ELSEVIER

Cognition 85 (2002) 1–36

COGNITION

www.elsevier.com/locate/cognit

What do you think I think you think?: Strategic reasoning in matrix games

Trey Hedden*, Jun Zhang*

Department of Psychology, Cognition and Perception Area, University of Michigan, 525 E. University Avenue, Ann Arbor, MI 48109-1109, USA

Received 20 February 2001; received in revised form 12 February 2002; accepted 13 March 2002

Abstract

In reasoning about strategic interpersonal situations, such as in playing games, an individual's representation of the situation often includes not only information about the goals and rules of the game, but also a mental model of other minds. Often such mental models involve a hierarchy of reflexive reasoning commonly employed in social situations ("What do you think I think you think…"), and may be related to the developmental notion of 'theory of mind'. In two experiments, the authors formally investigate such interpersonal recursive reasoning in college-age adults within the context of matrix games. Participants are asked to predict the moves of another player (experimenter's confederate) in a two-choice, sequential-move game that may terminate at various stages with different payoffs for each player. Participants are also asked to decide optimally on their own moves based on the prediction made. Errors concerning the prediction, or translation of those predictions into decisions about one's action, were recorded. Results demonstrate the existence of a "default" mental model about the other player in the game context that is dynamically modified as new evidence is accumulated. Predictions about this other player's behavior are, in general, used consistently in decision-making, though the opponent tends to be modeled, by default, to behave in a myopic fashion not anticipating the participant's own action.

Keywords: Recursive reasoning; Games; Mental models; Theory of mind; Backward induction

1. Introduction

It is widely held among cognitive psychologists that mental models (Johnson-Laird, 1983) are a means of understanding the manner in which people approach a wide range of problems involving decision-making and reasoning. Simply put, a person constructs a set

* Corresponding authors.
E-mail addresses: hedden@umich.edu (T. Hedden), junz@umich.edu (J. Zhang).

0010-0277/02/$ - see front matter
PII: S0010-0277(02)00054-9

of representations that model the possible states of the reasoning domain and may perform a set of operations on those representations. Depending upon the number of models involved and the order in which these models are constructed, particular types of problems may prove more difficult or more prone to errors (Evans, 1993; Johnson-Laird, 1983). It is possible that multiple models may be held simultaneously, although evidence from research in syllogistic reasoning suggests that this can be difficult for human reasoners (Johnson-Laird, 1983).

The utility of the mental models framework for understanding how people reason about a variety of situations suggests that this framework may yield insights into the manner in which mental entities are used to construct and construe the interpersonal world. There are many situations in which one may need to envision the mindset of another in order to reason about the potential consequences of a situation. A familiar example would be reasoning about a game of chess. What is the skill level of your opponent? Is she willing to make the knight for bishop trade? If not, will she anticipate the bishop fork that you can establish in two moves? These questions about the knowledge possessed by one's opponent and her current and future states of mind become highly relevant for winning the game, despite the constraints imposed by the game itself.

The deployment of mental models for decision-making in interpersonal settings has been experimentally studied in the paradigm of multi-player matrix games. Matrix games involve a set of payoffs to each player and a rule set by which players make decisions that determine the final payoffs. Such games have a long history in economic psychology, mathematical psychology, and the study of decision-making (Luce & Raiffa, 1957; von Neumann & Morgenstern, 1944). In the context of matrix games, mental models of others' minds may be used to determine a strategy, evaluate the actions of the opponent, and to modify the strategy based upon the match or mismatch between one's mental model of the opponent and the opponent's observed actions. Many such studies have been conducted, implicitly referencing the mental models constructed of others in the playing of matrix games (Colman, 1982). Following this tradition of experimental study of reasoning in two-person, two-strategy matrix games (Rapoport, Guyer, & Gordon, 1976), we explicitly probe mental models about an opponent's state of desire–belief in determining game behavior. We were interested in how the understanding of others' minds, sometimes referred to as a *theory of mind*, may serve as a basis for reasoning about interpersonal decisions. Further, we wanted to determine whether the construction and modification of mental models could be the mechanism by which such understanding of others' minds is translated into strategic actions.

Interest in mental models of others' minds has surged ever since the developmental literature on theory of mind (TOM) has investigated the understanding of others' minds in a variety of situations (Halford, 1993). TOM is an understanding of others as possessing mental lives, having desires, beliefs, and thoughts much like one's own (Estes, Wellman, & Woolley, 1989; Perner, 1991; Wellman, 1990, 1993; Woolley & Wellman, 1993). Even if development is understood, questions of the persistent use of TOM into adulthood are only beginning to be systematically investigated in experimental psychology. Several recent studies have explicitly looked at TOM in college students (Sabbagh & Taylor, 2000), older adults (Happe, Winner, & Brownell, 1998; Saltzman, Strauss, Hunter, & Archibald, 2000), and neurological patients (Happe, Malhi, & Checkly, 2001; Rowe,

Bullock, Polkey, & Morris, 2001; Stuss, Gallup, & Alexander, 2001; Surian & Siegal, 2001).

The major aim of the current investigation is to understand the structure and application of mental models in reasoning about the knowledge, desires, goals, and strategies of others. To this end, a set of games was developed that emulate a long tradition of experimental game research (e.g. Colman, 1982). The games used in this study are two-person asynchronous (or sequential) move games with finite stages (or horizon), also known as Stackelberg games in the literature on game theory (Osborne & Rubinstein, 1994). In short, a Stackelberg game is a two-player game in which one player, acting as the "leader", chooses an action from a set and then the other player, acting as the "follower" and informed of the leader's choice, chooses an action from a set, and so on, until the game terminates after finite moves and countermoves. It is said to be a game of complete and perfect information, in the sense that both players' payoffs and their previous moves are common knowledge. The Stackelberg game is naturally expressible as a decision tree (or an extensive form) where nodes represent decision points and paths represent the entire history of play (see Fig. 1). Because the game is of finite steps, one can work backwards (i.e. from the final choice point) to determine what the optimal strategy might be for each player when it is that player's turn to move, a process known as *backward induction*. Every finite extensive game with perfect information has a so-called sub-game perfect equilibrium, i.e. a combination of best-response strategies by the players from which they are unwilling to depart, a game theoretic result known as Kuhn's Theorem (Kuhn, 1953). Compared with the simultaneous-move (and therefore static) games used in much previous research (Colman, 1982; Rapoport et al., 1976), the games we have decided to use are based on a dynamic extension to the classical game theory that allows for sequential-move analysis (Aaftink, 1989; Brams, 1994; Brams & Wittman, 1981; Kilgour, 1984; Zagare, 1984). In these dynamic 2×2 games (Brams, 1994), the rules of play allow for a series of alternating moves and countermoves by the two players from the initial outcome selected. The concept of non-myopic equilibrium has been advanced, based on the idea that players can look ahead and anticipate where a process might conclude if the rules of

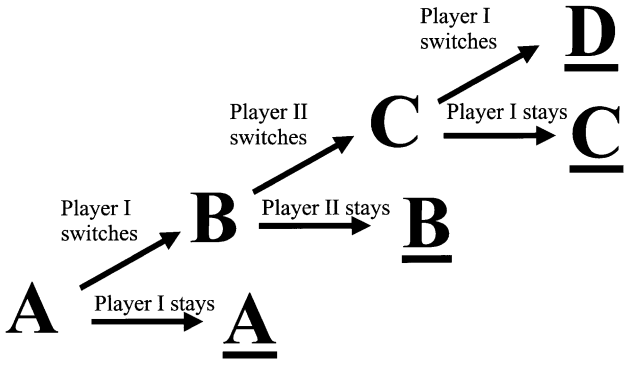

Fig. 1. A decision tree depicting the order of play and possible states within a game. Cells are indicated by letters. Possible terminal states are underlined.

the game allow a maximum of one, two, three or four sequential moves and countermoves altogether (called Rules I, II, III, and IV, respectively, in Zagare, 1984). There is evidence that even in certain static, simultaneous-move games, players tend to assume that the opponent has knowledge of what the player has chosen, as if the game were played out sequentially, a tactic termed the "Stackelberg heuristic" (Colman & Bacharach, 1997; Colman & Stirk, 1998). A systematic investigation of how people play these dynamic games will hopefully reveal the manner and extent in which mental models are engaged for interpersonal strategic decisions.

A simplified account of the games used in this study is as follows. First, all games for consideration in this study are of the 2×2 type, with two players each having two strategies that can result in a unique combination of payoffs for each player. These payoffs may always be ranked from worst to best by their numerical values, which, for simplicity, range from 1 to 4. Second, the game is understood to be a non-cooperative game with no communication allowed. Third, the initial state and the progression of the game are always specified in advance, where players take turns deciding to switch or stay. A game ends in a maximum of three such steps or decisions (i.e. played under Rule III of Zagare, 1984). Therefore, the first player can always be considered to have movement power, or the ability to control the final outcome more fully than the second player (see Fig. 1). In combination, these simplifying modifications enable the unambiguous delineation of different strategies represented by mental models, referred to as *orders of reasoning*. They also provide a framework in which the role of learning can be investigated.

A hierarchical classification similar to that used by Perner and Wimmer (1985) to describe the development of TOM was adopted to illustrate the structure of mental models in interpersonal reasoning. A zeroth-order model refers to the level of analysis that considers only one's own desires and the state of the world, having no understanding of the desires, beliefs, or thoughts of others. A first-order model is characterized by the understanding of oneself and others as having desires, beliefs and thoughts that would influence their behavior. A second-order model is characterized by the recognition that others also hold beliefs about the desires and beliefs of one's self, and therefore may act predictively for their own good. The predictions of oneself by others must be taken into account when predicting those others' behavior. Likewise, a third-order model can be defined as the mental model of others who hold a second-order model and therefore have beliefs about one's own beliefs about their beliefs, etc. From the perspective of mental models, the developmental hierarchy of TOM may be explained by the expansion of the ability to represent increasingly complex models. The reversion to lower-order reasoning is explained by the adoption of a mental model constructed from incomplete information or that focuses on an initial, but faulty, representation (Legrenzi & Girotto, 1996; Legrenzi, Girotto, & Johnson-Laird, 1993).

To identify the use of orders of reasoning based on strategic mental models of others in these games in separation from the "rational", or optimal, application of such reasoning for one's own decision-making, participants (always assigned to the role of the first player) were asked to first make a prediction about what the other player would do. After making this prediction, they were asked to actually make a decision. This methodology allows a dissociation between the mental model of the opponent's knowledge and strategy and the ability to make rational decisions (to stay or to switch) based on the output of that model.

The pivotal assumption in both classical game theory and the theory of moves is that players are "rational"; however, this term is reserved in this paper to refer to the optimal consistency between the mental model a participant holds (a "prediction") and the action the participant selects (a "decision"). The design provides a measurement of rationality in the play of the game, as well as a distinct assessment of the mental model used by a player without respect to the player's rational application of the model. By analyzing only the prediction and the decision made at the very first choice point (i.e. the root of a game tree) for Player I in each game, it is possible to classify the player as using either first-order or second-order reasoning, and to classify the player's decision as rational or not.

Two experiments were conducted in order to assess the applicability of a mental models approach to strategic reasoning about others in a matrix game setting. We were interested first in determining the type and structure of the mental model of the opponent employed by participants at the outset of the game, and second, whether such mental models are modifiable based on continued interaction with the opponent. Through observation of repeated predictions and decisions, it was possible to determine whether participants were truly modeling the mental states of their opponents as opposed to engaging in a backward induction strategy appropriate for any sequential decision task.

2. Experiment 1

In the first experiment, an opponent acts either as a zeroth-order reasoner or as a first-order reasoner, and his behavior either can be consistent with this assigned order of reasoning for all trials, or can switch the order of reasoning between the test blocks. Given the payoff structure of the games in the present design, and the fact that participants are instructed to play in a non-cooperative manner and to attempt only to maximize their own payoffs in each game, the following hypotheses are investigated. (1) In making predictions about the behavior of the opponent, participants will adopt at the outset a default mental model on which to base their reasoning. This default will be grounded in either the use of first-order or second-order reasoning, although no specification as to which would be chosen as the default was advanced. Such a default strategy leads to the specific hypothesis that at the beginning of the first test block, no differences should be observed between experimental conditions in which the opponent (Player II) behaves differently by adopting different (zeroth- and first-) orders of reasoning. Further, the observed score for predicting the opponent's behavior should be significantly different from chance. (2) Over a period of trials, players will deviate from this default model as they adapt to the behavior of the opponent and modify their mental model of the opponent as counter-evidence is presented (Gopnik & Wellman, 1994). Thus, players whose opponent is a zeroth-order reasoner will tend to stabilize at a first-order prediction, while those with a first-order opponent will tend to stabilize at a second-order prediction. (3) Rationality errors should remain relatively constant across experimental conditions, even when the opponent switches the order of reasoning between blocks of trials. Such constancy is indicative that decision-making based on one's predictions can be dissociated from the building of a mental model of the opponent. That is, despite the nature of the mental

model, which should change across experimental conditions, the decision-making process should remain unaffected.

Taken together, these three hypotheses should provide evidence for the use of mental models in reasoning within the game setting, and should speak to the structure of those mental models related to the depth of recursion intrinsic to reasoning about others' minds, which we will refer to, for the sake of brevity and clarity, as zeroth-order, first-order, and second-order TOM reasoning.

2.1. Methods

2.1.1. Participants

Participants were 70 (35 males, 35 females) undergraduate students at the University of Michigan who received class credit for their participation. The average age of the participants was 19.0 years. All participants gave informed consent prior to admission into the study and were properly debriefed (including the nature of the computerized opponent) at the conclusion of the experiment.

2.1.2. Materials

2.1.2.1. The game A sequential-move 2×2 matrix game was developed in which Player I and Player II take turns making choices that lead to the outcome of each game. The outcome of a game can be one of four cells (labeled A, B, C, and D), with each cell containing separate payoffs for Player I and Player II (see Fig. 2 and Appendix A). The players make decisions (also termed a "move") sequentially, so the game may progress as indicated by the arrow. Each game begins in cell A, and Player I may decide to stay in cell A or to switch to cell B. If Player I decides to switch to cell B, it then becomes Player II's turn, who must decide whether to stay in cell B or to switch to cell C. If Player II decides to switch, the turn then passes back to Player I, who makes the final decision of whether to stay in cell C or to switch to cell D. In either case, the game ends after this final decision. However, if at any stage either player decides to stay in the current cell during their turn, the game ends immediately. Fig. 1 depicts a decision tree with all possible outcomes, and Appendix A provides a sample game.

Both players receive their respective payoff indicated in the cell in which the game has ended. The payoffs in each cell consist of a pair of integers ranging from 1 to 4, often different for each player, to indicate the consequences to each player if the game ends in that cell. The payoff numbers correspond to the preference rankings of each player, with 4 indicating the most preferred outcome for a player and 1 indicating the least preferred outcome. All games consist of a unique "payoff structure", or a combination of possible payoffs in each cell for the players. This payoff structure is common knowledge to both players. Given the rules of this sequential-move game, the payoff structure is essential for optimal decision-making by the players. Indeed, the design of the current experiments concentrates on the pattern of these payoff structures that affect the reasoning and decision-making of the players.

2.1.2.2. Payoff structure Restricted to ordinal rankings of four outcomes for either player,

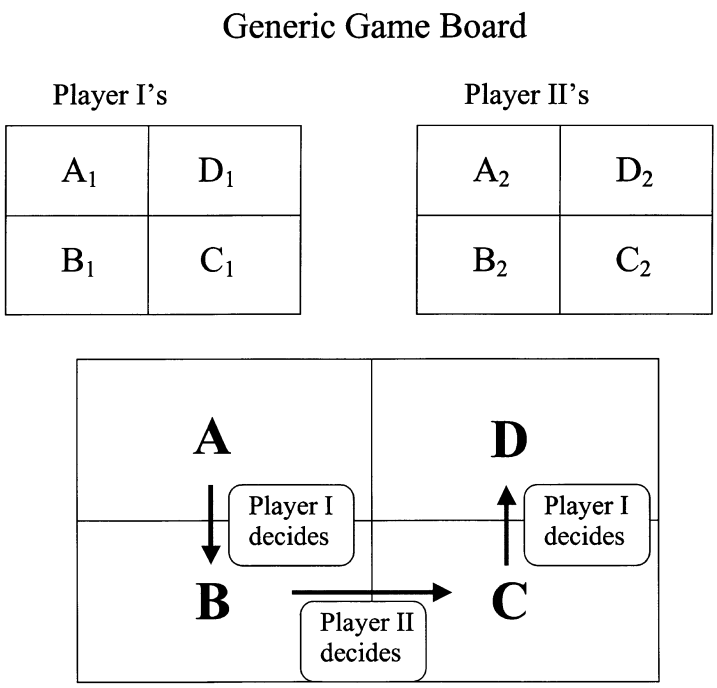

Fig. 2. Generic game structure, with arrows indicating progression of play.

there are a total of $4! \times 4! = 24 \times 24 = 576$ possible combinations of payoffs in 2×2 games. Of these, any game containing a 1 or a 4 in cell A for Player I is fundamentally uninteresting in this paradigm, since the worst or best outcome for that player is present at the outset of the game, rendering the decision to stay or to switch obvious. Similarly, any game containing a 1 or 4 in cell B for Player II provides the worst or best outcome at the first choice point for that player, and is therefore uninteresting for present purposes. Excluding all such games yields 144 possible payoff structures to be explored further.

There are 48 "trivial" games, in which the payoffs for Player II in cells C and D are both higher than the payoff in cell B (hence Player II will decide to switch regardless of other payoffs), or in which Player II's payoffs are lower in cells C and D than in cell B (Player II will stay in cell B regardless). These games are slightly different from the cases in which Player II has a 1 or 4 in cell B, due to the fact that in the latter cases Player II need not even examine the payoffs in cells C and D to make a decision. Such trivial games may be useful for familiarizing players with the rules and procedures of this non-cooperative, sequential-move game, and are thus included in the Training Block of the design. The remaining 96 games are interesting due to the necessity of engaging the players in reflexive reasoning about mental states.

Note that in this sequential-move game, Player I has control over the first move, and Player II controls the second move. If the game progresses to cell C, the choice for Player I is fairly obvious: stay if Player I's payoff is greater in cell C than in D, or switch if vice

versa. The more complex aspect lies in Player I's decision at the first choice point. Ideally, this decision depends upon what Player II will do if the game progresses to cell B. If Player II simply compares his payoffs in cells B and C to decide whether to stay or to switch, we term him a "myopic" player, because he engages in zeroth-order reasoning. If, on the other hand, Player II takes into account what Player I might do if the game reaches cell C, then we term Player II a "predictive" player, since he is engaging in first-order reasoning by forecasting Player I's behavior. Therefore, at the first choice point (cell A) for Player I, the optimal decision is connected to the ability to predict Player II's decision (at cell B) based upon reasoning about Player II's likely strategy. If Player I is equipped only with first-order TOM reasoning, he could only conceive of and correctly predict a myopic opponent (Player II) who uses zeroth-order reasoning. For Player I to be able to correctly predict a predictive decision on the part of Player II, it is required that Player I invoke second-order TOM reasoning. Furthermore, the rational actions of Player I depend upon whether he is faced with a myopic opponent or a predictive opponent.

Viewed in this way, the 144 matrix games can be classified according to Player I's decision at the first choice point (cell A), and Player II's decision (cell B). Since Player II may act in either a myopic or a predictive manner, and Player I may either predict Player II to be myopic or predictive and act accordingly, there are four possible combinations of actions associated with a game, which we have represented by a quadruplet (m, p, M, P). Here the variables can take only the binary values of 0 (representing a "stay" decision) or 1 (representing a "switch" decision) for a given game. The positions in the quadruplet occupied by m and p indicate the action of a myopic or a predictive Player II, respectively, while M and P indicate the rational action at the first choice point of a Player I who predicts that Player II is myopic or predictive. The quadruplet classifications of a subset of the games considered for use in this experiment are listed in Appendix B.

There are 96 games (including the trivial games) in which m and p are equal (and therefore M and P are also equal, as they follow from m and p for a rational player). These games are non-diagnostic with respect to the order of TOM reasoning. The quadruplets for such games are: (1,1,1,1), (1,1,0,0), (0,0,1,1), and (0,0,0,0). Precisely because they are non-diagnostic, and therefore yield no information about which order of TOM reasoning the opponent may be using, these games are useful for training purposes and as catch trials to test for heuristic strategies.

The remainder of the 48 games are diagnostic because Player I's first-order and second-order reasoning lead to diametrically opposed predictions and decisions. These games have been grouped by the pattern of their quadruplets, forming classes of strategically equivalent games, which have been labeled by "type" in Appendix B. In order to avoid the use of a simple heuristic strategy on the part of Player I (e.g. move on a 2 and stay on a 3 in cell A), these games were further divided according to whether Player I's payoff in cell A is a 2 or a 3. We call these 2-starting games and 3-starting games.

Within the context of the game design, it becomes possible to formally specify the hypothetical structure of mental models based on the various orders of reasoning. The payoffs of each player must be compared in a given order to make predictions and decisions using each order of reasoning. Representations of the structure of mental models based on zeroth-order, first-order, and second-order reasoning are displayed in Figs. 3 and 4. At each decision point, a comparison (represented by a colon) between one's payoff in

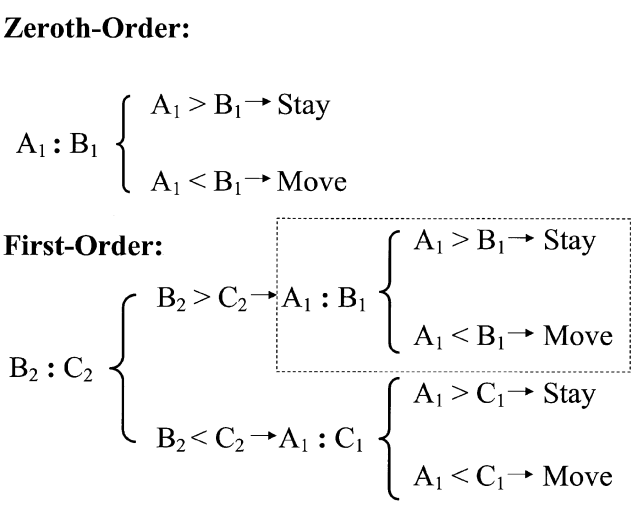

Fig. 3. Abstract structural representation of hierarchical mental models based on zeroth-order, and first-order reasoning. Colons represent comparison operations, arrows indicate the output of operations, and brackets indicate alternate states of affairs. Dashed boxes outline subroutines in each subsequent level of the hierarchy.

the various cells of the game must occur. An optimal decision to stay or to switch can then be made on the basis of that comparison. However, at the first-order of reasoning and higher, the optimal decision can only be made after performing additional comparison operations. Importantly, the necessary operations involved in each higher-order model subsume the lower-order models (as indicated by the dashed boxes in the figures). Thus, these models are fully hierarchical in structure.

2.1.2.3. Experimental blocks Three blocks of trials were assembled. The first block is the Training Block, consisting of 16–24 games. These games were drawn from the set of trivial games described above, plus two games from the set of initially excluded games containing a 1 in cell B for Player II (included for balancing considerations). Such trivial games are ideal for acquainting the participants (Player I) with the goal and rules of the game, while yielding no discriminative information about the TOM reasoning used by the opponent, since both a myopic and a predictive opponent (Player II) will behave identically. This allows the default model hypothesis to be tested after training, without unduly biasing the participants towards first-order or second-order reasoning. Hence, games in the Training Block should have no effect on the participant's initial model of the opponent. The games in the Training Block are evenly balanced with respect to the number of 2-starting and 3-starting games. Performed optimally, these games are also balanced for the number of stay and switch predictions about Player II and for the decisions of Player I. A training criterion that allowed no more than three rationality errors (defined below) in the last eight successive games of training was used. If the participant passed this criterion in the first 16 training games presented, he or she was given a brief recess before beginning Test Block 1. If the criterion was not passed after the

first 16 games, an additional four training games were administered. If the participant still did not pass the criterion, four more training games were presented, for a total of eight additional games (24 total training games).

Test Block 1 consisted of 16 3-starting games that are diagnostic for TOM reasoning; that is, the predictions for a zeroth-order ("myopic") opponent and a first-order ("predictive") opponent are exactly the opposite for each game in this block. These games were doubly balanced according to both the number of stay/switch predictions of Player II (for either order of TOM reasoning) and the number of stay/switch decisions by Player I, if performed optimally. Four non-diagnostic 3-starting games were selected as catch trials. Games for the catch trials have the same payoff structures as those used for the Training Block, and are hence uninformative (non-distinguishing) of the opponent's TOM order. These trials were used as a check against possible recognition of payoff structures merely as visual patterns by the participant. Such visual recognition could lead to heuristic strategies in decision-making. Due to the fact that in the catch trials, a myopic or a predictive opponent's reasoning will lead to the same action, these games should not affect the manner in which participants learn and employ an order of reasoning.

Test Block 2 consisted of 16 2-starting games that are diagnostic for TOM reasoning. These games are balanced according to the prediction that Player II will stay/switch when acting either myopically or predictively. However, a larger number of 2-starting games lead to a switch decision for Player I regardless of Player II's behavior. Such games have quadruplets of (0,1,1,1) and (1,0,1,1). Due to this fact, the games in Test Block 2 are

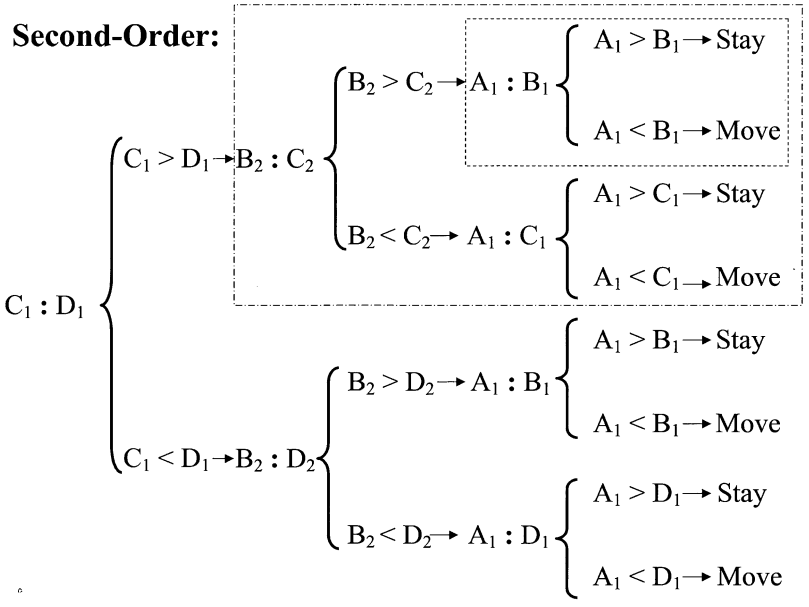

Fig. 4. Structural representation of a hierarchical mental model based on second-order reasoning. Colons represent comparison operations, arrows indicate the output of operations, and brackets indicate alternate states of affairs. Dashed boxes outline subroutines from prior levels of the hierarchy.

somewhat unbalanced (and unable to be balanced), favoring a switch decision at Player I's first choice point in a 3:1 ratio. Therefore, the four catch trials, all requiring a stay decision for Player I, were selected in order to more closely balance the number of stay/switch decisions. An optimal Player I using either first-order or second-order reasoning will decide to switch in 12 out of 20 games.

The rationale for dividing the 2-starting and 3-starting games across test blocks is to discourage a simple heuristic based on the participant's risk attitude. Such a heuristic would lead a participant (Player I) to always decide to switch when cell A contains a 2 for Player I, as if it is more probable that a higher outcome would be available, and to always decide to stay if cell A contains a 3 for Player I, as if it is unlikely that a better outcome could be attained. Notice that this strategy is heuristic in that it eliminates the need to analyze the entire payoff structure of each game (for oneself and for the opponent), and indeed, eliminates the necessity of making accurate predictions about the opponent. Such a heuristic strategy is discouraged by blocking the games according to whether Player I starts with a 2 or a 3 in cell A.

In both test blocks, the diagnostic games can be further grouped into four game types. All games within a type are strategically equivalent and yield the same predictions and decisions with respect to TOM reasoning. Thus, in addition to measuring how a mental model can be constructed and modified based on the opponent's order of TOM reasoning, the design allows an investigation of the effects that payoff structures may have on the use of TOM reasoning.

A fixed order of game presentation was adopted in each block to facilitate the analysis of learning over time and of the effect of game type. Each game is trial-unique, that is, presented only once during the entire sequence of games during an experiment (consisting of 16–24 games in the Training Block, 20 games in Test Block 1, and 20 games in Test Block 2).

2.1.2.4. Scoring Predictions of the opponent's decision (stay or switch) were scored against how a myopic or a predictive opponent would have behaved in each game. A prediction that the opponent will stay/switch according to zeroth-order (myopic) reasoning was scored as a 0, while a prediction consistent with a first-order (predictive) opponent was scored as a 1. Thus, as mean prediction scores approach 0, participants tend to use first-order TOM reasoning. As mean prediction scores approach 1, participants tend to adopt second-order reasoning. However, in the Training Block and for catch trials, the predictions were scored simply as either correct or incorrect. This is due to the fact that both TOM strategies lead to the same prediction in these games.

Based on his or her prediction of the opponent's move, the participant has an optimal (rational) decision as to whether to stay or to switch. The participant's actual choice was recorded and compared to the optimal decision given his/her prediction for that game. Any inconsistency was scored as a rationality error. Decisions were not scored separately from rationality errors, as the decisions can be completely reconstructed through a comparison of the predictions made and the rationality errors (if any), and hence are redundant. A rationality error occurs only when a participant fails to make the optimal decision given his or her prediction of Player II, whatever that prediction might be.

2.1.3. Design

Participants, always assigned the role of Player I, were randomly assigned to experimental conditions in a design with two factors. These conditions differed only in the order of reasoning employed by the opponent (Player II). The opponent would employ either zeroth-order (myopic) reasoning or first-order (predictive) reasoning in Test Block 1. Additionally, in Test Block 2, the opponent would either continue using the same order of reasoning as in Test Block 1, or switch to the other order of reasoning. Thus, for Test Block 1, there is a single factor with two levels (opponent strategy), while for Test Block 2 there are two factors (opponent strategy by strategy switch) with two levels each.

2.1.4. Procedure

A participant was first introduced to a confederate, ostensibly another student participating in the study. The participant was led to believe that the two of them would be playing a simple matrix game via two separate computer terminals connected to a common network. The participant watched as the confederate was led into a room designated for Player II and was then led into a room designated for Player I. The participant was instructed that the game was not intended to be cooperative and that the goal should be to earn as many points as possible in each game without regard for the number of points earned by the opponent. Points were determined by the final payoff to each player in the game-ending cell. After reading the instructions and playing an example game, the participant began the Training Block. All games were presented on a Power Macintosh 9500 using an AppleVision 1710AV display monitor. Responses were made via mouse clicks to on-screen prompts and recorded electronically.

Despite appearances, participants were actually playing against a computerized opponent programmed to use either zeroth-order or first-order reasoning. Random delays were built into the program at key points so that participants would believe that they were actually playing with an attentive confederate. Only 7.1% of participants indicated a suspicion that they were playing with the computer rather than a person on an exit questionnaire.

After the criterion had been achieved in the Training Block, or after all 24 games in the Training Block had been presented, Test Block 1 and Test Block 2 were presented. A brief intermission was given between blocks. At each game, participants were prompted in the following order to (1) make a prediction of the opponent's choice (to stay or to switch) at cell B, and (2) decide whether to stay or to switch at cell A. The first question provided a direct measure of the participant's mental model of the opponent, while the second question measured rational decision-making. These predictions and decisions for each game were subjected to further analysis.

After finishing all the games, forward and backward digit span measures of working memory were administered. An exit questionnaire consisting of open-ended questions about their own strategies, the opponent's strategies, and the nature of the experiment was administered. Participants were then thanked and debriefed, including information about the nature of the computerized opponent and the necessity of this deception.

2.2. Results and discussion

2.2.1. Training Block

In the Training Block, prediction errors and rationality errors were scored in order to assess basic understanding of the game, namely, the non-cooperative, complete-information, and sequential-move characteristics. Because participants may differ in their speed of acquaintance with the rules of the game, the last 16 games presented to a participant were used for this assessment. Participants with greater than five rationality errors were excluded from further analyses. Due to the simplicity of the predictions in the training Block (the opponent's payoffs in cells C and D are either both higher or both lower than his payoff in cell B), participants with two or more prediction errors were also excluded for failure to understand the game. This left a total of 52 participants to be included in further analyses.

Games in the Training Block were designed to familiarize the participants with the rules and procedures of the game situation. For the last 16 training games presented to a player, the mean number of errors made in predicting the opponent (whether the opponent will stay or switch in cell B) and in making a rational decision (translating one's prediction of the opponent into a choice) are reported in Table 1. No differences were found across the four conditions in a MANOVA analysis (largest $F(3, 47) = 2.15$, $P > 0.10$). Therefore, the baseline condition of participants in each group (which will constitute the experimental conditions in the test blocks) can be considered equivalent after training.

2.2.2. Test Block 1

For the initial analyses of Test Block 1 performance, the 16 diagnostic games were grouped into sets of four successively presented games. Each set contained one game from each of the four game types (due to our design), allowing the effect of type to be factored out of the analyses. Set positions 1 through 4 reflect the first, second, third, and fourth occurrence of each game type (I, II, III, and IV) in the block. Each set position consists of one game of each type averaged together in order to minimize any effects of game type and

Table 1
Proportion of prediction errors and rationality errors in the Training Block for Experiment 1[a]

Experimental condition	Prediction errors		Rationality errors	
	M	SE	*M*	SE
Myopic opponent	0.012	0.005	0.113	0.021
No-switch	0.017	0.009	0.108	0.035
Switch	0.008	0.006	0.117	0.026
Predictive opponent	0.010	0.005	0.096	0.021
No-switch	0.000	0.000	0.089	0.031
Switch	0.018	0.018	0.103	0.030
Total	0.011	0.003	0.105	0.015

[a] The last 16 games in the Training Block for each participant are included. Statistics are reported for the experimental conditions and the population. Myopic and predictive opponents refer to the strategy used by the opponent in Test Block 1, while switch and no-switch refer to whether or not the opponent changes behavioral strategy between Test Block 1 and Test Block 2.

allow a graded calculation of the scores. The catch trials were excluded from this portion of the analysis because they provide no information as to which order of reasoning was used. Due to the equivalent predictions and decisions across strategies in these catch trial games, they should not affect performance on the remainder of the games.

2.2.2.1. Prediction scores In Test Block 1, the hypotheses of interest are (1) that participants will hold a default model of the opponent, and (2) that this model is dynamic, changing as information about the opponent's strategy becomes available. Performance on the initial games in the series should reveal whether there is a default model utilized by participants, or if they merely respond randomly. The effect of the opponent's strategy (either "myopic" or "predictive") on the participants' use of TOM reasoning should become evident through the examination of prediction scores in games across time. For games in later positions within the block, the opponent's behavior should have a larger effect.

A mean prediction score was calculated as the proportion of predictions made using second-order TOM reasoning over one set, i.e. a group of four successive games. As this mean score approaches 0, participants tend to use first-order reasoning in making predictions about Player II's behavior (modeling the opponent as "myopic"). As it approaches 1, participants tend to use second-order reasoning (modeling the opponent as "predictive"). Fig. 5 shows the prediction scores when the opponent (confederate) is behaving myopically or predictively throughout Test Block 1.

To test the statistical significance of the prediction scores at different time points of the test block and for different opponent behaviors, game set position was entered as a within-subject variable, and opponent behavior or strategy (zeroth-order vs. first-order) was entered as the between-subjects variable in a repeated-measures ANOVA on the mean prediction score at each position. The main effect of opponent TOM strategy was found to be significant ($F(1,50) = 5.72$, $P < 0.03$), indicating that the opponent's behavior does

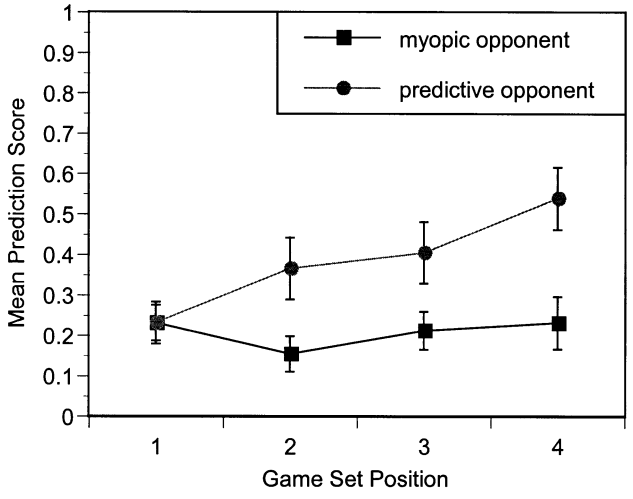

Fig. 5. Mean prediction scores for each game set position in Test Block 1.

affect the participant's use of TOM reasoning over the entire series of games. The main effect of game set position was also significant ($F(3,48) = 5.93$, $P < 0.005$), as was the interaction between opponent behavior and set position ($F(3,48) = 6.34$, $P < 0.005$). Fig. 5 depicts the behavior by position interaction. The main effect of position demonstrates that learning occurs over the series based on the feedback of how the opponent actually behaves, whereas the interaction indicates that the direction of this learning was dictated by the opponent's TOM strategy. Further, the linear trend contrast within this interaction was highly significant ($F(1,50) = 13.62$, $P < 0.001$), suggesting that the opponent's strategy caused a successively greater effect as participants became more experienced at playing the game. In order to determine the trends within this interaction, a priori *t*-tests were used to supplement the ANOVA analysis. At game set position 1, the means are identical ($t(1,50) = 0.00$). This suggests that the same default model of the opponent is used by all participants. In order to test whether this default model differs significantly from random responding, a one-sample *t*-test compared the prediction score at set position 1 with a value of 0.5, which would indicate random guessing. This test was significant ($t(1,51) = -8.03$, $P < 0.001$), supporting the hypothesis that a certain default model of the opponent is being used; in this case, the opponent is predicted to use a zeroth-order TOM strategy.

The hypothesis that the mental model held by participants is dynamic, i.e. modifiable on the basis of counter-evidence, is supported by the interaction between set position and opponent strategy. The prediction scores for participants playing against a myopic opponent do not change from set position 1 to position 4 ($t(1,25) = 0.00$), indicating that they continue to use the default zeroth-order model effectively. On the other hand, participants playing against a predictive opponent exhibit an increase in prediction scores from set position 1 to position 4 ($t(1,25) = -4.92$, $P < 0.001$), supporting the view that they are shifting their strategy toward second-order TOM reasoning in response to counter-evidence that the opponent is not behaving according to the default model.

Taken together, these results support the hypothesis that participants began the task of predicting the opponent with a default model biased toward viewing the opponent as engaging in zeroth-order reasoning; the participants are hence engaged in first-order reasoning themselves. However, by the end of Test Block 1, participants playing against an opponent who consistently uses first-order reasoning are able to make an adjustment to the opponent's behavior and begin to make predictions by using second-order reasoning, thereby supporting the hypothesis that TOM reasoning is adaptive. Nevertheless, this transition should be characterized as incomplete by the end of Test Block 1, as the mean score never rises above 0.54, and only 69% of participants in this condition make predictions using second-order reasoning in half or more of the last four games (in the last set position).

2.2.2.2. Rationality errors Rationality in our game context was defined as the ability to use one's predictions of the opponent in making optimal decisions for oneself. Rationality errors measure the extent of the participants' rational decision-making given their prediction of the opponent's behavior. Fig. 6 displays the mean rationality errors when playing against the myopic (zeroth-order) and predictive (first-order) opponent for all four game set positions in Test Block 1.

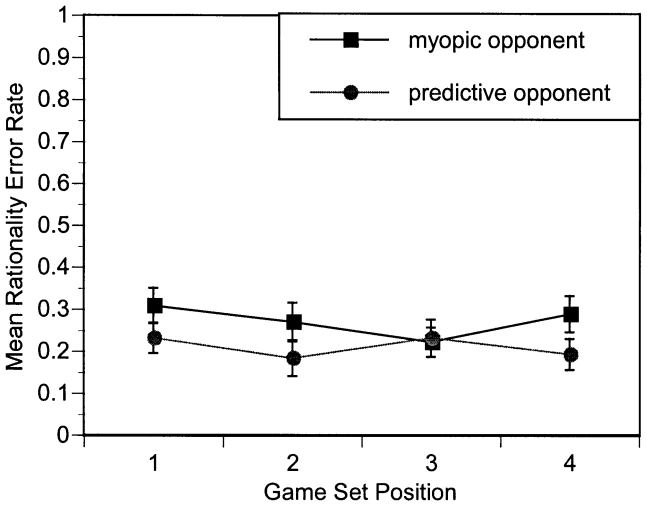

Fig. 6. Mean rationality errors for each game set position in Test Block 1.

The repeated-measures ANOVA on rationality errors yielded no significant differences for either the position ($F(3,48) = 0.92$, $P > 0.4$) or opponent strategy ($F(1,50) = 1.95$, $P > 0.15$) factors. Neither was the position by opponent strategy interaction significant ($F(3,48) = 1.31$, $P > 0.25$), demonstrating that rationality errors did not differ across the conditions. This supports the hypothesis that rationality errors are unaffected by the opponent's behavior (more accurately, the participants' particular mental model of the opponent). Further, this suggests a dual-stage process in which the mental model is first constructed and accessed to predict the opponent's behavior, followed by a second stage of applying this prediction (regardless of the model used to reach the prediction) to generate one's own decision.

2.2.3. Test Block 2

In the second test block, an opponent (our confederate) either maintains his TOM strategy/behavior as that in Test Block 1 ("no-switch condition") or switches to the other TOM strategy/behavior for this entire block ("switch condition"). The primary area of interest is the ability of participants to dynamically change their mental model of the opponent in response to such changes. For the switch condition, changes in the opponent's strategy were either from the myopic one to the predictive one, or vice versa. This provides an opportunity for a strong test of the hypotheses that (1) participants are able to adapt to changes in the opponent's strategy by making on-line modifications of their mental model about the opponent, and (2) rationality (as we defined it) should remain unaffected by such changes in the mental model.

2.2.3.1. Prediction scores The repeated-measures ANOVA on the participant's prediction scores (scored in the same manner as in Test Block 1) yields a significant interaction between Test Block 1 behavior and the behavioral switch ($F(1,48) = 13.15$, $P < 0.001$)

for performance averaged across all four positions. As this interaction is equivalent to testing for effects of the opponent's strategy in Test Block 2, this result demonstrates that the opponent's behavior has a significant effect regardless of whether there has been a behavioral switch or not. For tests involving the within-subjects factor of game set position, the three-way interaction between opponent's strategy in Test Block 1, strategy switch, and position was highly significant ($F(3, 46) = 11.02$, $P < 0.001$). This interaction is depicted in Fig. 7, where the prediction scores were plotted separately for those conditions when the opponent behaves consistently across the two test blocks (Fig. 7A) and when the opponent switches his behavior between these blocks (Fig. 7B). More

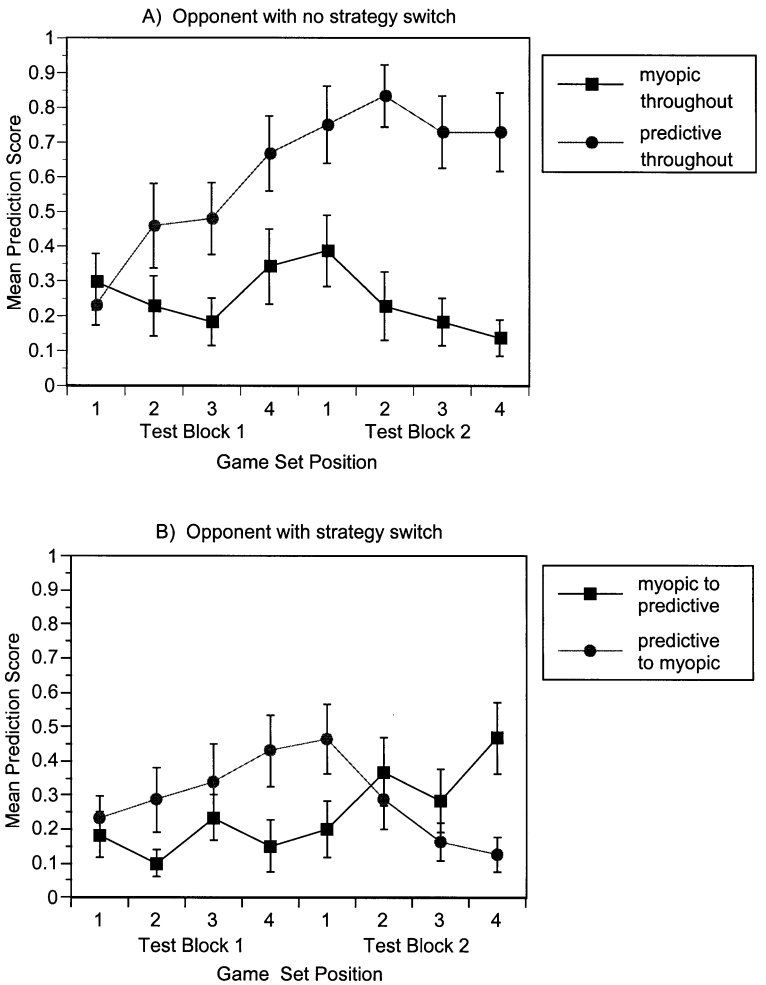

Fig. 7. Mean prediction scores across Test Blocks 1 and 2. (A) Conditions in which no behavioral switch (NS) by the opponent occurred. (B) Conditions in which the opponent's behavior switched (S) between Test Block 1 and Test Block 2.

participants facing a consistently myopic opponent stabilize in the use of first-order reasoning, while more participants facing a predictive opponent stabilize in the use of second-order reasoning. However, if a behavioral switch on the part of the opponent occurs, there is a crossover interaction such that the prediction scores follow the behavioral strategy of the opponent, indicating that participants dynamically adapt to the opponent's behavior. Furthermore, the participants are more readily adapted to a shift in the opponent's behavior when the opponent switched from the first-order behavior to the zeroth-order behavior than vice versa. This is consistent with the findings in Test Block 1 that participants tend to have a default model of the opponent as using the zeroth-order strategy. It should be noted that this effect appears to be largely driven by the continuing improvement in the predictions made by participants playing against a consistently predictive opponent. By comparing the prediction scores at position 4 of Test Block 2 where the opponent switches from the myopic to the predictive behavior with position 4 in Test Block 1 for participants in the predictive-opponent condition, one can see that the scores are roughly equivalent. Therefore, it would appear that there is no switching cost for either type of strategy shift, but that participants return to the default model easily, while further improvement is possible in the case where the opponent engages in first-order reasoning.

2.2.3.2. Rationality errors Despite the possible change of the opponent's behavior and therefore the TOM model of the opponent, the analysis of the rationality errors in Test Block 2 yielded no significant interactions with set position for either the opponent's strategy/behavior in Test Block 1 ($F(3,46) = 0.047,\ P > 0.98$), for the behavioral switch factor ($F(3,46) = 0.21,\ P > 0.89$), or for the three-way interaction ($F(3,46) = 1.25,\ P > 0.3$). Neither was the main effect of set position significant ($F(3,46) = 1.51,\ P > 0.22$). These results support the hypothesis that participants are able to utilize their predictions of the opponent to formulate consistent decisions, and that this process is unaffected even in the face of changes in their model of the opponent.

2.2.4. Game type

For this portion of the analyses, individual games were grouped by type (labeled I, II, III, and IV as listed in Appendix B) rather than by their positions in a block. Each type occurs exactly four times spaced more or less evenly throughout each test block of trials. Individual game types were examined in order to determine whether particular game structures affect the predictions and rational decisions of participants.

2.2.4.1. Prediction scores The repeated-measures ANOVA for prediction scores in Test Block 1 using opponent behavior as the between-subjects variable and game type as the within-subjects variable yielded a significant main effect of opponent's behavior across all game types ($F(1,50) = 5.91,\ P < 0.02$). This demonstrates that participants playing against a first-order opponent had higher prediction scores overall, which was true for all game types. In addition, the main effect of game type was significant ($F(4,47) = 10.08$, $P < 0.001$). This is to say that different types of games, due to their particular structures, induce different propensities for using second-order reasoning. For instance, for game type IV, there is a high payoff to both players (a 4 for both, or a 4 and a 3) in cell D, indicating

that it would be beneficial for both if the game were to end in cell D (we call this the "4–4" heuristic). This heuristic, through the recognition of structure, may induce participants toward a "move" prediction on the part of the opponent, and a "move" decision on the part of themselves, though sometimes this is only wishful thinking when facing a myopic opponent. Although such structural aspects of the games may influence participants' prediction of the opponent's behavior, it would appear that the effects of game type are independent from the effect of the opponent's strategy since the game type by opponent behavior interaction is non-significant ($F(4,47) = 0.85,\ P > 0.5$). The lack of an interaction between the two factors indicates that the opponent's behavior, or participants' experience with it, provided an additive influence to their prediction scores.

The same analysis was applied to games in Test Block 2. Game type was again found to be highly significant ($F(4,47) = 10.15,\ P < 0.001$) in the analysis of prediction scores. The main effect of the opponent's strategy was significant when averaged across all game types ($F(1,50) = 9.31, P < 0.005$), while the interaction between the opponent's strategy in Test Block 2 and game type was not significant ($F(4,47) = 1.55, P > 0.2$). Thus, the Test Block 2 performance for game type coincides with the findings in Test Block 1: game structure and opponent behavior are two independent factors that influence the prediction score of a participant. The results of the analyses from Test Block 1 and Test Block 2 are represented graphically in Fig. 8.

Catch trials were expected to yield no differences between conditions due to the fact that the predictions were identical for both the zeroth-order and first-order opponent strategies in these games. Indeed, one-way ANOVA tests demonstrated no difference on catch trials in Test Block 1 ($F(1,50) = 1.32, P > 0.25$) or in Test Block 2 ($F(1,50) = 0.89, P > 0.3$).

2.2.4.2. Rationality errors A repeated-measures ANOVA on rationality errors found no significant differences for the main effect of opponent strategy in either Test Block 1 ($F(1,50) = 1.07,\ P > 0.3$) or in Test Block 2 ($F(1,50) = 1.96,\ P > 0.15$). The interaction between opponent strategy and type was also not significant in Test Block 1 ($F(4,47) = 1.86,\ P > 0.1$) or in Test Block 2 ($F(4,47) = 1.76,\ P > 0.15$). Thus, it

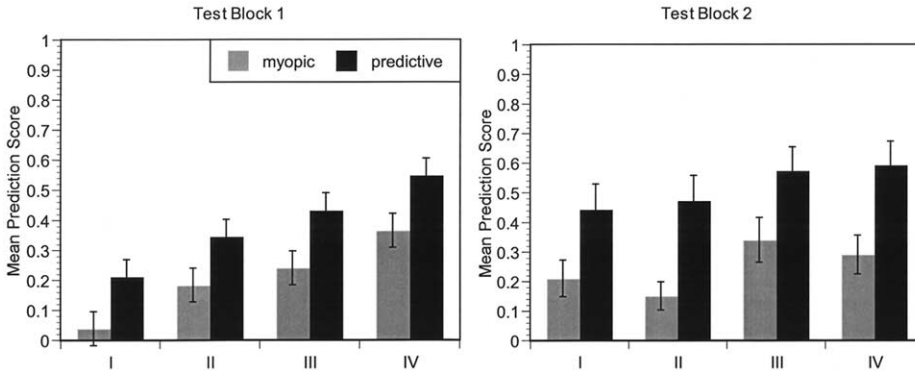

Fig. 8. Mean prediction scores for each game type in Test Blocks 1 and 2. The ordering of types is for graphic purposes only, and should not be taken as a progressive trend. Rather, it is important only to note that there are differences among types.

appears that opponent behavior does not affect rationality errors committed by participants across game types. However, there was a difference among game types in both Test Block 1 ($F(4, 47) = 13.23$, $P < 0.001$) and in Test Block 2 ($F(4, 47) = 11.63$, $P < 0.001$). This led us to a more in-depth analysis of these rationality errors. We coded the rationality errors contingent upon the prediction each participant made about the opponent. Thus, all rationality errors committed when a participant made a first-order prediction were recorded for each game type within each block, as were all rationality errors associated with a second-order prediction. These errors are graphed in Fig. 9.

Type II games in Test Block 1 appear to have a disproportionately large number of rationality errors associated with predictions of the opponent being myopic. To understand this, we examined the payoff structure underlying games of this type. In Type II games (an example of which is given in Fig. 10), the opponent's payoffs are always D < B < C, whereas the participant starts with a 3 and has a higher payoff (a 4) only in cell D. If the opponent is modeled as a predictive player, and therefore not predicted to move away from cell B, most participants correctly decide to stay in cell A, and a low rationality error rate is observed. However, if the opponent is modeled as a myopic player who would move to cell C, then the rational decision for our participants ought to be to move from cell A to cell B in order to finally reach cell D. Instead, about half of the time, participants chose to stay in cell A despite their mental model of an opponent who would move. It appears that they failed to consider that they will then have a second turn to move to cell D for a more favorable outcome after all. This amounts to a comparison of cell A and cell C in making their initial decision, thereby committing a rationality error when the optimal decision should have been based on a comparison of cell A and cell D (hence, we refer to this error as being due to the "ignore-D" heuristic). In other words, it appears that participants are not engaged in fully enacting the possible consequences of their opponent's action and making contingency plans based on these predicted consequences.

However, in Test Block 2, this particular rationality error has dissipated. A close examination of game types in Test Block 2 reveals that, because all the games start with a payoff of 2 for the participant ("2-starting games" by our design), the decision to move away or to stay in cell A is a much simpler one; participants can often ignore the opponent's payoff and strategy in making their decision. This is because there is only one

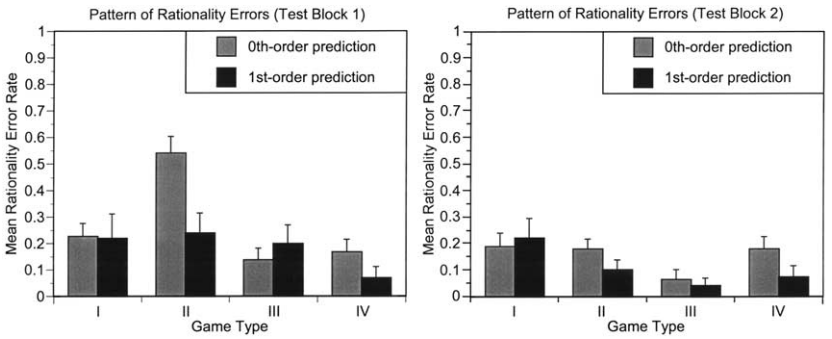

Fig. 9. Mean rationality errors in Test Blocks 1 and 2. Errors were scored contingent upon the prediction about the opponent (whether the opponent will use zeroth-order or first-order reasoning) made by a participant.

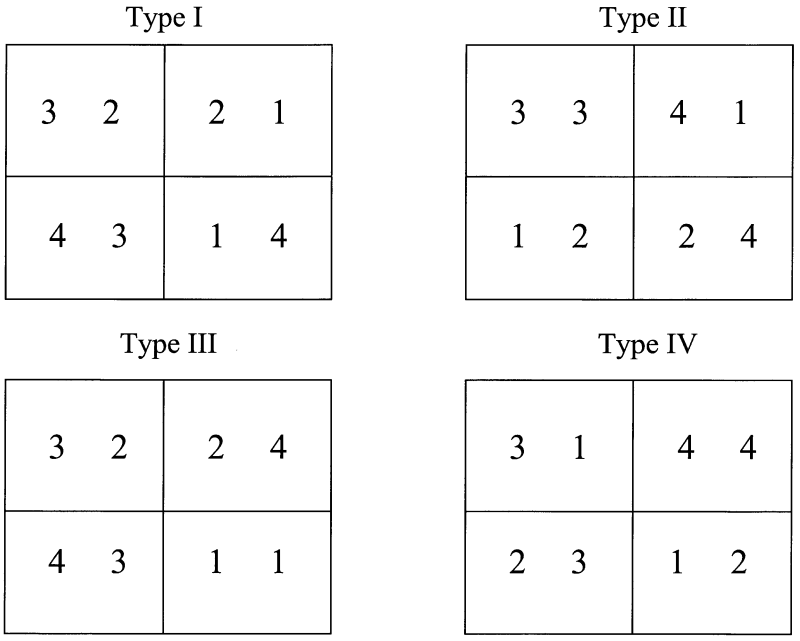

Fig. 10. Sample games from Test Block 1 displaying the characteristic structure of each game type.

outcome having a lower value (a 1) for the participant. If that payoff of 1 falls in either cell C or cell D, the participant can always decide to move (thereby insuring a payoff of 3 or 4) regardless of what the opponent might do. This can be termed an "always-move" heuristic. Unfortunately, the situation when this heuristic is applicable includes certain Type II games in Test Block 2 where the particular rationality error might have been committed (as in Test Block 1).

Taken together, the data from the game type analyses support a view that both game structure and opponent's behavior influence the participant's prediction scores significantly but independently. In making decisions, participants appear to be able to understand the consequences of predicted outcomes (as indicated by the fairly constant rate of rationality errors), albeit sometimes not fully (as indicated by the game type II rationality errors).

3. Experiment 2

Experiment 1 demonstrated that participants exhibit a default mental model based upon first-order TOM reasoning, and that this model can be modified as counter-evidence becomes available. Our paradigm delineates two stages of processing: a stage in which a mental model is constructed to predict the opponent's behavior, followed by a second stage where the prediction provides a basis for making optimal decisions. However, although the hypothesis that modifiable mental models are employed received strong support, no direct evidence (e.g. response times) was provided to demonstrate that these

models are in fact hierarchically structured (with increasing complexity). Nor was any indication provided as to why the first-order model was predominately established as the default.

Experiment 2 was designed with three major aims in mind. One, to replicate the main findings of Experiment 1, namely, the results supporting a default model that is updated when presented with new information over time. Two, to attempt to influence the order of TOM used as the default model. In order to accomplish this, a manipulation of the opponent's appearance was instituted and the specific hypothesis that a more intelligent appearing opponent would invoke a higher-order default model was advanced. Three, to validate the hierarchical structure of mental models based on increasing orders of TOM reasoning. To this end, reaction times were collected to test the hypothesis that higher-order reasoning should take longer precisely because it involves all mental operations supporting lower orders of TOM reasoning (as seen in Fig. 4). Further, this increase in reaction time should be observed only in the process of making predictions, but not for making decisions. This is due to the two-stage nature of our task. One first predicts what the opponent will do, invoking different orders of TOM reasoning, and then bases the decision on the output of this prediction. Regardless of the order of reasoning used to reach the prediction, the decision process should remain constant.

3.1. Method

3.1.1. Participants

Participants were 70 (36 males, 34 females) undergraduate University of Michigan students who received class credit for their participation. The average age of the participants was 19.2 years. All participants gave informed consent prior to admission into the study and were provided with information about the nature of the computerized opponent at the conclusion of the experiment.

3.1.2. Materials

The materials were identical to those used in Experiment 1.

3.1.3. Design

The design was similar to that of Experiment 1. However, one manipulation was deleted. Namely, no strategy switch occurred on the part of Player II (confederate) after Test Block 1 in any condition; the opponent adopts the same (myopic or predictive) behavior in both test blocks. An additional manipulation was instituted instead. Participants were introduced to one of three opponents at the outset of the experimental session. The opponent was either presented as an intelligent human, as a not-so-intelligent human, or as a computer. The overall design therefore involves two between-subjects factors, confederate type (three levels) by opponent strategy (two levels).

3.1.4. Procedure

Participants were randomly assigned to conditions. Upon arriving at the experimental session, those participants in the "computer" condition were told that they would be playing a game with a computerized opponent programmed to play like a human. Parti-

cipants in the human confederate conditions were told that another student was also scheduled for that session and were asked to wait until that student arrived. The confederate then entered the room and followed a script. The confederate in the "intelligent" condition, while carrying a mathematics book, apologized for being late, saying that he had been tutoring a student in calculus and that the session had run long. The confederate in the "unintelligent" condition, while carrying a supermarket tabloid, said that she had been at her calculus tutor's and the session ran long. At the experimenter's prompting, the confederate and the participant introduced themselves. The intelligent confederate portrayed himself as an engineering student involved in the chess club and a member of the Honors College. The unintelligent confederate portrayed herself as an undeclared major who "just likes to hang out" and found the Introduction to Psychology class "really hard". The confederate conditions were designed to utilize stereotypes to invoke a perception of intelligence or unintelligence that would hopefully influence the manner in which participants behaved when playing a TOM game with that opponent.

Procedures were the same as in Experiment 1, with the exception that participants in the human confederate conditions were administered questionnaires asking for ratings of the opponent's intelligence, appearance, and friendliness both before and after the games were played. Ratings were given on a Likert-type scale ranging from 0 to 10. Of participants in the human confederate conditions, 12.5% indicated a suspicion that the opponent was actually a computer on an exit questionnaire.

3.2. Results and discussion

3.2.1. Impression questionnaires

The mean ratings given by participants in each confederate condition are shown in Table 2. A multivariate ANOVA showed no effect of confederate appearance on attractiveness ($F(1,46) = 1.79$, $P > 0.15$) or friendliness ($F(1,46) < 1$) at either time. The expected difference in intelligence ratings was significant at the first measurement ($F(1,46) = 10.84$, $P < 0.005$). However, at the second measurement, after the playing of the games, this difference was no longer significant ($F(1,46) = 1.63$, $P > 0.2$). Hence, the confederate manipulation appears to have affected the participants' perceptions of the opponent's intelligence at the outset of the session in the expected direction.

Table 2
Ratings for intelligence, attractiveness, and friendliness of the confederate in Experiment 2[a]

Confederate		Rating scale score					
		Intelligence		Attractiveness		Friendliness	
		Time 1	Time 2	Time 1	Time 2	Time 1	Time 2
Intelligent	*M*	7.96	6.17	4.00	3.91	6.74	5.96
	SE	(0.24)	(0.48)	(0.39)	(0.36)	(0.25)	(0.39)
Unintelligent	*M*	6.56	5.72	4.64	4.68	6.60	6.20
	SE	(0.34)	(0.43)	(0.43)	(0.44)	(0.32)	(0.32)

[a] Ratings were obtained both before (Time 1) and after (Time 2) playing the series of games.

3.2.2. Training Block

Prediction errors and rationality errors in the Training Block were again used to exclude participants who failed to understand the nature of the game from further analyses. Participants with greater than seven rationality errors or greater than three prediction errors were excluded from further analyses. A total of 61 participants remained.

For the last 16 training games presented to a player, the mean number of errors made in predicting the opponent (whether the opponent will stay or switch in cell B) and in making a rational decision (translating one's prediction of the opponent into a decision) are reported in Table 3. No differences were found across the conditions in a MANOVA analysis (all $F < 1$). Therefore, the baseline condition of participants in each group can be considered equivalent after training.

3.2.3. Test Blocks 1 and 2

Because an opponent behaves consistently (whether myopic or predictive) in both Test Blocks 1 and 2 in any condition, the two blocks can be considered as one larger block. All subsequent analyses included games from both test blocks in the game type and set position factors (recall that a set position reflects the average of one occurrence of each game type).

3.2.3.1. Prediction scores Effects involving the confederate appearance manipulation did not approach significance at the first set position ($F(2, 55) = 1.49$, $P > 0.20$), nor across positions ($F(14, 100) < 1$). Neither did interactions between confederate appearance and opponent strategy reach significance ($F(2, 55) = 1.77$, $P > 0.18$). Although confederate appearance did influence the intelligence judgments made by participants (see above), it failed to influence the default model or the manner in which this model is modified as the games are played. Hence, all subsequent analyses combine across this factor and involve only the opponent strategy conditions.

This study provides a clean replication of the results found in Experiment 1. Again, a repeated-measures ANOVA involving opponent strategy and game set position resulted in a main effect of opponent strategy ($F(1, 59) = 21.55$, $P < 0.0005$), a main effect of set position ($F(7, 53) = 2.61$, $P < 0.03$), and a significant interaction ($F(7, 53) = 7.46$, $P < 0.0005$). The means used in this analysis are graphed in Fig. 11. The hypothesis predicting a default mental model was supported, as the means at the first position did not differ from one another ($t(59) = 0.01$), and they were significantly below the chance

Table 3
Proportion of prediction and rationality errors in the Training Block for Experiment 2[a]

Experimental condition	Prediction errors		Rationality errors	
	M	SE	*M*	SE
Myopic opponent	0.023	0.007	0.133	0.023
Predictive opponent	0.026	0.010	0.121	0.019
Total	0.025	0.006	0.127	0.015

[a] The last 16 games of the Training Block for each participant are included.

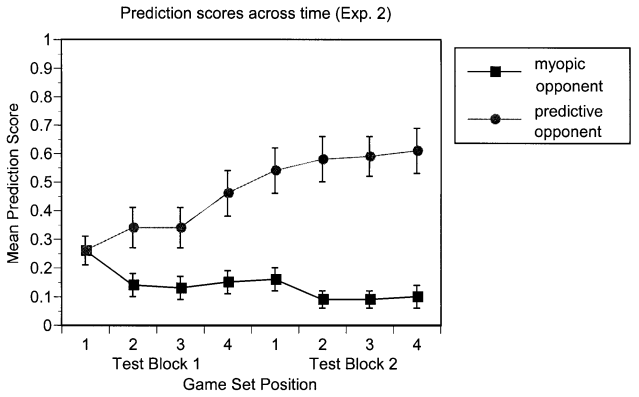

Fig. 11. Mean prediction scores for each game set position in Test Blocks 1 and 2 for Experiment 2.

level of 0.5 ($t(60) = 6.90$, $P < 0.0005$). This indicates again that participants used a default first-order TOM model of Player II. The hypothesis that this model is dynamic, changing as counter-evidence becomes available, is supported by the main effect of set position and the interaction of opponent strategy and position.

3.2.3.2. Rationality errors Again, confederate appearance did not influence the number of rationality errors, having no main effect ($F(2, 55) < 1$) or interactions ($F(2, 55) = 1.08$, $P > 0.30$). This factor was therefore dropped from all subsequent analyses.

The results of the analyses of rationality errors largely replicated those of Experiment 1. A repeated-measures ANOVA found no significant effects of opponent strategy ($F(1, 59) < 1$), and a marginal effect for the interaction of strategy and position ($F(7, 53) = 2.04$, $P < 0.07$). The main effect of game set position was significant ($F(7, 53) = 6.74$, $P < 0.0005$). The mean rationality errors are displayed in Fig. 12.

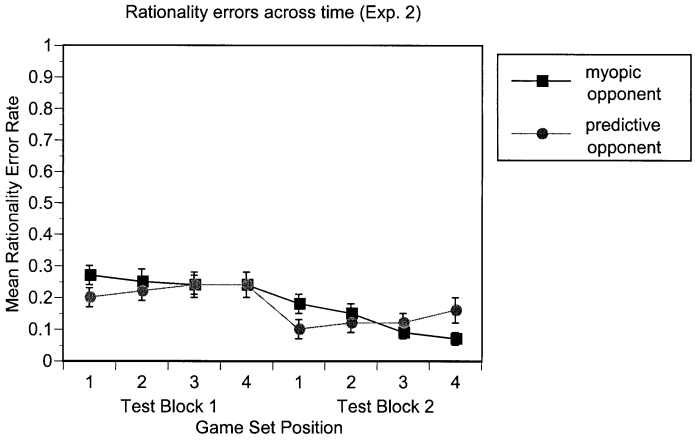

Fig. 12. Mean rationality errors for each game set position in Test Blocks 1 and 2 for Experiment 2.

3.2.4. Game type

3.2.4.1. Prediction scores The game type analyses produced results similar to those found in Experiment 1. A repeated-measures ANOVA with type as the within-subjects variable found no evidence for an opponent strategy by game type interaction ($F(3,57) < 1$). The main effect of game type was significant ($F(3,57) = 12.14$, $P < 0.001$), as was the main effect of opponent strategy ($F(1,59) = 21.55$, $P < 0.001$). To reiterate the conclusions from Experiment 1, participants playing against a myopic opponent had lower overall prediction scores than did those playing against a predictive opponent, and this was true for all game types. The game structures, on the other hand, provide independent influences to the participant's prediction scores, possibly through heuristic processing (e.g. the 4–4 heuristic for game type IV).

3.2.4.2. Rationality errors A repeated-measures ANOVA on rationality errors for game types found no significant interaction between opponent strategy and game type ($F(3,57) < 1$), and no main effect of opponent strategy ($F(1,59) < 1$). The main effect of game type was significant ($F(3,57) = 23.09, P < 0.001$). This pattern of results mirrors the pattern observed in Experiment 1.

3.2.5. Reaction time

Reaction times for all predictions and decisions were collected. In order to obtain meaningful reaction time data, each participant's predictions were coded as either first-order or second-order predictions. Mean reaction times for each type of prediction were then computed. The same procedure was used to obtain mean reaction times for decisions under each type of prediction. Fig. 13 shows a scatter-plot, demonstrating the relationship between mean reaction times for first-order and for second-order predictions for each individual participant. All points below the diagonal represent participants who took longer to make a second-order prediction than to make a first-order prediction. Those above the diagonal took longer to make a first-order prediction. Clearly, the majority of participants fall below the diagonal.

When the individual data are combined (in Fig. 14) to yield an overall mean reaction time for each order of prediction, the data unambiguously show that second-order predictions took longer to produce than did first-order predictions ($t(48) = 6.83$, $P < 0.0005$), but there was no corresponding difference in decision times ($t(48) < 1$). Hence, the two-stage hypothesis claiming that decisions are based on predictions (but independent of the model used to reach the predictions) is supported, as is the hypothesis that the orders of TOM are hierarchical. That is, the procedures involved in second-order reasoning necessarily encompass a subroutine consisting of first-order reasoning.

3.2.6. Correlational analysis

By analyzing performance on the entire series of games in Test Blocks 1 and 2, it is possible to correlate each participant's behavior with a number of possible behavioral strategies that have been identified, including strategies that may be masked by the orthogonal design employed in the current studies.

A number of strategies were considered for this analysis. The optimal set of decisions

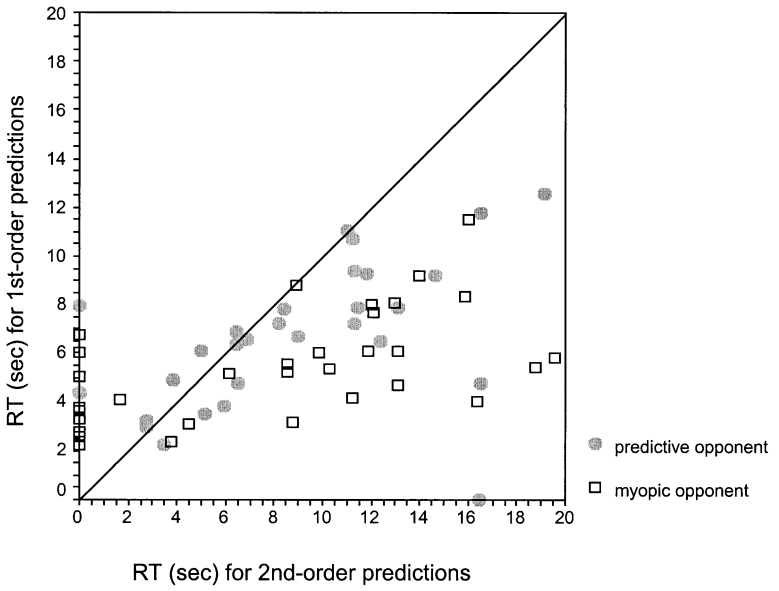

Fig. 13. Scatter-plot of reaction times to predict the opponent's choice. Time to make a prediction that the opponent will use zeroth-order reasoning (a 1st-order prediction) is plotted against time to predict the opponent will use first-order reasoning (a 2nd-order prediction) for each participant.

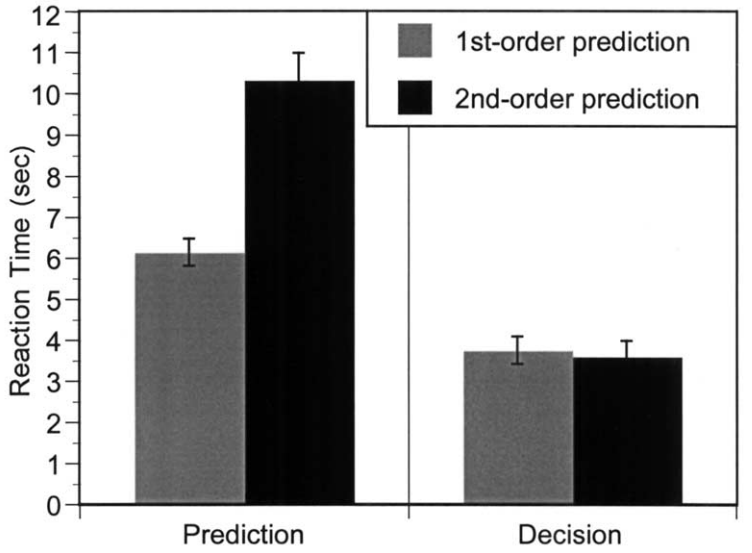

Fig. 14. Mean reaction times to make a prediction about the opponent and to decide whether to stay in or move away from cell A. Reaction times are separately coded for predictions using and decisions following first-order or second-order reasoning.

that each such strategy predicts was correlated with the set of actual decisions of each participant. By selecting the largest correlation value, each participant was classified as using the corresponding strategy. The number of participants using each strategy in the myopic and predictive conditions was then calculated. The results of this calculation are shown in Table 4. It can be seen that in the myopic-opponent condition, the most popular strategies involved first-order reasoning, although several variations on this strategy were observed. It appears that many participants supplemented first-order reasoning with at least one heuristic. These include the 4–4 heuristic (discussed above) and the ignore-D heuristic, in which participants fail to take the cell D payoffs into account while projecting the consequences of a decision. A few participants tended to use a conjunction of the 4–4 and ignore-D heuristics, while no participants used the simple heuristics of always choosing to move from or stay in cell A. For the predictive condition, the most popular strategy was based on second-order reasoning. However, there is a considerable spill-over to the first-order strategy and its variants – many participants fail to establish second-order reasoning even at the last game set position (see Fig. 11). This difference between the conditions was significant in a Chi-square analysis ($\chi^2(5) = 21.02$, $P < 0.001$). This can be taken to confirm our emphasis on the first-order and second-order TOM strategies in the previous analyses. The intrinsic correlations of each behavioral strategy with the first-order and second-order strategies intended by our orthogonal design are also shown in Table 4. (Note that our experimental design allows a -1.00 correlation between first-order and second-order predictions, but that the decisions based on these predictions only allow a -0.60 correlation, as eight out of 16 diagnostic games in Test Block 2 yield a "switch" decision for either a first-order or a second-order prediction.) Although intercorrelations

Table 4
Correlational analysis of TOM strategies and heuristics for Experiment 2[a]

Possible behavioral strategy	Experimental condition		Intercorrelation	
	Myopic	Predictive	First-order	Second-order
Always move	0	0	0.23	− 0.14
Always stay	0	0	− 0.23	0.14
Zeroth-order	5	2	0.26	0.26
First-order	12	3	1.00	− 0.60
First-ignore-D	9	6	0.60	− 0.47
First + 4–4	0	1	0.49	− 0.29
First-ignore-D + 4–4	0	2	− 0.07	− 0.07
Second-order	0	13	− 0.60	1.00

[a] The series of decisions made by individual participants in Test Blocks 1 and 2 were correlated with the decisions provided by each behavioral strategy. The behavioral strategy with the largest correlation was identified as the predominant strategy employed by the participant. The number of participants using each corresponding strategy is given for the two experimental conditions. Heuristic strategies examined include the 4–4 heuristic and the ignore-D heuristic in conjunction with use of first-order reasoning, as well as the simpler heuristics of always deciding to move or always deciding to stay. Intercorrelations of each possible behavioral strategy with the first-order and second-order TOM strategies are given on the right (a consequence of the selection of games for Test Blocks 1 and 2).

among strategies may produce difficulties for this analysis, the gains in the ability to parse the use of particular strategies by individuals may provide valuable information.

4. General discussion

The results of these two experiments show that (1) participants begin with a default mental model of a myopic opponent who employs zeroth-order reasoning, (2) the mental model used by a participant is influenced by the (myopic or predictive) behavior of the opponent within the game setting, and (3) this mental model is dynamic in nature, responding to changes in the strategies adopted by the opponent. With regard to the decision process, participants are able to make rational decisions based on a mental model of the opponent's predicted behavior independent of what kind of opponent (myopic or predictive) their model predicts. However, game structure was found to affect predictions as an independent factor, possibly through heuristic processing. Further, a specific kind of rationality error was revealed that appears to be due to a failure to fully comprehend the consequences of the opponent's behavior; future choice options that become available if the opponent behaves exactly as predicted are ignored by participants who exhibit a difficulty in planning ahead by more than one hypothetical step of reasoning.

Results from both experiments provide support for the formation of a default mental model, namely, that participants, on average, adopt a first-order mental model, predicting a myopic opponent at the beginning of Test Block 1. It should be noted that we cannot conclude that the first-order model is the common default for college-age adults. Our results may be specific to the task and the game setting employed in the experiments. Furthermore, the default may change with factors of experience, motivation, expectations, etc. For example, results from normal controls (average age of 42 years) in a study of frontal patients demonstrated that first-order inferences in a TOM task were correctly completed by about 90% of participants, while second-order inferences were correctly completed by approximately 70% of the sample (Rowe et al., 2001). However, a manipulation of conceptualization of the opponent failed to influence the choice of default model in the present study, although the manipulation was powerful enough to influence intelligence judgments about the opponent. It is also possible that the Training Block contained a preponderance of games in which cell D was not reached. This may have conditioned participants to ignore cell D (the "ignore-D" heuristic) even in cases when it was necessary for an optimal analysis, as revealed by rationality errors committed in game type II in Test Block 1.

There is clear evidence that although the first-order model may be the default for most participants, many are capable of adapting to the use of second-order reasoning. This is shown by the convergence toward the second-order strategy by those participants playing against a first-order (predictive) opponent. Because participants playing against a zeroth-order (myopic) opponent were already using the default first-order strategy and were thus behaving optimally against such an opponent, no such adjustment was required in this condition. However, the updating of mental models, on average, appears to be fairly slow and incomplete. The slowness with which participants were able to change from a first-order model to a second-order model may be indicative of confirmation bias (Evans, 1987, 1993). That is, participants construct a default mental model using first-order reasoning

(therefore viewing the opponent as engaging in myopic reasoning), and perseverate or focus on this initial representation (Legrenzi & Girotto, 1996; Legrenzi et al., 1993). This focusing may be at least partially due to confirmation bias, in which the participants fail to take account of counter-evidence provided by the opponent's behavior across trials. It is also possible that participants use this default because they represent the gaming situation as an inductive or probabilistic, rather than a deductive, task (Oaksford & Chater, 2001). Whereas a normative deductive account would lead to a second-order default model, a probabilistic account might regard the likelihood of ending in a given cell as uncertain, rather than deductively knowable when playing against a fully rational opponent. Oaksford and Chater (2001) have noted that 90% of university students appear to use probabilistic reasoning in normatively deductive situations such as the selection task (Stanovich & West, 1998).

There may be two main objections to our approach to probing the recursive modeling of others' minds (and their models of our minds, and so on). The first objection concerns the use of matrix-type games, as some authors have in the past debated the general suitability of such games for investigating higher-order TOM (Perner & Wimmer, 1985). For instance, deception games (e.g. the "windows task"), in which a child attempts to strategically misdirect a confederate's choice in order to obtain a reward (Samuels, Brooks, & Frye, 1996), tend to confound decisions made using zeroth-order and second-order reasoning. As a remedy, either introspective reports are used to supplement the game analysis (Perner, 1979), or a different paradigm is employed in which the child answers questions about characters in a story (Baron-Cohen, 1989; Perner & Wimmer, 1985). The present design is able to get around such intrinsic limitations of static 2×2 games and allows for payoff structures that uniquely determine a single order of TOM without resorting to introspective reports or justifications to distinguish between orders. In contrast to a prior study by Perner (1979), in which the opponent's choice was determined by a dominating strategy, in our critical games, no dominating strategy exists, and the opponent's choice can only be predicted by explicitly modeling whether or not the opponent is predicting the participant's own choices.

The second objection is in regards to an alternative account of the data obtained from our paradigm, namely, that participants may use backward induction to determine their preferred course of action. This is the strategy generally assumed by most research involving game = theoretic situations (Aumann, 1995; Carbone & Hey, 2001). Under backward induction, participants would examine the payoff structure of the game, determine how the game would proceed at the last choice point, and reason through the extensive form in reverse direction to find the optimal decision for the first choice point. Such reasoning, mandated by the rational assumption under situations of complete information (Aumann, 1995), and requiring that only first-order reasoning *about payoffs* at each choice point be applied successively, could in theory produce behavior indistinguishable from that of a true second-order mental model that involves reflexive reasoning *about beliefs* (i.e. beliefs about beliefs).

However, three aspects of the present data suggest that participants were not employing backward induction in their play of the games. First, the observed behavior at the outset of the game is best described by the first-order mental model, which appears to serve as a default for most participants in the absence of feedback about the opponent's strategy. If

backward induction were the predominant strategy, the default behavior should conform instead to an outcome as predicted by second-order reasoning. Second, the behavior and predictions of the participants changed over time in response to the behavior of their opponent. Participants encountering an opponent who switched strategies between blocks were able to adapt by changing their mental model of the opponent. If backward induction had been employed, participants would likely have thought that the opponent did not understand the game and would not know how to respond. In contrast, participants were quite able to adapt to a changing opponent, even when that opponent's observed behavior would have to be considered irrational from a backward induction perspective. Third, an "ignore-D" heuristic involving failures to look ahead to future cells was identifiable (see discussion of Experiment 1 – Section 2.2.4.2) and appears to have been used by a substantial subset of participants. Backward induction, as it starts from the last cell in the series of play, would not lead to errors based on a failure to look ahead, but rather to errors involving failures to look backwards (i.e. to consider earlier cells).

Others have noted that backward induction not only fails to account for a variety of human reasoning and decision-making behavior, but that it is not robust in gaming situations where parity and certainty do not exist (Brams & Kilgour, 1998; Carbone & Hey, 2001). Of relevance, parity between the two players is not present in the current set of games (i.e. one player has greater control over the outcome of the game). Instead of using backward induction, participants appeared to construct a dynamic, and eventually, recursive mental model that accounted for the strategy used by the opponent in accordance with the opponent's observed behavior. This mental model appears to be based primarily upon a notion of what the opponent thinks about the gaming situation, including the opponent's understanding of the nature of play, desirable outcomes, and anticipation of the player's own strategy. For instance, a first-order mental model of the opponent relies upon a representation of the strategy governing the opponent's myopic behavior, which happens to be based on incomplete or faulty knowledge of the player's own strategy. The payoffs of a game alone do not determine a participant's course of action, which must account for the payoffs in conjunction with a prediction of the strategic order of reasoning used by the opponent. For these reasons, we believe that the mental models employed by participants in this gaming situation measure the application of TOM reasoning in a game setting.

Finally, our results indicate that game structures influence performance. Hence, the game structures used in these experiments are catalogued in Appendix B. All diagnostic games require a non-myopic player to take a calculated risk and switch to a lower payoff temporarily (in cell B), or to overcome the temptation of myopically switching to a higher payoff (in cell B or C). Rationality errors were constant across all opponent conditions and in the face of switches in the opponent's behavior. This supplies evidence that the predictions of the opponent, once made, are translated into one's own decisions about how to play the game. However, certain game structures promote the use of heuristics in making decisions (e.g. Type IV games cause a higher probability that participants will conform to second-order reasoning via the 4–4 heuristic), while structural characteristics in other games may influence only the decision stage, causing an increase in rationality errors (through the ignore-D heuristic). If the game matrices can be successfully adopted for use with young children (and the use of other matrix games with children points to this possibility), it would provide a more rigorous and quantitative measure of TOM devel-

opment and strategic thinking than has been available in the past. This would allow the paradigm to be used to study the development of the use of mental models as cognitive structures for processing information about other minds.

Acknowledgements

The authors would like to acknowledge Colleen Seifert and several anonymous reviewers for their comments on an earlier draft, and thank Sara Deringer, Ki Goosens, Alex Kurakin, Cristina Oliva, and Brooke Rossi for their assistance in data collection.

Appendix A

The figure below displays a sample game, replacing the generic symbols (c.f. Fig. 2) for payoffs with values specific to a given game (Fig. 15). The game begins in cell A, where Player I (the participant) has the first turn. In the play of this game, Player I would first predict what Player II (the opponent) will do if given the opportunity (this is formally known as the sub-game analysis). In order to do this, Player I should first examine cell B and note the payoff to Player II, in this case a 3. One possibility would be that Player II, having compared this payoff with his payoff in cell C, a 4, decides to move to the larger value. In this case, Player II would behave according to what we have classified as a zeroth-order strategy. A second possibility is that Player II will realize that by moving to cell C, Player I will get a second turn, and will most likely move to cell D, which contains a 4 for Player I. In this case, Player II would end the game with only a 2. Therefore, while

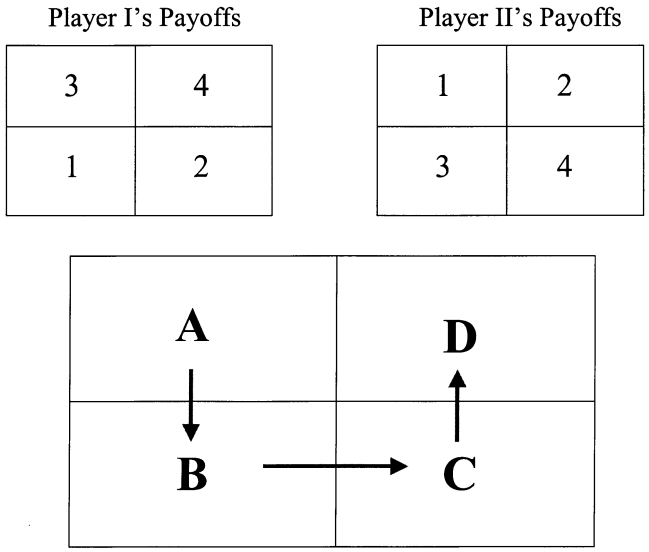

Fig. 15. A sample game in which the generic payoffs from Fig. 2 have been replaced with payoffs specific to an individual game.

still in cell B, Player II is likely to decide to stay there, thus terminating the game in cell B and receiving a 3. In this case, Player II would have behaved according to a first-order strategy. In this scenario, the reasoning that Player I just went through would fall under the first-order and second-order classification, respectively, because Player I's prediction about Player II was based on modeling the latter as using a zeroth-order or a first-order strategy.

Dependent upon the prediction made about his opponent, Player I must then decide what to actually do, stay in cell A to end the game and receive a 3, or move to cell B and allow the game to continue. If Player I predicts that Player II will move from cell B to cell C, the rational decision is to also move. On the other hand, if Player I predicts that Player II will stay in cell B, the rational decision is to stay in cell A. Hence, a participant's decision cannot be made on the basis of the payoffs alone, but must take into consideration Player II's likely strategy.

Appendix B

Position in series	Payoffs to Player I	Payoffs to Player II	Quadruplet	Game type
	A B C D	A B C D	*m p M P*	
Training Block				
1	2 4 3 1	1 2 4 3	1 1 1 1	Trivial
2	3 4 2 1	1 2 3 4	1 1 0 0	Trivial
3	2 1 4 3	4 3 1 2	0 0 0 0	Trivial
4	3 4 1 2	4 3 1 2	0 0 1 1	Trivial
5	2 3 4 1	1 2 3 4	1 1 1 1	Trivial
6	3 4 1 2	4 3 2 1	0 0 1 1	Trivial
7	2 1 3 4	4 3 1 2	0 0 0 0	Trivial
8	3 4 2 1	1 2 4 3	1 1 0 0	Trivial
9	3 4 2 1	4 3 2 1	0 0 1 1	Trivial
10	2 1 4 3	4 3 2 1	0 0 0 0	Trivial
11	3 4 1 2	1 2 4 3	1 1 0 0	Trivial
12	2 4 3 1	1 2 3 4	1 1 1 1	Trivial
13	3 4 2 1	4 3 1 2	0 0 1 1	Trivial
14	3 4 1 2	1 2 3 4	1 1 0 0	Trivial
15	2 3 4 1	1 2 4 3	1 1 1 1	Trivial
16	2 1 3 4	4 3 2 1	0 0 0 0	Trivial
17	3 2 4 1	4 3 1 2	0 0 0 0	Trivial
18	2 4 1 3	4 3 2 1	0 0 1 1	Trivial
19	2 1 4 3	1 2 3 4	1 1 1 1	Trivial
20	3 4 1 2	4 1 2 3	1 1 0 0	Trivial
21	2 1 3 4	1 2 4 3	1 1 1 1	Trivial
22	2 3 1 4	4 3 2 1	0 0 1 1	Trivial
23	3 4 2 1	4 1 2 3	1 1 0 0	Trivial
24	3 1 4 2	4 3 2 1	0 0 0 0	Trivial
Test Block 1				
1	3 4 1 2	2 3 4 1	1 0 0 1	I
2	3 4 1 2	3 2 1 4	0 1 1 0	III

(*continued*)

Position in series	Payoffs to Player I	Payoffs to Player II	Quadruplet	Game type
	A B C D	A B C D	*m p M P*	
3	3 4 2 1	3 2 1 4	0 0 1 1	Catch
4	3 2 1 4	4 2 1 3	0 1 0 1	IV
5	3 1 2 4	1 3 4 2	1 0 1 0	II
6	3 4 1 2	4 2 3 1	1 0 0 1	I
7	3 2 1 4	1 3 2 4	0 1 0 1	IV
8	3 1 2 4	3 2 4 1	1 0 1 0	II
9	3 1 4 2	4 2 3 1	1 1 1 1	Catch
10	3 4 1 2	2 3 1 4	0 1 1 0	III
11	3 2 1 4	3 2 4 1	1 0 1 0	II
12	3 1 2 4	2 3 1 4	0 1 0 1	IV
13	3 2 4 1	1 3 2 4	0 0 0 0	Catch
14	3 4 1 2	3 2 4 1	1 0 0 1	I
15	3 4 1 2	1 3 2 4	0 1 1 0	III
16	3 1 2 4	4 2 1 3	0 1 0 1	IV
17	3 2 1 4	2 3 4 1	1 0 1 0	II
18	3 4 1 2	4 2 1 3	0 1 1 0	III
19	3 4 2 1	1 3 4 2	1 1 0 0	Catch
20	3 4 1 2	1 3 4 2	1 0 0 1	I
Test Block 2				
1	2 4 1 3	4 2 1 3	0 1 1 1	III
2	2 1 3 4	1 3 4 2	1 0 1 0	I
3	2 3 1 4	2 3 4 1	1 0 1 1	II
4	2 1 4 3	4 2 1 3	0 0 0 0	Catch
5	2 1 3 4	3 2 1 4	0 1 0 1	IV
6	2 3 1 4	1 3 4 2	1 0 1 1	II
7	2 1 3 4	2 3 4 1	1 0 1 0	I
8	2 3 1 4	4 2 1 3	0 1 1 1	III
9	2 1 4 3	2 3 1 4	0 0 0 0	Catch
10	2 1 3 4	1 3 2 4	0 1 0 1	IV
11	2 1 3 4	4 2 3 1	1 0 1 0	I
12	2 4 1 3	3 2 4 1	1 0 1 1	II
13	2 1 4 3	1 3 2 4	0 0 0 0	Catch
14	2 1 3 4	4 2 1 3	0 1 0 1	IV
15	2 4 1 3	2 3 1 4	0 1 1 1	III
16	2 1 3 4	3 2 4 1	1 0 1 0	I
17	2 1 3 4	2 3 1 4	0 1 0 1	IV
18	2 1 4 3	3 2 1 4	0 0 0 0	Catch
19	2 4 1 3	1 3 2 4	0 1 1 1	III
20	2 3 1 4	4 2 3 1	1 0 1 1	II

References

Aaftink, J. (1989). Far-sighted equilibria in 2 × 2, non-cooperative, repeated games. *Theory and Decision*, *16*, 175–192.

Aumann, R. J. (1995). Backward induction and common knowledge of rationality. *Games & Economic Behavior*, *8*, 6–19.

Baron-Cohen, S. (1989). The autistic child's theory of mind: a case of specific developmental delay. *Journal of Child Psychology and Psychiatry*, *30*, 285–297.

Brams, S. J. (1994). *Theory of moves*. Cambridge: Cambridge University Press.

Brams, S. J., & Kilgour, D. M. (1998). Backward induction is not robust: the parity problem and the uncertainty problem. *Theory and Decision*, *45*, 263–289.

Brams, S. J., & Wittman, D. (1981). Nonmyopic equilibria in 2×2 games. *Conflict Management and Peace Science*, *6*, 36–62.

Carbone, E., & Hey, J. D. (2001). A test of the principle of optimality. *Theory and Decision*, *50*, 263–281.

Colman, A. M. (1982). *Game theory and experimental games*. Oxford: Pergamon Press.

Colman, A. M., & Bacharach, M. (1997). Payoff dominance and the Stackelberg heuristic. *Theory and Decision*, *43*, 1–19.

Colman, A. M., & Stirk, J. A. (1998). Stackelberg reasoning in mixed-motive games: an experimental investigation. *Journal of Economic Psychology*, *19*, 279–293.

Estes, D., Wellman, H. M., & Woolley, J. D. (1989). Children's understanding of mental phenomena. In H. W. Reese (Ed.), *Advances in child development and behavior* (pp. 41–87), Vol. 22. San Diego, CA: Academic Press.

Evans, J. (1987). Beliefs and expectations as causes of judgmental bias. In G. Wright (Ed.), *Judgmental forecasting* (pp. 31–47). Chichester: Wiley.

Evans, J. (1993). The cognitive psychology of reasoning: an introduction. *Quarterly Journal of Experimental Psychology*, *46A*, 561–567.

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: domain specificity in cognition and culture* (pp. 257–293). Cambridge: Cambridge University Press.

Halford, G. S. (1993). *Children's understanding: the development of mental models*. Hillsdale, NJ: Lawrence Erlbaum.

Happe, F., Malhi, G. S., & Checkly, S. (2001). Acquired mind-blindness following frontal lobe surgery? A single case study of impaired 'theory of mind' in a patient treated with stereotactic anterior capsulotomy. *Neuropsychologia*, *39*, 83–90.

Happe, F., Winner, E., & Brownell, H. (1998). The getting of wisdom: theory of mind in old age. *Developmental Psychology*, *34*, 358–362.

Johnson-Laird, P. N. (1983). *Mental models: towards a cognitive science of language, inference, and consciousness*, Cambridge, MA: Harvard University Press.

Kilgour, D. M. (1984). Equilibria for far-sighted players. *Theory and Decision*, *16*, 135–157.

Kuhn, H. W. (1953). Extensive games and the problem of information. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games* (pp. 193–216), Vol. 2. Princeton, NJ: Princeton University Press.

Legrenzi, P., & Girotto, V. (1996). Mental models in reasoning and decision-making processes. In J. Oakhill & A. Garnham (Eds.), *Mental models in cognitive science* (pp. 95–117). Hove: Psychology Press.

Legrenzi, P., Girotto, V., & Johnson-Laird, P. N. (1993). Focusing in reasoning and decision-making. *Cognition*, *49*, 37–66.

Luce, R. D., & Raiffa, H. (1957). *Games and decisions: introduction and critical survey*. New York: Dover.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, *5*, 349–357.

Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory*. Cambridge, MA: MIT Press.

Perner, J. (1979). Young children's preoccupation with their own payoffs in strategic analysis of 2×2 games. *Developmental Psychology*, *15*, 204–213.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: Bradford Books/MIT Press.

Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that…" Attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology*, *39*, 437–471.

Rapoport, A., Guyer, M. J., & Gordon, D. G. (1976). *The 2 × 2 game*, Ann Arbor, MI: University of Michigan Press.

Rowe, A. D., Bullock, P. R., Polkey, C. E., & Morris, R. G. (2001). 'Theory of mind' impairments and their relationship to executive functioning following frontal lobe excisions. *Brain*, *124*, 600–616.

Sabbagh, M. A., & Taylor, M. (2000). Neural correlates of theory-of-mind reasoning: an event-related potential study. *Psychological Science*, *11*, 46–50.

Saltzman, J., Strauss, E., Hunter, M., & Archibald, S. (2000). Theory of mind and executive functions in normal human aging and Parkinson's disease. *Journal of the International Neuropsychological Society*, *6*, 781–788.

Samuels, M. C., Brooks, P. J., & Frye, D. (1996). Strategic game playing in children through the windows task. *British Journal of Developmental Psychology*, *14*, 159–172.

Stanovich, K. E., & West, R. F. (1998). Cognitive ability and variation in selection task performance. *Thinking and Reasoning*, *4*, 193–230.

Stuss, D. T., Gallup, G. G., & Alexander, M. P. (2001). The frontal lobes are necessary for 'theory of mind'. *Brain*, *124*, 279–286.

Surian, L., & Siegal, M. (2001). Sources of performance on theory of mind tasks in right hemisphere-damaged patients. *Brain & Language*, *78*, 224–232.

Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.

Wellman, H. M. (1993). Early understanding of mind: the normal case. In S. Baron-Cohen, H. Tager-Flusberg & D. Cohen (Eds.), *Understanding other minds: perspectives from autism* (pp. 10–39). Oxford: Oxford University Press.

Woolley, J. D., & Wellman, H. M. (1993). Origin and truth: young children's understanding of imaginary mental representations. *Child Development*, *64*, 1–17.

Zagare, F. C. (1984). Limited-move equilibrium in 2×2 games. *Theory and Decision*, *16*, 1–19.