta_1 x^{t-1}_{ch} + \ldots + \eta_n x^{t-1}_{ch} x^{t-1}_{ch} + \nu_1 x^{t-1}_{us} + \ldots + \nu_n x^{t-n}_{us} \\ &= \zeta_{us}(x^t_{ch}, \ldots, x^{t-n}_{ch}, x^{t-1}_{us}, \ldots, x^{t-n}_{us}) \end{aligned} \tag{38}$$

and, if coordination is present, deviations of x^t_{us} from $\zeta_{us}(x^t_{ch}, \ldots, x^{t-n}_{ch}, x^{t-1}_{us}, \ldots, x^{t-n}_{us})$ should lead to imposition of a punishment by China. We thus test for coordination by estimating the deviation of x^t_{us} from $\zeta_{us}(x^t_{ch}, \ldots, x^{t-n}_{ch}, x^{t-1}_{us}, \ldots, x^{t-n}_{us})$ and by computing the residuals of a regression of the U.S. move of date t against China's concurrent and past six moves as well as the past six U.S. moves. The absolute value of these residuals

$$\varepsilon^t_{us} = |x^t_{us} - \zeta_{us}(x^t_{ch}, \ldots, x^{t-n}_{ch}, x^{t-1}_{us}, \ldots, x^{t-1}_{us})| \tag{39}$$

may be interpreted by China as a symptom of U.S. deviation from a common conjecture and may be the basis of a punishment π_{us}. If this is the case, we would expect a significant positive weight μ_{us} on those residuals in the determination of the U.S. discounted payoff.[67] A similarly defined ε^t_{ch} should have significant positive weight μ_{ch} in the determination of China's discounted payoff if a coordination effect can be

[67] In other words, we assume that $\pi_{us} = \mu_{us}|x^t_{us} - \zeta_{us}(x^t_{ch}, \ldots, x^{t-n}_{ch}, x^{t-1}_{us}, \ldots, x^{t-n}_{us})|$, with $\mu_{us} \geq 0$.

attributed to the U.S. strategy. Generalizing (17) to allow for such coordination yields

$$g_{us}(x^t_{ch}, X^{t-1}, X^{t-2}, \ldots, X^{t-n}) - \mu_{us}\varepsilon^t_{us}$$
$$= x^t_{us} - b_{us}x^t_{ch} + \omega_{us}g_{us}(x^{t+1}_{ch}, X^t, X^{t-1}, \ldots, X^{t-n+1}) \qquad (40)$$

and, as a straightforward generalization of (18), implies estimating

$$\begin{aligned}\omega_{us}\gamma_{ch}x^{t+1}_{ch} = {} & (1 - \omega_{us}\alpha^1_{ch})x^t_{us} + (\gamma^{ch} - b_{us} - \omega_{us}\beta^1_{ch})x^t_{ch} \\ & + (\alpha^1_{ch} - \omega_{us}\alpha^2_{ch})x^{t-1}_{us} + (\beta^1_{ch} - \omega_{us}\beta^2_{ch})x^{t-1}_{ch} \\ & + (\alpha^{n-1}_{ch} - \omega_{us}\alpha^n_{ch})x^{t-n+1}_{us} + (\beta^{n-1}_{ch} - \omega_{us}\beta^n_{ch})x^{t-n+1}_{ch} \\ & + \alpha^n_{ch}x^{t-n}_{us} + \beta^n_{ch}x^{t-n}_{ch} + (1 - \omega_{us})k_{ch} + \mu_{us}\varepsilon^t_{us}. \qquad (41)\end{aligned}$$

A μ_{us} that is not significantly different from zero is evidence against a coordination effect and in favor of pure countervailing play. Our system estimation for the period 1972 to 1979 yields

$$\begin{aligned} x^{t+1}_{ch} &= \underset{(.68)}{2.27} - \underset{(.10)}{.24}\, x^{t-1}_{ch} + \underset{(.09)}{.15}\, x^{t-4}_{ch} + \underset{(.12)}{.32}\, x^{t-3}_{us} - \underset{(.26)}{.18}\, \varepsilon^t_{us} \\ x^{t+1}_{us} &= \underset{(.09)}{.31}\, x^{t-1}_{us} - \underset{(.08)}{.18}\, x^{t-1}_{ch} - \underset{(.07)}{.22}\, x^{t-2}_{ch} - \underset{(.07)}{.15}\, x^{t-4}_{ch} \\ &\quad + \underset{(.065)}{.074}\, x^{t-5}_{ch} - \underset{(.11)}{.02}\, \varepsilon^t_{ch}. \qquad (42)\end{aligned}$$

For the period 1979 to 1988 our system estimates are as follows:

$$\begin{aligned} x^{t+1}_{ch} &= \underset{(.61)}{1.22} + \underset{(.08)}{.19}\, x^{t-1}_{ch} + \underset{(.08)}{.20}\, x^{t-3}_{us} + \underset{(.14)}{.02}\, \varepsilon^t_{us} \\ x^{t+1}_{us} &= \underset{(.66)}{-1.65} + \underset{(.09)}{.18}\, x^{t-2}_{ch} - \underset{(.17)}{.06}\, \varepsilon^t_{ch}. \qquad (43)\end{aligned}$$

In all cases the coefficients found on residuals ε^t_{us} and ε^t_{ch} are insignificant. We conclude therefore that the data do not support the hypothesis of coordination effects for either of the periods we consider.