# A Window of Cognition: Eyetracking the Reasoning Process in Spatial Beauty Contest Games[*]

Chun-Ting Chen, Chen-Ying Huang and Joseph Tao-yi Wang

February 5, 2013

## Abstract

We study the reasoning process people utilize to reach a decision in an environment where final choices are well understood, the associated theory is procedural, and the decision-making process is observable. In particular, we introduce a two-person "beauty contest" game played spatially on a two-dimensional plane. Players choose locations and are rewarded by hitting "targets" dependent on opponents' locations. By tracking subjects' eye movements (termed the lookups), we infer their reasoning process and classify subjects into various types based on a level-$k$ model. More than half of the subjects' classifications coincide with their classifications using final choices alone, supporting a literal interpretation of the level-$k$ model for subject's reasoning process. When choice data is noisy, lookup data could provide additional separation of types.

**Keywords** beauty contest game, level-$k$ model, best response hierarchy, cognitive hierarchy

**JEL** C91, C72, D87

[*] Department of Economics, National Taiwan University, 21 Hsu-Chow Road, Taipei 100, Taiwan. Chen: r94323016@ntu.edu.tw, cuc230@psu.edu; Huang: chenying@ntu.edu.tw; Wang: josephw@ntu.edu.tw (corresponding author).Research support was provided by the National Science Council of Taiwan (grant 96-2415-H-002-006). Joseph thanks the advice, guidance and support of Colin F. Camerer. We thank Ming Hsu for valuable suggestions that direct us to eyetracking. We thank comments from Vincent Crawford, Rosemarie Nagel, Matthew Shum, Yi-Ting Chen, Shih-Hsun Hsu, Ching-Kang Ing, Chung-Ming Kuan, and the audience of the ESA 2008 International Meeting and North American Region Meeting, TEA 2008 Annual Meeting, 2009 Stony Brook Workshop on Behavioral Game Theory, AEA 2010 Annual Meeting, 2010 KEEL conference, the 12th BDRM conference and the 2010 World Congress of the Econometric Society.

# I Introduction

Since Samuelson [1938] developed the theory of revealed preferences, economic theory has been focusing on interpreting people's observed choices as directly reflecting their personal preferences, usually unobserved by outsiders. Based on the theoretical predictions, empirical researchers then collect data either from natural occurring or controlled environments, and construct econometric models to analyze it. The revealed preference approach has achieved tremendous success by simply assuming utility optimization. Nonetheless, this focus on final choices (and the preferences they reflect) does not exclude the possibility of analyzing the decision-making process in the middle. Just as modern theories of the firm open up the black box of profit maximization and explore the effect of contracts and organizational structures within the firm, there is no reason why economic theory cannot consider the reasoning process prior to the final decision, especially when it is potentially observable and can help make better predictions.

In many cases, the economic theory could potentially suggest a procedure by which people calculate and reason to determine what is the best. When economic theories provide clear predictions on the underlying decision-making process, it is natural to ask whether one could test these predictions using some form of empirical data. For example, in extensive form games, subgame perfect equilibrium is typically solved by backward induction, a procedure that can be carried out (and therefore tested) step-by-step by players of the game. Hence, Camerer et al. [1993] and Johnson et al. [2002] employ a mouse-tracking technology called "mouselab" to test predictions of backward induction, and find evidence against it even in three-stage bargaining games. In addition to testing predictions, one could also use a procedural theory to analyze how different reasoning processes can lead to systematically different behavior. For example, Krajbich, Armel and Rangel [2010] consider an attentional drift-diffusion model and demonstrate how different decision thresholds can lead to specific premature choices in an individual decision-making problem. More recently, Koszegi and Szeidl [2013] consider the possibility that people focus on certain attributes of available options, and hence, become prone to present bias and time inconsistency problems.

In this paper, we attempt to study the reasoning process as well as final choices in a game-theoretic environment. In particular, we consider the reasoning process people utilize to reach a decision, in which they perform different levels of strategic reasoning. To conduct this alternative research strategy of studying the decision-making process, there are three important requirements on the task to use. First, we need a setting in which final choices are well understood and mature theories exist to explain how choices are made. This is because if there is still no consensus regarding which theory best explains final choices and why, it is conceivably harder to come up with satisfactory hypotheses on reasoning processes to base tests on. Secondly, to make a plausible hypothesis on reasoning, we want the associated economic theory to be more procedural. In other words, there is room that if the theory is taken literally, it makes predictions on not only choices, but also a particular reasoning process that leads to the final choice. Finally, we require some data collection method that will allow us to observe the reasoning process and for that purpose the task used has to suit the method.

We design a new set of games, termed two-person spatial beauty contest games, to analyze individual's reasoning process by observing lookup patterns with video-based eyetracking, meeting all three requirements as follows. This new set of games, as its name suggests, is essentially a graphical simplification of the $p$-beauty contest games for two players.[1] It is known that initial responses in the $p$-beauty contest games can be well explained by theories of heterogeneous levels of rationality such as the level-$k$ model.[2] Since level-$k$ models can predict choices well in these guessing games, the first requirement that mature theory exists to explai final choices is met. Logically the next question should be on whether they can also predict the reasoning processes. A key in the level-$k$ model is that players of higher levels of rationality best respond to players of lower levels, who in turn best respond to players of even lower levels and so on. This best response procedural hierarchy is the perfect candidate for modeling the reasoning process of a subject prior to making the final choice, since in a two-person game, the final choice should be a best response to the subject's belief regarding the other player's choice, which in turn is a best response to the subject's belief about the other player's belief about her choice, and so on.[3] In other words, to figure out which choice to make, a subject has to go through a particular best response hierarchical procedure. Thus, the second requirement is squarely met since by taking the level-$k$ model procedurally, one can come up with a natural hypothesis regarding the reasoning process. Lastly, the graphical representation of the spatial beauty contest games induces subjects to go through this hierarchical procedure of best responses by counting on the computer screen (instead of reasoning in their minds), leaving footprints that the experimenter can trace, and thus the third requirement is met.

We eyetrack each subject's reasoning process by recording the entire sequence of locations she looks at. In other words, we record not only her final choice, but also every

[1] Nagel [1995], Ho, Camerer and Weigelt [1998] studied the $p$-beauty contest game. Variants of two-person guessing games are studied by Costa-Gomes and Crawford [2006] and Grosskopf and Nagel [2008]. However, unlike the two person guessing game considered in Grosskopf and Nagel [2008], choosing the boundary is *not* a dominant strategy in our spatial beauty contest game.

[2] Level-$k$ models are proposed and applied by Stahl and Wilson [1995], Nagel [1995], and Costa-Gomes and Crawford [2006]. A related model, the cognitive hierarchy model is proposed by Camerer, Ho and Chong [2004].

[3] To avoid confusion, the subject is denoted by her while her opponent is denoted by him.

location the subject has ever fixated at in an experimental trial real-time. Following the convention, we call this real-time fixation data the "lookups" even though there is really nothing to be looked up in our experiment. When a subject reasons through a particular best response hierarchy, designated by her level-$k$ type, each step of thinking is characterized as a "state." To describe changes between the thinking states of a subject, we construct a constrained Markov-switching model between these states. Eye fixations conditional on each thinking state are then modeled to allow for logit errors due to imprecise eyetracking or peripheral vision. We classify subjects into various level-$k$ types based on maximum likelihood estimation using individual lookup data. Moreover, we adopt an empirical likelihood ratio test for non-nested but overlapping models proposed by Vuong [1989] to ensure the distinctive separation of the estimated type from other competing types. Results show that among the seventeen subjects we tracked, one follows the level-0 ($L0$) best response hierarchy the closest with her lookups, six follow the level-1 ($L1$) hierarchy, four follow the level-2 ($L2$) hierarchy, another four follow the level-3 ($L3$) hierarchy, and the remaining two follow the equilibrium ($EQ$) best response hierarchy, which coincides with level-4 ($L4$) hierarchy in most games of our experiment. Treating the $EQ$ type as having a thinking step of 4, the average thinking step is 2.00, in line with results of other $p$-beauty contest games.

If the level-$k$ model can predict not only choices but also reasoning processes well, the estimated level of a player when we analyze her lookups should coincide with her level when we analyze her choices alone since $k$ reflects her strategic sophistication. To check whether the lookup data indeed align well with choice data, we classify subjects by using their final choice data only. We find that choice-based and lookup-based classifications are pretty consistent, classifying ten of the seventeen subjects as the same type. Consistency between choice-based and lookup-based classifications suggests that for a high percentage of subjects, if their lookups are classified as a particular level-$k$ type, their final choices follow the prediction of that level-$k$ type as well. This is a strong support to a literal interpretation of the level-$k$ model to explain subjects' reasoning process and final choice altogether in the spatial beauty contest game. It means that the corresponding best response hierarchy implied by each level-$k$ type is literally carried out by subjects.

We look further into the subtle difference between lookup and choice data even though for the majority of subjects they align well. Among the seven subjects whose two classifications differ, for all but one subject, the choice-based level-$k$ types are not robust to a (nonparametric) bootstrap procedure, having a misclassification rate of at least 18% if one resamples the choice data and performs the same estimation. On the other hand, for the ten subjects whose two classifications are the same, the average misclassification rate is less than 5%. The difference is significant, having a $p$-value of 0.0123 according to the Mann-Whitney-Wilcoxon rank sum test. In other words, when the two classifications differ, it is when the choice data is noisy. When the two classifications agree, choice data is quite robust. This leaves open the possibility that lookup data may help classify subjects more sharply since when they differ, choice data is noisy and thus there is room to improve choice estimation.

Even when the level based on lookups and that based on choices differ, the level based on lookups does a reasonable job in predicting choices and is thus a viable alternative to the choice-based type. In fact, for six out of seven subjects whose two classifications differ, their types based on analyzing lookups predict final choices reasonably well, ranking second in terms of likelihood.[4] According to a bootstrap procedure, their lookup-based types are also the second most successful types in predicting choices. Moreover, we demonstrate how lookups indeed provide better classification when choice-based estimation is not robust through an out-of-sample prediction exercise. We estimate the models with 2/3 of the trials and predict the final choices of the remaining trials on the nine subjects whose final choices are not robust according to the bootstrap procedure. We show that the lookup-based model is superior in terms of both mean square errors and economic value (Camerer, Ho and Chong, 2004). To sum up, when the classifications based on lookups and choices differ, the lookup type predicts choices reasonably well. Moreover, when the choice data is noisy, we can predict the later choices of a subject better by her earlier lookup data than by her earlier choice data. In other words, looking into players' reasoning process gives us valuable information if we are to classify them properly.[5]

In the related literature, some experimental studies do attempt to investigate "information search" patterns in games, in order to capture part of the reasoning process. In addition to Camerer et al. [1993] and Johnson et al. [2002], Costa-Gomes, Crawford and Broseta [2001] and Costa-Gomes and Crawford [2006] also employ the mouse-tracking technology "mouselab" to study payoff lookups in normal form games and information search in two-person guessing games. Gabaix, Laibson, Moloche and Weinberg [2006] also use mouselab to observe information acquisition and analyze aggregate information search patterns to test a heuristic "directed cognition" model. More recently, Wang, Spezio and Camerer [2010] employ eyetracking to observe the decision-making process of a deceptive sender in sender-receiver games. In all these studies some information must be withheld, and "looked-up" by subjects during the experiment. Hence, these studies rely on information search to infer certain stages of the reasoning process, instead of directly observing the entire process itself. Our paper differs from these previous attempts by observing

[4] The last subject's type based on lookups ranked third. The most successful type is of course the one based on analyzing choices.

[5] Even if we focus on the seven subjects whose two classifications differ, the lookup-based model is still superior in terms of mean square errors and is comparable in economic value.

lookup patterns when there is no explicit hidden information to be acquired. We directly observe the reasoning process instead of making an inference on it. To the best of our knowledge, this is the first paper analyzing the reasoning process directly and comparing it with final choice. Specifically, it is the graphical feature of our design that makes direct observations of reasoning processes possible. This points to the importance of tailoring games for tracking decision-making. The structure of the $p$-beauty contest games implies a best response hierarchy of reasoning which can be fully exploited in our spatial design. In other less-structured games, some viable hypotheses concerning the reasoning process have to be formed and specific designs have to be tailor made so that these reasoning processes can be directly observed. This leaves open an interesting direction for future research.[6]

The remaining of the paper is structured as follows: Section A describes the spatial beauty contest game and its theoretical predictions; Section B describes details of the experiment; Section III reports aggregate statistics on lookups; Section IV reports classification results from the Markov-switching model based on lookups; Section V compares classification results with those based on final choices alone. Section VI concludes.

# II The Experiment

## A The Spatial Beauty Contest Game

We now introduce our design, the equilibrium prediction, the prediction by the level-$k$ model and formulate the hypotheses which will be tested. To create a spatial version of the $p$-beauty contest game, we reduce the number of players to two, so that we can display the action space of all players on the computer screen visually. Players choose locations (instead of numbers) simultaneously on a 2-dimensional plane attempting to hit one's target location determined by the opponent's choice. The target location is defined as a relative location to the other player's choice of location by a pair of coordinates $(x, y)$. We use the standard Euclidean coordinate system. For instance, $(0, -2)$, means the target location of a player is "two steps below the opponent," and $(-4, 0)$ means the target location of a player is "four steps to the left of the opponent." These targets are common knowledge to the players. Payoffs are determined by how "far" (the sum of horizontal distance and vertical distance) a player is away from the target. The larger this distance is, the lower her payoff is. Players can only choose locations on a given grid

[6] Several recent level-$k$ papers estimate population mixture models to infer the fraction of level-$k$ types within the population (Burchardi and Penczynski [2011]). Instead of investigating the population mixture of types, we focus on how well individual lookup patterns correspond to a particular level-$k$ best response hierarchy in an environment where we already know the level-$k$ model predicts aggregate subject behavior fairly well.

map, though one's target may fall outside if the opponent is close to or on the boundary.[7] For example, consider the 7 × 7 grid map in Figure I. For the purpose of illustration, suppose a player's opponent has chosen the center location labeled O ($(0,0)$) and the player's target is $(-4,0)$. Then to hit her target, she has to choose location $(-4,0)$. But location $(-4,0)$ is not on the map, while choosing location $(-3,0)$ is optimal among all 49 feasible choices because location $(-3,0)$ is the only feasible location that is one step from location $(-4,0)$.[8]

The spatial beauty contest game is essentially a spatial version of Costa-Gomes and Crawford [2006]'s asymmetric two-person guessing games, in which one subject would like to choose $\alpha$ of her opponent's choice and her opponent would like to choose $\beta$ of her choice. Hence, similar to Costa-Gomes and Crawford [2006], the equilibrium prediction of this spatial beauty contest game is determined by the targets of both players. For example, if the targets of the two players are $(0,2)$ and $(4,0)$ respectively, the equilibrium consists of both players choosing the Top-Right corner of the map. This conceptually coincides with a player hitting the lower bound in the two-person guessing game of Costa-Gomes and Crawford [2006] where $\alpha\beta$ is less than 1, or all choosing zero in the $p$-beauty contest game where $p$ is less than 1.[9] Note that in general the equilibrium need not be at the corner since targets can have opposite signs. For example, when the targets are $(4,-2)$ and $(-2,4)$ played on a $7\times7$ grid map, the equilibrium locations for the two players are both two steps away from the corner (labeled as E1 and E2 for the two players respectively in Figure I).

We derive the equilibrium predictions for the general case as follows. Formally, consider a spatial beauty contest game with targets $(a_1,b_1)$ and $(a_2,b_2)$. With some abuse of notation, suppose player $i$ chooses location $(x_i,y_i)$ on a map $G$ satisfying $(x_i,y_i)\in G\equiv\{-X,-X+1,...,X\}\times\{-Y,-Y+1,...,Y\}$ where $(0,0)$ is the center of the map. For instance, $(x_i,y_i)=(X,Y)$ means player $i$ chooses the Top-Right corner of the map. The other player $-i$ also chooses a location $(x_{-i},y_{-i})$ on the same map: $(x_{-i},y_{-i})\in G$. The payoff to player $i$ in this game is:

$$p_i(x_i,y_i;x_{-i},y_{-i};a_i,b_i)=\bar{s}-(|x_i-(x_{-i}+a_i)|+|y_i-(y_{-i}+b_i)|)$$

where $\bar{s}$ is a constant. Notice that payoffs are decreasing in the number of steps a player is away from her target, which in turn depends on the choice of the other player. There is no

[7] Similar designs of 3 × 3 games could also be found in Kuo et al. [2009]. They addressed different issues.

[8] For instance, to go from location $(-3,1)$ to $(-4,0)$, one has to travel one step left and one step down and hence the distance is 2.

[9] However, choosing the Top-Right corner is *not* a dominant strategy, unlike in the symmetric two-person guessing game analyzed by Grosskopf and Nagel [2008].

interaction between the choices of $x_i$ and $y_i$. Hence the maximization can be obtained by choosing $x_i$ and $y_i$ separately to minimize the two absolute value terms. We thus consider the case for $x_i$ only. The case for $y_i$ is analogous.[10]

To ensure uniqueness, in all our experimental trials, $a_i + a_{-i} \neq 0$.[11] Without loss of generality, we assume that $a_i + a_{-i} < 0$ so that the overall trend is to move leftward.[12] Suppose $a_1 < 0$. If $a_1 a_2 < 0$, implying player 1 would like to move leftward but player 2 would like to move rightward, since the overall trend is to move leftward, it is straightforward to see that the force of equilibrium would make player 1 hit the lower bound while player 2 will best respond to that. The equilibrium choices of both, denoted by $(x_1^e, x_2^e)$, are characterized by $x_1^e = -X$ and $x_2^e = -X + a_2$.[13] If $a_1 a_2 \geq 0$, since both players would like to move leftward, they will both hit the lower bound. The equilibrium is characterized by $x_1^e = x_2^e = -X$. To summarize, when $a_1 + a_2 < 0$, only the player whose target is greater than zero will not hit the lower bound. Therefore, as a spatial analog to Observation 1 of Costa-Gomes and Crawford [2006], we obtain:

**Proposition 1**

In a spatial beauty contest game with targets $(a_1, b_1)$ and $(a_2, b_2)$ where two players each choose a location $(x_i, y_i) \in G$ satisfying $G \equiv \{-X, -X+1, ..., X\} \times \{-Y, -Y+1, ..., Y\}$, $-2X \leq a_1, a_2 \leq 2X$ and $-2Y \leq b_1, b_2 \leq 2Y$, the equilibrium choices $(x_i^e, y_i^e)$ are characterized by: ($I\{\cdot\}$ is the indicator function)

$$\begin{cases} x_i^e = -X + a_i \cdot I\{a_i > 0\} & \text{if } a_i + a_{-i} < 0 \\ x_i^e = X + a_i \cdot I\{a_i < 0\} & \text{if } a_i + a_{-i} > 0 \end{cases}$$

and

$$\begin{cases} y_i^e = -Y + b_i \cdot I\{b_i > 0\} & \text{if } b_i + b_{-i} < 0 \\ y_i^e = Y + b_i \cdot I\{b_i < 0\} & \text{if } b_i + b_{-i} > 0 \end{cases}$$

In addition to the equilibrium prediction, one may also specify various level-$k$ predictions. First, we need to determine the anchoring $L0$ player who is non-strategic or

[10] As an illustrative example, consider $a_1 = -2$ and $a_2 = +1$, indicating that player 1 wants to be two steps to the left of player 2, while player 2 wants to be one step to the right of player 1.

[11] Suppose $a_1 = -2$ and $a_2 = +2$. Any location where player 1 is two steps to the left of player 2 is an equilibrium since player 2 is then two steps to the right of player 1. Note that this corresponds to the case where $\alpha\beta = 1$ in the two-person guessing game of Costa-Gomes and Crawford [2006]. If $a_i = -a_{-i} = a$, any feasible $x_i, x_{-i}$ satisfying $x_i - x_{-i} = a$ constitutes an equilibrium.

[12] In the illustrative example of $a_1 = -2$ and $a_2 = +1$, $(-2) + 1 < 0$. Due to symmetry, all other cases are isomorphic to this case.

[13] In the illustrative example of $a_1 = -2$ and $a_2 = +1$, the equilibrium is $(x_1^e, x_2^e) = (-X, -X+1)$. We impose $a_i \leq 2X$ for all games in the experiment, thus we do not need to worry about the possibility that $x_i^e$ lies outside the upper bound $X$ (i.e., $x_i^e = -X + a_i > X$). In general, if $a_i > 2X$, player $i$ would hit the upper bound and thus $x_i^e = X$. Similarly, we assume $-2X \leq a_i$, so we need not worry about the possibility that $x_i^e$ lies outside the lower bound $-X$ (i.e., $x_i^e = X + a_i < -X$).

naïve. This is usually done by assuming players choosing randomly.[14] In a spatial setting, Reutskaja et al. [2011] find the center location focal, while Crawford and Iriberri [2007*a*] define $L0$ players as being drawn toward focal points in the non-neutral display of choices. In addition, due to a drift-correction procedure of the eyetracker (fixating on a dot at the center and hitting a button or key) prior to every trial, the center location is the first fixation of every trial. Therefore, a natural assumption here is that an $L0$ player will either choose any location on the map randomly (according to the uniform distribution), which is on average the center $(0, 0)$, or will simply choose the center. An $L1$ player $i$ with target $(a_i, b_i)$ would best respond to an $L0$ opponent who either chooses the center on average or exactly chooses the center, and as a von Neumann-Morgenstern utility maximizer, would choose the same location against these two opponents.[15] If an $L0$ player chooses (on average) the center, to best respond, an $L1$ player would choose the location $(a_i, b_i)$ unless $X$, $Y$ is too small so that it is not feasible.[16] Similarly, for an $L2$ opponent $j$ with the target $(a_j, b_j)$ to best respond to an $L1$ player $i$ who chooses $(a_i, b_i)$, he would choose $(a_i + a_j, b_i + b_j)$ when $X$, $Y$ is large enough. Repeating this procedure, one can determine the best responses of all higher level-$k$ ($Lk$) types. Figure I shows the various level-$k$ predictions of a $7 \times 7$ spatial beauty contest game for two players with targets $(4, -2)$ and $(-2, 4)$.

To account for the possibility that one's target may fall outside the map, we define the adjusted choice $R(X, Y; (x, y))$. Formally, the adjusted choice is given by

$$R(X, Y; (x, y)) \equiv (\min\{X, \max\{-X, x\}\}, \min\{Y, \max\{-Y, y\}\}).$$

In words, if the ideal best response which hits the target is location $(x, y)$, the adjusted choice $(\tilde{x}, \tilde{y}) \equiv R(X, Y; (x, y))$ gives us the closest feasible location on the map so the choice $(\tilde{x}, \tilde{y})$ is constrained to lie within the range $\tilde{x} \in \{-X, -X+1, ..., X\}$, $\tilde{y} \in \{-Y, -Y+1, ..., Y\}$. This adjusted choice is the best feasible choice on the map since payoffs are decreasing in the distance between the ideal best response (target) and the final choice. Moreover, as shown in Supplementary Appendix A2, since the grid map is of a finite size, eventually when $k$ for a level-$k$ type is large enough, the $Lk$ prediction will coincide with the equilibrium. To summarize, we have

**Proposition 2**

[14] See Costa-Gomes, Crawford and Broseta [2001], Camerer, Ho and Chong [2004], Costa-Gomes and Crawford [2006] and Crawford and Iriberri [2007*b*].

[15] See proof in Supplementary Appendix A1.This is true because our payoff structure is point symmetric by $(0, 0)$ over the grid map. Hence, it makes no difference for an $L1$ opponent whether we assume an $L0$ player chooses exactly the center, or randomly (on average the center). In our estimation, we assume $L0$ chooses the center but incorporates random $L0$ as a special case (when the logit parameter is zero).

[16] In this case, an $L1$ player would choose the closest feasible location.

Consider a spatial beauty contest game with targets $(a_1, b_1)$ and $(a_2, b_2)$ where two players choose locations $(x_1, y_1)$, $(x_2, y_2)$ satisfying $(x_i,y_i) \in G \equiv \{-X, -X+1, ..., X\} \times \{-Y, -Y+1, ..., Y\}$, $-2X \leq a_1, a_2 \leq 2X$ and $-2Y \leq b_1, b_2 \leq 2Y$. Denote the choice of a level-$k$ player $i$ by $(x_i^k, y_i^k)$, then $(x_1^0, y_1^0) = (x_2^0, y_2^0) \equiv (0, 0)$ and

1. $(x_i^k, y_i^k) = R\left(X, Y; (a_i + x_{-i}^{k-1}, b_i + y_{-i}^{k-1})\right)$ for $k = 1, 2, ...$
2. there exists a smallest positive integer $\overline{k}$ such that for all $k \geq \overline{k}$, $(x_i^k, y_i^k) = (x_i^e, y_i^e)$.

Proof.
See Supplementary Appendix A2.

In Table I we list all the 24 spatial beauty contest games used in the experiment, their various level-$k$ predictions, equilibrium predictions and the minimum $\overline{k}$'s. Notice that in the first 12 games, targets of each player are 1 dimensional while in the last 12 games, targets are 2 dimensional. Also, Games $(2m-1)$ and $(2m)$ (where $m = 1, 2, \ldots, 12$) are the same but with reversed roles of the two players, so for instance, Games 1 and 2 are the same, Games 3 and 4 are the same, etc.

The $\overline{k}$'s for our 24 games are almost always 4, but some are 3 (Games 1, 10, 17), 5 (Games 5, 11, 12) or 6 (Game 6). This indicates that as long as we include level-$k$ types with $k$ up to 3 and the equilibrium type, we will not miss the higher level-$k$ types much since higher types coincide with the equilibrium most of the time. Moreover, as evident in Table I, different levels make different predictions. In other words, various levels are strongly separated on the map.[17] The level-$k$ model predicts what final choices are made for each level $k$. This is formulated in Hypothesis 1.

**Hypothesis 1 (Final Choice)** *Consider a series of one-shot spatial beauty contest games without feedback, $n = 1, 2, \ldots, N$, each with targets $(a_{1,n}, b_{1,n})$ and $(a_{2,n}, b_{2,n})$ where two players choose locations $(x_{1,n}, y_{1,n})$, $(x_{2,n}, y_{2,n})$ satisfying $(x_{i,n}, y_{i,n}) \in G_n \equiv \{-X_n, -X_n + 1, \cdots, X_n\} \times \{-Y_n, -Y_n + 1, \cdots, Y_n\}$, $-2X_n \leq a_{1,n}, a_{2,n} \leq 2X_n$, and $-2Y_n \leq b_{1,n}, b_{2,n} \leq 2Y_n$. A level-k subject i's choice for game n, denoted $(x_{i,n}^k, y_{i,n}^k)$ is $(x_{i,n}^k, y_{i,n}^k) = R(X_n, Y_n; (a_{i,n} + x_{-i,n}^{k-1}, b_{i,n} + y_{-i,n}^{k-1}))$ as defined in Proposition 2, and this k is constant across games.*

Since our games are spatial, players can literally count using their eyes how many steps on the map they have to move to hit their targets. Thus, a natural way to use lookups is to take the level-$k$ reasoning processes literally in the following sense. Take an $L2$ player as an example, the level-$k$ model implies that she best responds to an $L1$ opponent, who in turn best responds to an $L0$. Therefore, for the $L2$ player to make a final choice, she

[17] The only exceptions are $L3$ and $EQ$ in Games 1, 10, 17, $L2$ and $L3$ in Games 2, 6, 9, and $L2$ and $EQ$ in Game 18. See the underlined predictions in Table I.

has to first figure out what an $L0$ would choose since her opponent thinks of her as an $L0$. She then needs to figure out what her opponent, an $L1$, would choose. Finally, she has to make a choice as an $L2$. It is possible that this process is carried out solely in the mind of a player. Yet since the games are spatial, one can simply figure all these out by looking at and counting on the map. This has the advantage of reducing much memory load and being much more straightforward. If this hypothesis is true, an $L2$ player would look at the center (where an $L0$ player would choose), her opponent's $L1$ choice and her own final choice as an $L2$. In other words, the hotspots of an $L2$ player in her lookups would consist of these three locations on the map. This is probably the most natural prediction on the lookup data one can make when the underlying model is the level-$k$ model. Hence we formulate Hypothesis 2 and base our econometric analysis of lookups on this.

**Hypothesis 2 (Lookup)** *Consider a series of one-shot spatial beauty contest games with targets $(a_{1,n}, b_{1,n})$ and $(a_{2,n}, b_{2,n})$ where two players choose locations $(x_{1,n}, y_{1,n})$, $(x_{2,n}, y_{2,n})$ satisfying $(x_{i,n}, y_{i,n}) \in G_n \equiv \{-X_n, -X_n+1, \cdots, X_n\} \times \{-Y_n, -Y_n+1, \cdots, Y_n\}$, $-2X_n \leq a_{1,n}, a_{2,n} \leq 2X_n$, and $-2Y_n \leq b_{1,n}, b_{2,n} \leq 2Y_n$ played without feedback. Denote the choice of a level-$k$ player $i$ by $(x^k_{i,n}, y^k_{i,n})$. Assuming one carries out the reasoning process on the map, a level-$k$ subject $i$ will also:*

*a.* **(Duration of Lookups):** *Fixate at the following locations in the level-$k$ best response hierarchy $(x^0_{\cdot,n}, y^0_{\cdot,n})$ ($L0$ player's choices), ..., $(x^{k-2}_{i,n}, y^{k-2}_{i,n})$ (own $L(k-2)$ player's choice), $(x^{k-1}_{-i,n}, y^{k-1}_{-i,n})$ (opponent $L(k-1)$ player's choice), $(x^k_{i,n}, y^k_{i,n})$ (own $Lk$ player's choice) associated with that particular $k$ longer than random.*[18]

*b.* **(Sequence of Lookups):** *Have fixation sequences for each game $n$ with many transitions from $(x^{K-1}_{-i,n}, y^{K-1}_{-i,n})$ to $(x^K_{i,n}, y^K_{i,n})$ for $K = k, k-2, ...$, and transitions from $(x^{K-1}_{i,n}, y^{K-1}_{i,n})$ to $(x^K_{-i,n}, y^K_{-i,n})$ for $K = k-1, k-3, ...$ (steps of the associated level-$k$ best response hierarchy).*

## B Experimental Procedure

We conduct 24 spatial beauty contest games (with various targets and map sizes) randomly ordered without feedback at the Social Science Experimental Laboratory (SSEL), California Institute of Technology. Each game is played twice, once on the two-dimensional grid map as shown in Figure II (which we denote as the GRAPH presentation), the other time as two one-dimensional choices chosen separately (see Figure III, denoted as the SEPARATE presentation).[19] Half of the subjects are shown the two-dimensional grid

[18] The player subscript of $(x^0_{\cdot,n}, y^0_{\cdot,n})$ is dropped since both $L0$ players choose the center.

[19] Note that these two presentations are mathematically identical. However, the GRAPH presentation allows us to trace the decision-making process through observing the lookups.

maps first in trials 1-24 and the two one-dimensional choices later in trials 25-48, while the rest are shown the two one-dimensional choices first (trials 1-24) and the maps later (trials 25-48). The results of the two presentations are quite similar, so we focus on the results of the two-dimensional presentation.[20]

In addition to recording subjects' final choices, we also employ Eyelink II eyetrackers (SR-research Inc.) to track the entire decision process before the final choice is made. The experiment is programmed using the Psychophysics Toolbox of Matlab (Brainard, 1997), which includes the Video Toolbox (Pelli, 1997) and the Eyelink Toolbox (Cornelissen et al., 2002). For every 4 milliseconds, the eyetracker records the location one's eyes are looking at on the screen and one's pupil sizes. Location accuracy is guaranteed by first calibrating subjects' eyetracking patterns (video images and cornea reflections of the eyes) when they fixate at certain locations on the screen (typically 9 points), interpolating this calibration to all possible locations, and validating it with another set of similar locations. Since there is no hidden information in this game, the main goal of eyetracking is not to record information search. Instead, the goal is to capture how subjects reason before making their decision and to test whether they think through the best response hierarchy implied by a literal interpretation of the level-$k$ model.

Before each game, a drift correction is performed in which subjects fixate at the center of the screen and hit a button (or space bar). This realigns the calibration at the center of the screen. During each game, when subjects use their eyes to fixate at a location, the eyetracker sends the current location back to the display computer, and the display computer lights up the location (real time) in red (as Figures 2 and 3 show). Seeing this red location, if subjects decide to choose that location, they could hit the space bar. Subjects are then asked to confirm their choices ("Are you sure?"). They then have a chance to confirm their choice ("YES") or restart the process ("NO") by looking at the bottom left or right corners of the screen.

In each session, two subjects were recruited to be eyetracked. Since there was no feedback, each subject was eyetracked in a separate room individually and their results were matched with the other subject at end of the experiment. Three trials were randomly drawn from the 48 trials played to be paid. Average payment is US$15.24 plus a show-up fee of US$20. A sample of the instructions can be found in the Supplementary Appendix. Due to insufficient showup of eligible subjects, three sessions were conducted with only one subject eyetracked, and their results matched with a subject from a different session. Hence, we have eyetracking data for 17 subjects.

[20] A comparison of the final choices under these two representations is shown in Supplementary Table 2. None of the subjects' two sets of final choices differ significantly.

# III Lookup Summary Statistics

We first summarize subjects' lookups to test Hypothesis 2a, namely, subjects do look at and count on the map during their reasoning process. Then, we analyze subjects' lookups with a constrained Markov-switching model to classify them into various level-$k$ types to test Hypothesis 2b. As a part of the estimation, we employ Vuong's test for non-nested but overlapping models to ensure separation between competing types.

According to Hypothesis 2a, subjects will spend more time at locations corresponding to the thinking steps of a particular best response hierarchy. We present aggregate data regarding empirical lookups for all 24 Spatial Beauty Contest games in Supplementary Figures 1 through 24. For each game, we calculate the percentage of time a subject spent on each location. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. The level-$k$ choice predictions are labeled as O, L1, L2, L3, E for each game.

If Hypothesis 2a were true, the empirical lookups would concentrate on locations predicted by the level-$k$ best response hierarchy. For some games, many big circles in Supplementary Figures 1–24 do fall on various locations corresponding to the thinking steps of the level-$k$ best response hierarchy.[21] However, there seems to be a lot of noise in the lookup data: Many locations *other than* those specified in the best response hierarchy are also looked up.

We attempt to quantify this concentration of attention. First, we define *Hit* area for every level-$k$ type as the minimal convex set enveloping the locations predicted by this level-$k$ type's best response hierarchy in game $n$. For instance, for an $L2$ subject $i$ (with opponent $-i$), the best response hierarchy consists of $(x^0_{\cdot,n}, y^0_{\cdot,n})$, $(x^1_{-i,n}, y^1_{-i,n})$, $(x^2_{i,n}, y^2_{i,n})$. Thus we can construct a minimal convex set enveloping these three locations. We then take the union of *Hit* areas of all level-$k$ types and see if subjects' lookups are indeed within the union. Figure IV shows an example of *Hit* areas for various level-$k$ types in a $7 \times 7$ spatial beauty contest game with target $(4, -2)$ and the opponent's target $(-2, 4)$ (Game 16).

Figure V shows the empirical percentage of time spent on the union of *Hit* areas, or hit time, denoted as $h_t$. Across the 24 games, average hit time is 0.62, ranging from $h_t = 0.81$ (in Game 9), to $h_t = 0.36$ (in Game 21). However, hit time depends on the

[21] However, not all locations are looked up. This is likely because the error structure of high speed video-based eyetracking is very different from the error structure of mouse-tracking (such as MouseLab). In particular, eyetrackers have imprecise spatial resolution due to imperfect calibration and peripheral vision, but little temporal error (usually 250 or more samples per second). In contrast, mouse-tracking has very precise spatial resolution for cursor locations and mouse clicks, but movements of the mouse cursor need not correspond to movements of the eye. Hybrid methods are a promising direction for future research.

size of the area. Even if subjects scan over the map uniformly, the empirical hit time would not be zero. Instead, it would be proportional to the size percentage of the union of *Hit* areas, or hit area size, denoted as $h_{as}$. To correct for this hit area size bias, we calculate Selten [1991]'s linear "difference measure of predicted success," $h_t - h_{as}$, i.e. the difference between empirical hit time and hit area size, and report it in Figure VI. Note that if subjects scan randomly over the map, the percentage of time she spends on the union of the *Hit* areas will roughly equal the hit area size. By subtracting the hit area size, we can evaluate how high the empirical hit time is compared with random scanning over the map. These measures are all positive (except for Game 21), strongly rejecting the null hypothesis of random lookups. The $p$-value of one sample t-test is 0.0001, suggesting that subjects indeed spend a disproportionately long time on the union of *Hit* areas. In fact, sometimes subjects have hit time nearly 1. For example, Figure VII shows the lookups of subject 2 in round 17, acting as a Member B. The diameter of each fixation circle is proportional to the length of each lookup. Note that these circles fall almost exclusively on the best response hierarchy of an $L2$, which is exactly her level-$k$ type (based on lookups) according to the fifth column of Table II.

To sum up, the aggregate result is largely consistent with Hypothesis 2a that subjects look at locations of the level-$k$ best response hierarchy longer than random scanning would imply, although the data is noisy. We next turn to test Hypothesis 2b and consider whether individual lookup data can be used to classify subjects into various level-$k$ types.

# IV A Markov-Switching Model for Level-$k$ Reasoning

## A The State Space

According to Hypothesis 2b, a level-$k$ type subject $i$ goes through a particular best response hierarchy associated with her level-$k$ type during the reasoning process, and carries out transitions from $\left(x_{-i,n}^{K-1}, y_{-i,n}^{K-1}\right)$ to $\left(x_{i,n}^{K}, y_{i,n}^{K}\right)$, for $K = k, k-2, \cdots$, and transitions from $\left(x_{i,n}^{K-1}, y_{i,n}^{K-1}\right)$ to $\left(x_{-i,n}^{K}, y_{-i,n}^{K}\right)$ for $K = k-1, k-3, \cdots$. Taking level-2 as an example, the two key transition steps are from $(x_{i,n}^{0}, y_{i,n}^{0})$ to $(x_{-i,n}^{1}, y_{-i,n}^{1})$, thinking as a level-1 opponent, best-responding to her as a level-0 player and from $(x_{-i,n}^{1}, y_{-i,n}^{1})$ to $(x_{i,n}^{2}, y_{i,n}^{2})$, thinking as a level-2 player, best-responding to a level-1 opponent. Hence, the reasoning process of a level-2 subject $i$ consists of three stages. First, she would fixate at $(x_{i,n}^{0}, y_{i,n}^{0})$ since she believes her opponent is level-1, who believes she is level-0. Then, she would fixate at $(x_{-i,n}^{1}, y_{-i,n}^{1})$, thinking through her opponent's choice as a level-1 best responding to a level-0. Finally, she would best respond to the belief that her opponent is a level-1 by making her choice fixating at $(x_{i,n}^2, y_{i,n}^2)$. These reasoning processes are gone through in the mind of a subject and may be reflected in her lookups.

We define each stage of the reasoning process as a state. The states are in the mind of a subject. If she is a level-2, then according to the best response hierarchy of reasoning, in her mind, there are three states. To distinguish a state regarding beliefs about self from beliefs about the opponent, if a state is about the opponent, we indicate it by a minus sign. Thus, for a level-2 player, three states, namely $s = 0$ (fixating at the location of $(x_{i,n}^0, y_{i,n}^0)$ since she thinks her opponent thinks she is a level-0), $s = -1$ (fixating at the location of $(x_{-i,n}^1, y_{-i,n}^1)$ since she thinks her opponent is a level-1), and $s = 2$ (fixating at the location of $(x_{i,n}^2, y_{i,n}^2)$ since she is a level-2), are expected to be passed through during the reasoning process of a level-2 subject. We hasten to point out that these states are in the mind of a subject. It is not the level of a player. Take a level-2 subject as an example. Her level, according to the level-$k$ model, is 2. But there are three states, $s = 0$, $s = -1$, and $s = 2$, in her mind. Which state she is in depends on what she is currently reasoning about. A level-2 subject could be at state $s = -1$ because at that point of time, she is thinking about what her opponent would choose, who is a level-1 according to the best response hierarchy. However, this state $s = -1$ is not to be confused with $k = 1$ for a level-1 subject (whose states of thinking consist of $s = -0$ and $s = 1$).

More generally, for a level-$k$ subject, define $s = k$ as the highest state indicating that she is contemplating a choice by fixating at the location $(x_{i,n}^k, y_{i,n}^k)$, best responding to an opponent of level-$(k-1)$. Imagining what an opponent of level-$(k-1)$ would do, state $s = -(k-1)$ is defined as the second highest state when her fixation is at the location $(x_{-i,n}^{k-1}, y_{-i,n}^{k-1})$ contemplating her opponent's choice by best responding to herself as a level-$(k-2)$.[22] Lower states $s = k-2, s = -(k-3), ...,$ etc. are defined similarly. Then, steps of reasoning of a subject's best response hierarchy of Hypothesis 2b (associated with a particular "$k$") can be expressed as "$0, \ldots, k-2, -(k-1), k$." We regard these $(k+1)$ steps of reasoning as the $(k+1)$ states of the mind for a level-$k$ player $i$. Hence, for a level-$k$ subject, state space $\Omega_k$ consists of all thinking steps in the best response hierarchy of this particular level-$k$ type. Thus, $\Omega_k = \{0, ..., -(k-3), k-2, -(k-1), k\}$.

## B The Constrained Markov Transition Process

To account for the transitions of states within a subject's mind, we employ a Markov-switching model by Hamilton [1989] and characterize the transition of states by a Markov transition matrix. Instead of requiring a level-$k$ subject to "strictly" obey a monotonic order of level-$k$ thinking going from lower states to higher states, we allow subjects to

[22] We use the minus sign $(-)$ to refer to players contemplating about their opponent. Note that the lowest state 0 can be about one's own or the opponent. Thus the state 0 and $-0$ should be distinguished. For the ease of exposition, we do not make this distinction and call the lowest state 0.

move back from higher states to lower states. This is to account for the possibilities that subjects may go back to double check as may be typical in experiments. However, since a level-$k$ player best responds to a level-$(k-1)$ opponent, it is difficult to imagine a subject jumping from the reasoning state of say $s = (k-2)$ to that of $s = k$ without first going through the reasoning state of $s = -(k-1)$. Thus, we restrict the probabilities for all transitions that involve a jump in states to be zero.[23]

Specifically, suppose the subject is a particular level-$k$. Let $S_t$ be the random variable representing subject's state at time $t$, drawn from the state space

$$\Omega_k = \{0, ..., -(k-3), k-2, -(k-1), k\}.$$

Let the realization of the state at time $t$ be $s_t$. Denote the state history up to time $t$ by $\mathcal{S}^t \equiv \{s_1, ..., s_{t-1}, s_t\}$.[24] Since lookups may be serially correlated, we model this by estimating a constrained Markov stationary transition matrix of states. Let the transition probability from state $S_{t-1} = s_{t-1}$ to $S_t = s_t$ be

$$\Pr(S_t = s_t | S_{t-1} = s_{t-1}) = \pi_{s_{t-1} \to s_t}. \tag{1}$$

Thus, the state transition matrices $\theta_k$ for level-$k$ types for $k \in \{0, 1, 2, 3, 4\}$ are

$$\theta_0 = (\pi_{0\to0}) = (1), \theta_1 = \begin{pmatrix} \pi_{0\to0} & \pi_{0\to1} \\ \pi_{1\to0} & \pi_{1\to1} \end{pmatrix}, \theta_2 = \begin{pmatrix} \pi_{0\to0} & \pi_{0\to-1} & 0 \\ \pi_{-1\to0} & \pi_{-1\to-1} & \pi_{-1\to2} \\ \pi_{2\to0} & \pi_{2\to-1} & \pi_{2\to2} \end{pmatrix},$$

$$\theta_3 = \begin{pmatrix} \pi_{0\to0} & \pi_{0\to1} & 0 & 0 \\ \pi_{1\to0} & \pi_{1\to1} & \pi_{1\to-2} & 0 \\ \pi_{-2\to0} & \pi_{-2\to1} & \pi_{-2\to-2} & \pi_{-2\to3} \\ \pi_{3\to0} & \pi_{3\to1} & \pi_{3\to-2} & \pi_{3\to3} \end{pmatrix}, \theta_4 = \begin{pmatrix} \pi_{0\to0} & \pi_{0\to-1} & 0 & 0 & 0 \\ \pi_{-1\to0} & \pi_{-1\to-1} & \pi_{-1\to2} & 0 & 0 \\ \pi_{2\to0} & \pi_{2\to-1} & \pi_{2\to2} & \pi_{2\to-3} & 0 \\ \pi_{-3\to0} & \pi_{-3\to-1} & \pi_{-3\to2} & \pi_{-3\to-3} & \pi_{-3\to4} \\ \pi_{4\to0} & \pi_{4\to-1} & \pi_{4\to2} & \pi_{4\to-3} & \pi_{4\to4} \end{pmatrix}.$$

Note that the upper triangle where the column number is greater than one plus the row number is restricted to zero since we do not allow for jumps.

## C From States to Lookups

When a subject is in a particular state, her reasoning will be reflected in the lookups which we can track. Recall that for each game $n$, $G_n$ is the map on which she can fixate at.

[23] Estimation results without such restrictions are similar to the results presented below and are provided in Supplementary Table 4: 12 of the 17 subjects are classified as the same level-$k$ lookup type.

[24] In the experiment, subjects could look at the entire computer screen. Here, we only consider lookups that fall on the grid map and drop the rest.

Define a state-to-lookup mapping $l_n^k : \Omega_k \to G_n$ which assigns each state $s$ a corresponding lookup location on the map $G_n$ according to the level-$k$ model.[25] Suppose a level-2 player is inferred to be in state $s = -1$, then by the mapping $l_n^2$, her lookup should fall exactly on the location $l_n^2(-1)$. In words, when a level-2 player is in state $s = -1$, she is thinking about what her opponent as a level-1 would choose. Hence, the state-to-lookup mapping $l_n^2(-1)$ should be on the location a level-1 opponent would choose. If her lookup is not on that location, we interpret this as an error. We assume a logit error structure so that looking at locations farther away from $l_n^2(-1)$ is less likely.

Formally, the lookup sequence in trial $n$ is a time series over $t = 1, ..., T_n$ where $T_n$ is the number of her lookups in this game $n$. Because of the logit error, a level-$k$ subject may not look at a location with certainty. Therefore, at the $t$-th lookup, let the random variable $\mathbf{R}_n^t$ be the probabilistic lookup location in $G_n$ and its realization be $r_n^t$. Denote the lookup history up to time $t$ by $\mathcal{R}_n^t \equiv \{r_n^1, \ldots, r_n^{t-1}, r_n^t\}$.

Conditional on $S_t = s_t$, the probability distribution of a level-$k$ subject's probabilistic lookup $\mathbf{R}_n^t$ is assumed to follow a logit error quantal response model (centered at $l_n^k(s_t)$), independent of lookup history $\mathcal{R}_n^{t-1}$. In other words,

$$\Pr(\mathbf{R}_n^t = r_n^t | S_t = s_t, \mathcal{R}_n^{t-1}) = \frac{\exp\left(-\lambda_k \left\| r_n^t - l_n^k(s_t) \right\|\right)}{\sum\limits_{g \in G_n} \exp\left(-\lambda_k \left\| g - l_n^k(s_t) \right\|\right)}. \tag{2}$$

where $\lambda_k \in [0, \infty)$ is the precision parameter. If $\lambda_k = 0$, the subject randomly looks at locations in $G_n$. As $\lambda_k \to \infty$, her lookups concentrate on the lookup location $l_n^k(s_t)$ predicted by the state $s_t$ of a level-$k$.

Combining the state transition matrix and the logit error, we can calculate the probability of observing lookup $r_n^t$ conditional on past lookup history $\mathcal{R}_n^{t-1}$:

$$\Pr(\mathbf{R}_n^t = r_n^t | \mathcal{R}_n^{t-1}) = \sum_{s_t \in \Omega_k} \Pr(S_t = s_t | \mathcal{R}_n^{t-1}) \cdot \Pr(\mathbf{R}_n^t = r_n^t | S_t = s_t, \mathcal{R}_n^{t-1}) \tag{1}$$

[25] For instance, if a level-2 player with target $(4, -2)$ in game $n = 16$ (player 1 as shown in Figure I) is at state $s = 0$ at a point of time, the mapping $l_{16}^2$ would give us the location $l_{16}^2(0) = (0, 0)$ which a level-0 player would choose (O in Figure I) since at this particular point of time, she is thinking about what her opponent thinks she would choose as a level-0. Similarly, if a level-2 player is in state $-1$, then the $l_{16}^2$ mapping would give us the location $l_{16}^2(-1) = (-2, 3)$ which a level-1 opponent would choose ($\mathbf{L1}_2$ in Figure I) since at this particular point of time, she is thinking about what her opponent would choose as a level-1. Finally, if a level-2 player 1 is in state 2, then the mapping $l_{16}^2$ would give us the location $l_{16}^2(2) = (2, 1)$ which a level-2 subject would choose ($\mathbf{L2}_1$ in Figure I) since at this particular point of time, she is thinking about her choice as a level-2.

where

$$\begin{aligned}
&\Pr(S_t = s_t|\mathcal{R}_n^{t-1}) \\
&\quad = \sum_{s_{t-1}\in\Omega_k} \Pr(S_{t-1} = s_{t-1}|\mathcal{R}_n^{t-1}) \cdot \Pr(S_t = s_t|S_{t-1} = s_{t-1}, \mathcal{R}_n^{t-1}) \\
&\quad = \sum_{s_{t-1}\in\Omega_k} \Pr(S_{t-1} = s_{t-1}|\mathcal{R}_n^{t-1}) \cdot \pi_{s_{t-1}\to s_t} \\
&\quad = \sum_{s_{t-1}\in\Omega_k} \frac{\Pr(S_{t-1} = s_{t-1}|\mathcal{R}_n^{t-2}) \Pr(\mathbf{R}_n^{t-1} = r_n^{t-1}|S_{t-1} = s_{t-1}, \mathcal{R}_n^{t-2})}{\Pr(\mathbf{R}_n^{t-1} = r_n^{t-1}|\mathcal{R}_n^{t-2})} \cdot \pi_{s_{t-1}\to s_t}. \qquad (2)
\end{aligned}$$

The second equality in equation (2) follows since according to the Markov property, $S_{t-1} = s_{t-1}$ is sufficient to predict $S_t = s_t$. Note that equation (2) depends on the Markov transition matrix. Meanwhile, the second term on the right hand side of equation (1) ($\Pr(\mathbf{R}_n^t = r_n^t|S_t = s_t, \mathcal{R}_n^{t-1})$) depends on the logit error. Notice that all the terms on the last line of equation (2) are now expressed with the time index moving backwards by one period. Hence, for a given game $n$, coupled with the initial distribution of states, the joint density of a level-$k$ subject's empirical lookups, denoted by

$$\begin{aligned}
f_n^k(r_n^1, ..., r_n^{T_n-1}, r_n^{T_n}) &\equiv \Pr(r_n^1, ..., r_n^{T_n-1}, r_n^{T_n}) \\
&= \Pr(r_n^1) \Pr(r_n^2|r_n^1) \Pr(r_n^3|r_n^1, r_n^2) ... \Pr(r_n^{T_n}|r_n^1, r_n^2, ..., r_n^{T_n-1}),
\end{aligned}$$

can be derived.[26] The log likelihood over all 24 trials is thus

$$L(\lambda_k, \theta_k) = \ln \left[ \prod_{n=1}^{24} f_n^k(r_n^1, ..., r_n^{T_n-1}, r_n^{T_n}) \right]. \qquad (3)$$

Since level-$k$ reasoning starts from the lowest state (here state 0), we assume this initial distribution of states degenerates to a mass point at the lowest state corresponding to level-0 (of herself if $k$ is even and of her opponent if $k$ is odd). With this assumption, we estimate the precision parameter $\lambda_k$ and the constrained Markov transition matrix $\theta_k$ using maximum likelihood estimation for each $k$, and classify subjects into the particular level-$k$ type which has the largest likelihood.

To summarize, for each level $k$, we estimate a state transition matrix and a precision parameter for the logit error. Thus for a given initial distribution of the states, we know the probability distribution of states at any point of time using the state transition matrix. Moreover, at any point of time, the mapping $l_n^k$ from the state to the lookup gives us the lookup location corresponding to any state when there is no error. Coupled with the error

[26] See Supplementary Appendix A3 for a formal derivation.

structure, we can calculate the probability distribution of various errors and therefore the distribution of predicted lookup locations. We then maximize the likelihood to explain the entire observed sequence of lookups. We do this for various levels. The final step is to select the $k$ in various level-$k$ types to best explain the observed sequence of lookups for each subject.

## D Vuong's Test for Non-Nested but Overlapping Models

The above econometric model may be plagued by an overfitting problem since higher level-$k$ types have more states and hence more parameters. It is not surprising if one discovers that models with more parameters fit better. In particular, the Markov-switching model for level-$k$ has $(k+1)$ states with a $(k+1)\times(k+1)$ transition matrix. This gives the model $\left[\frac{k(k+3)}{2}\right]$ parameters in the transition matrix alone.[27] For example, a level-2 subject has 3 states 0, $-1$, and 2 and five (Markov) parameters, but a level-1 subject has only 2 states 0 and 1 and two (Markov) parameters. Hence, we need to make sure our estimation does not select higher levels merely because it contains more states and more parameters. However, usual tests for model restrictions may not apply, since the parameters involved in different level-$k$ types could be non-nested. In particular, the state space of a level-2 subject $\{0,-1,2\}$ and the states of a level-1 subject $\{0,1\}$ are not nested. Yet, the state space of a level-1 type, $\{0,1\}$, is nested in the state space of a level-3 type, $\{0,1,-2,3\}$. In order to evaluate the classification, we use Vuong's test for non-nested but overlapping models (1989).[28]

Let $Lk^*$ be the type which has the largest likelihood with corresponding parameters $(\lambda_{k^*},\theta_{k^*})$. Let $Lk^a$ be an alternative type with corresponding parameters $(\lambda_{k^a},\theta_{k^a})$. In our case $Lk^*$ is the type with the largest likelihood based on lookups. The alternative type $Lk^a$ is the type having the next largest likelihood among all lower level types.[29] If according to Vuong's test, $Lk^*$ is a better model than $Lk^a$, we can be assured that the maximum likelihood criterion does not pick up the reported type by mere chance. Thus, we conclude that the lookup-based type is $Lk^*$. If instead we find that according to Vuong's test, $Lk^*$ and $Lk^a$ are equally good, then we conservatively classify the subject as the second largest lower type $Lk^a$.

Table II shows the results of the maximum likelihood estimation and Vuong's test

[27] Since each row sums up to one and elements with the column index greater than the row index plus one are zero, we have in total $(k+1)(k+1)-(k+1)-[k(k-1)]/2=[k(k+3)]/2$ parameters.

[28] See Supplementary Appendix A4 for the details of Vuong's test for non-nested but overlapping models. Note that this is the generalized version of the well-known "nested" Vuong's test.

[29] Recall that the reason why we look at Vuong's test is to avoid overfitting. Hence, if the alternative type has a larger transition matrix (more parameters) but a lower likelihood, there is no point to perform a test, since $Lk^*$ will not suffer from the problem of overfitting because it has fewer parameters but has a higher likelihood. This leads us to consider only lower level types as the alternative type.

for each subject. For each subject, we list her $Lk^*$ type, her $Lk^a$ type, her Vuong's test statistic, and her lookup-based type according to Vuong's test in order. Six of the seventeen subjects (subjects 1, 5, 6, 8, 11, 13) pass Vuong's test and have their lookup-based type as $Lk^*$. The remaining eleven subjects are conservatively classified as $Lk^a$. The overall results are summarized in column (A) of Table III. After employing Vuong's test, the type distribution for $(L0, L1, L2, L3, EQ)$ is $(1, 6, 4, 4, 2)$.[30] The distribution is slightly higher than typical type distributions reported in previous studies. In particular, there are two $EQ$'s and four $L3$'s, accounting for more than one third of the data. Treating the $EQ$ type as having a thinking step of 4, we find that the average number of thinking steps is 2.00, in line with results of the standard $p$-beauty contest games using Caltech subjects, but higher than normal subjects.[31] Neither employing Hansen [1992]'s test (to avoid nuisance parameter problems), nor iteratively applying Vuong's test (until the likelihood of the current type is significantly higher than that of the next alternative) alters the distribution of level-$k$ types by much (see A4 and Supplementary Table 3).

Up to now, we have shown that lookups do fall on the hotspots of the best response hierarchy (Hypothesis 2a). Classifying subjects based on lookups (Hypothesis 2b) gives us a reasonable level of sophistication as argued above. However, one might still wonder whether the results reported in Table II is due to a misspecification of possible types. After all, many assumptions are required for Hypothesis 2b to hold. We take up this issue now. Our argument is that if we take the level-$k$ theory literally to interpret underlying reasoning process, the classification based on lookups should match well with the classification using final choices alone since the level $k$ reflects a player's sophistication.

# V Matching Up with Final Choices

We first classify subjects using their final choices and compare classifications based on choices to those based on lookups. We point out the similarity between these two classification results. Finally we address how lookup data could help classify subjects when the choice data is noisy.

Following the literature, we classify individual subjects into various level-$k$ types based on final choices alone. Supplementary Appendix A5 provides details of the maximum

[30] Ignoring the two pseudo-17 subjects (subjects 3 and 17, both classified as $L1$) whose choices suggest non-compliance to level-$k$ theory, the type distribution for $(L0, L1, L2, L3, EQ)$ is $(1, 4, 4, 4, 2)$. For pseudotypes, refer to Costa-Gomes and Crawford [2006].

[31] Camerer [1997] reports that Caltech students play an average of 21.88 in a $p$-beauty contest game with $p = 0.7$. This is between $L2$'s choice of 24.5 and $L3$'s choice of 17.15. Higher than typical distributions could also result from the spatial beauty contest game being intuitive and not requiring mathematical multiplication (as compared with say, the standard $p$-beauty contest game), as Chou et al. [2009] show that a graphical presentation of the standard $p$-beauty contest game yields results closer to equilibrium.

likelihood estimation and pseudotype test we adopt from Costa-Gomes and Crawford [2006], and subject-by-subject results are reported in the sixth column of Table II. The idea of the pseudotype is to treat each subject's choices as a possible type. This is to examine whether there are clusters of subjects whose choices resemble each other's and thus predict other's choices in the cluster better than the pre-specified level-k types. Since we have 17 subjects, we include 17 pseudotypes, each constructed from one of our subject's choices in 24 trials. The aggregate distribution of types (with or without the pseudotype test) are reported in column (B) and (C) of Table III. In Table III, the choice-based and lookup-based classification results look similar. The choice results indicate slightly more steps of reasoning ($2.12 - 2.13$ for choice-based types instead of 2.00 for lookup-based types). This suggests that the lookup-based estimation (and the underlying Hypothesis 2b) is in the right ballpark. In fact, if we consider the classification results on a subject-by-subject basis, the similarity between the two estimations are even more evident. As reported in Table II, overall, for ten out of the seventeen subjects, their lookup-based types and the choice-based types are the same. In other words, for most subjects, when their choices reflect a particular level of sophistication, their lookup data suggests the same level of sophistication. Such alignment in classification results would be surprising if one thought Hypothesis 2b was too strong a claim. This supports a literal interpretation of the level-$k$ model. When a subject's choice data indicates a particular level of sophistication, her lookups suggest that the best response hierarchy of that level is carried out when she reasons.

Since the classification based on lookups and that based on choices align, we next turn to discuss the subtle differences between them. We evaluate the robustness of individual choice-based classification by performing bootstrap. This is a departure from past literature such as Costa-Gomes and Crawford [2006], as they do not consider whether the maximum likelihood estimation has enough power to distinguish between various types. For example, reading from Supplementary Table 1, for subject 14, the log likelihood is $-98.89$ for $L0$, $-84.17$ for $L1$, $-96.99$ for $L2$, $-76.67$ for $L3$, and $-74.45$ for $EQ$. Maximum likelihood estimation classifies her as $EQ$, although the likelihood of $L3$ is also close. In this case, classifying this subject as $EQ$ based on maximum likelihood alone may be questionable. To the best of our knowledge, there has not been any proposed test in experimental economics for evaluating the robustness of maximum likelihood-based type classifications. Hence we propose a bootstrap procedure (Efron [1979]; Efron and Tibshirani [1994]) to deal with the issue of robustness.[32] Imagine that from the maximum likelihood estimation, a subject is classified as a particular level-$k$ type with the logit

[32] Costa-Gomes and Crawford [2006] do use various information criteria to perform the horse-race. However, this still fails to address how much the runner-up is "close" to the winner.

error parameter $\lambda_k$. Draw (with replacement) 24 new trials out of the original dataset and re-estimate her $k$ and $\lambda_k$. We do this 1000 times to generate the discrete distribution of $k$ and the distribution of $\lambda_k$. Then, we evaluate the robustness of $k$ by looking at the distribution of $k$. Each level-$k$ type estimated from a re-sampled dataset that is not the same as her original level-$k$ type is viewed as a "misclassification," and counted against the original classification $k$. By calculating the total misclassification rate (out of 1000 re-samples), we can measure the robustness of the original classification. This bootstrap procedure is in the spirit of the test reported in Salmon [2001], which evaluates the robustness of the parameters estimated in a EWA learning model using simulated data.

The results of this bootstrap procedure are listed in Table IV. For each subject, we report the bootstrap distribution of $k$ (the number of times a subject is classified into $L0$, $L1$, $L2$, $L3$ or $EQ$ in the 1000 resampled datasets). The bootstrap misclassification rate (percentage of times classifying the subject as a type different from her original type) is listed in the last column. For example, subject 14 is originally classified as $EQ$, but is only re-classified as $EQ$ 587 times during the bootstrap procedure. Subject 14 is instead classified as $L3$ 228 times and as $L1$ 185 times. Hence, the distribution on the number of times that subject 14 is classified into $L0$, $L1$, $L2$, $L3$ or $EQ$ in the 1000 resampled datasets is $(0, 185, 0, 228, 587)$ and the corresponding misclassification rate is 0.413.

The bootstrap results align surprisingly well with whether the lookup-based classifications match their choice-based types. In particular, for the ten subjects whose two classifications match, all but three of them have (choice-based) bootstrap misclassification rates lower than 0.05, suggesting that their classifications are truly sharp.[33] In contrast, for six of the remaining seven subjects whose two classifications do not match, their choice-based type have bootstrap misclassification rates higher than 18.4%, suggesting that misclassifying these subjects into the wrong types using choice data alone (due to insignificantly larger likelihoods) is possible. The difference is significant, having a $p$-value of 0.0123 according to Mann-Whitney-Wilcoxon rank sum test. To sum up, when the lookup-based types match the choice-based types, it is when the choice-based classification is quite sharp. In contrast, when they differ, the classification based on choice is not that sharp, suggesting that for these subjects, choice data may not be enough.

In this case, one wonders whether lookup data could provide additional separation of types to predict choices. A closer look at Table IV (see the type underlined) indicates for ten subjects, when we resample their choices, the level they are most frequently classified into in the 1000 resampled choice datasets is exactly their level classified using their

[33] One of these three subjects (subject 17) fails the pseudotype test and is unlikely to resemble any of the level-$k$ types. The remaining two subjects (subjects 2 and 4) have a misclassification rate of 0.076 and 0.110. These are marginally higher than 0.05.

lookups.[34] For six other subjects, their lookup-based type is the one they are second most frequently classified into.[35] In fact, these subjects' lookup-based type also rank second in terms of likelihood based on choices.[36] A subject's lookup-based type is classified using her lookups, not using her choices. The high predictability of choices by her lookup-based type suggests that the lookup-based type is a viable alternative for predicting choices even when the lookup-based types differ from the choice-based types.

In order to evaluate whether lookup data can indeed improve classification, we perform an out-of-sample prediction horse-race between the lookup-based and choice-based types. Note that our lookup-based model makes predictions on lookups, not on final choice per se. However, we can first classify individual subjects into a particular level-$k$ type based on either lookups or choices using two thirds of the trials, and see how well the classified level-$k$ type predicts the final choices of the remaining one third of trials. In particular, for each subject, we classify her as a level-$k^l_{16}$ type based on lookups (using the first 16 sequences of lookups) and a level-$k^c_{16}$ type based on final choices (using the first 16 final choices) respectively. We then use these particular $k$'s (one for lookup, the other for choice) to predict final choices of the last eight trials. Since we are mainly interested in how lookup data can provide additional separation of types (to predict behavior) when choice data is insufficient, we group subjects into those whose choice-based classification is robust (having bootstrap misclassification rates greater than 0.05 as reported in the right panel of Table II), and those who is not.

To compare the prediction power of the two models, we report mean square errors of the predicted choices for the lookup-based and choice-based models. In particular, suppose a subject chose location $g_n = (x_n, y_n)$ in trial $n$, while the lookup-based and choice-based models predicted $(x^l_n, y^l_n)$ and $(x^c_n, y^c_n)$. Then, the mean square errors of the two models are $\left|x_n - x^l_n\right|^2 + \left|y_n - y^l_n\right|^2$ and $|x_n - x^c_n|^2 + |y_n - y^c_n|^2$ respectively. As reported in Table V, though overall performance of the two models are comparable, among the nine subjects whose choice-based types are not robust, the lookup-based model has a better mean square error of 5.75 (compared with 8.67 for the choice-based model) predicting the last eight trials.[37] A Wilcoxon sign rank test shows that this difference is marginally significant ($p = 0.0781$).[38]

To see how significant this gain in prediction power is, we calculate the "economic

[34] They are subjects 1, 2, 4, 5, 7, 10, 12, 13, 16, 17 (those whose two classifications match).

[35] They are subjects 3, 6, 8, 9, 11, 15.

[36] Refer to the likelihood double underlined in Supplementary Table 1.

[37] Even among the "robust" subjects, subject 7 is the only one whose lookup-based model has a much larger mean square error than the choice-based model.

[38] If we focus only on the seven subjects whose two classifications differ, the lookup-based model still has a better mean square error of 6.55 (compared with 8.68 for the choice-based model), though not statistically significant.

value" (cf. Camerer, Ho and Chong, 2004) of the two models, to evaluate how much these predictions could potentially add to the opponent's payoffs. In particular, we calculate the opponent's payoffs had they followed these models and best responded to the model predictions, $\pi^{Follow}$, and see how much an opponent can gain in addition to his actual payoffs, $\pi^{Actual}$, in the experiment. The economic value is the percentage of this gain, compared with the maximum gain possible, $\pi^{BR}$: (Note that economic values could be negative if the model performs worse than actual subjects.)

$$EV = \frac{\pi^{Follow} - \pi^{Actual}}{\pi^{BR} - \pi^{Actual}}$$

Results in the last two columns of Table V show that both choice-based and lookup-based models have good predictive power (compared to actual subjects) and can (on average) increase opponent payoffs by $39-41\%$. Moreover, the bootstrap robustness test indeed evaluates choice-based models well—the second panel of Table V show that for the robustness subjects, the average economic value for the choice-based model is 56.3%, higher than the lookup-based model (42.0%). On the other hand, the lookup-based model is a good compliment, especially when choice data is not good enough: As shown in the the first panel of Table V, for the non-robust subjects, the average economic value for the lookup-based model is 40.4%, compared with 24.3% for the choice-based model. In other words, among the subjects whose choice-based type is not robust to bootstrap, had the opponent known her lookup-based level, his payoffs could be increased by 40.4%. As a comparison, had the opponent known her choice-based level, his payoffs could be increased by 24.3%.

To summarize, these results show that lookup data can help us confirm classification results based on choices alone and even provide better classification results when choice-based classifications are not robust. Moreover, lookup data provide a chance to put the level-$k$ model to an ultimate test, asking if the model can not only predict final choices, but also describe the decision-making process employed by subjects by going through the best response hierarchy specified in Hypothesis 2b. Results in Table II show that the level-$k$ model does indeed hold up under this test for our spatial beauty contest games. One ought to keep in mind that explaining the reasoning process is a hard one, if not harder than explaining choices. Seeing in our dataset, for more than a half of subjects, their lookup-based types are aligned with their choice-based types should be read as a strong support to the level-$k$ model. This may be due to the graphical nature of the spatial beauty contest games. How general this result is should be tested in future experiments in which the reasoning process can somehow be analyzed.

# VI Conclusion

We introduce a new spatial beauty contest game in which the process of reasoning can be tracked, and provide theoretical predictions based on the equilibrium and a literal interpretation of the level-$k$ theory. The theoretical predictions of the level-$k$ model yield a plausible hypothesis on the decision-making process when the game is actually played. We then conduct laboratory experiments using video-based eyetracking technology to test this conjecture, and fit the eyetracking data on lookups using a constrained Markov-switching model of level-$k$ reasoning. Results show that based on lookups, experimental subjects' lookup sequences could be classified into following various level-$k$ best response hierarchies, which for more than a half of them coincide with types that they were classified into using final choices alone. Moreover, when the two classifications differ, most of the choice-based types are not robust to bootstrap, indicating that we might have misclassified them due to insignificantly larger likelihoods. In fact, lookup-based types often come out second (if not first) in the bootstrap procedure. Finally, for all subjects whose choice-based models are not robust to bootstrap, an out of sample prediction exercise shows that lookup-based models predict final choices better. This suggests that studying the reasoning process (such as through eyetracking lookups) can indeed help us understand economic behavior (such as individual's final choices) better.

Analyzing reasoning processes is a hard task. The spatial beauty contest game is designed to fully exploit the structure of the $p$-beauty contest so that subjects are induced to literally count on the map to carry out their reasoning as implied by the best response hierarchy of a level-$k$ theory. The high percentage of subjects whose classifications based on lookups and choices align could be read as a support to the level-$k$ model as a complete theory of reasoning and choice altogether in the spatial beauty contest game. Whether this holds true for more general games remains to be seen. Nevertheless, the paper points out a possibility of analyzing reasoning before arriving at choices. A design exploiting the structure of the game and is ideal for the tracking technology used seems to be indispensable.

Pennsylvania State University
National Taiwan University
National Taiwan University

# VII References

Brainard, D. H. [1997], 'The psychophysics toolbox', *Spatial Vision* **10**, 433–436.

Burchardi, K. B. and Penczynski, S. P. [2011], Out of your mind: Eliciting individual reasoning in one shot games.

Camerer, C. F. [1997], 'Progress in behavioral game theory', *Journal of Economic Perspectives* **11**(4), 167–188.

Camerer, C. F., Ho, T.-H. and Chong, J.-K. [2004], 'A cognitive hierarchy model of games', *Quarterly Journal of Economics* **119**(3), 861–898.

Camerer, C. F., Johnson, E., Rymon, T. and Sen, S. [1993], *Cognition and Framing in Sequential Bargaining for Gains and Losses*, MIT Press, Cambridge, pp. 27–47.

Chou, E., McConnell, M., Nagel, R. and Plott, C. R. [2009], 'The control of game form recognition in experiments: Understanding dominant strategy failures in a simple two person "guessing" game', *Experimental Econoimcs* **12**(2), 159–179.

Cornelissen, F. W., Peters, E. M. and Palmer, J. [2002], 'The eyelink toolbox: Eye tracking with matlab and the psychophysics toolbox', *Behavior Research Methods, Instruments and Computers* **34**, 613–617.

Costa-Gomes, M. A. and Crawford, V. P. [2006], 'Cognition and behavior in two-person guessing games: An experimental study', *American Economic Review* **96**(5), 1737–1768.

Costa-Gomes, M., Crawford, V. P. and Broseta, B. [2001], 'Cognition and behavior in normal-form games: An experimental study', *Econometrica* **69**(5), 1193–1235.

Crawford, V. P. and Iriberri, N. [2007*a*], 'Fatal attraction: Salience, naivete, and sophistication in experimental hide-and-seek games', *American Economic Review* **97**(5), 1731–1750.

Crawford, V. P. and Iriberri, N. [2007*b*], 'Level-k auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions?', *Econometrica* **75**(6), 1721–1770.

Efron, B. [1979], 'Bootstrap methods: Another look at the jackknife', *The Annals of Statistics* **7**(1), 1–26.

Efron, B. and Tibshirani, R. J. [1994], *An Introduction to the Bootstrap*, Chapman and Hall/CRC Monographs on Statistics and Applied Probability, Chapman and Hall/CRC.

Gabaix, X., Laibson, D., Moloche, G. and Weinberg, S. [2006], 'Costly information acquisition: Experimental analysis of a boundedly rational model', *American Economic Review* **96**(4), 1043–1068.

Grosskopf, B. and Nagel, R. [2008], ‘The two-person beauty contest’, *Games and Economic Behavior* **62**(1), 93–99.

Hamilton, J. D. [1989], ‘A new approach to the economic analysis of nonstationary time series and the business cycle’, *Econometrica* **57**(2), 357–384.

Hansen, B. E. [1992], ‘The likelihood ratio test under nonstandard conditions: Testing the markov switching model of gnp’, *Journal of Applied Econometrics* **7**(S1), S61–S82.

Ho, T. H., Camerer, C. F. and Weigelt, K. [1998], ‘Iterated dominance and iterated best response in experimental ”p-beauty contests”’, *American Economic Review* **88**(4), 947–969.

Johnson, E. J., Camerer, C., Sen, S. and Rymon, T. [2002], ‘Detecting failures of backward induction: Monitoring information search in sequential bargaining’, *Journal of Economic Theory* **104**(1), 16–47.

Koszegi, B. and Szeidl, A. [2013], ‘A model of focusing in economic choice’, *Quarterly Journal of Economics* **128**(1), forthcoming.

Krajbich, I., Armel, C. and Rangel, A. [2010], ‘Visual fixations and the computation and comparison of value in simple choice’, *Nature Neuroscience* **13**(10), 1292–1298. 10.1038/nn.2635.

Kuo, W.-J., Sjostrom, T., Chen, Y.-P., Wang, Y.-H. and Huang, C.-Y. [2009], ‘Intuition and deliberation: Two systems for strategizing in the brain’, *Science* **324**(5926), 519–522.

Nagel, R. [1995], ‘Unraveling in guessing games: An experimental study’, *American Economic Review* **85**(5), 1313–1326. FLA 00028282 American Economic Association Copyright 1995 American Economic Association.

Pelli, D. G. [1997], ‘The videotoolbox software for visual psychophysics: Transforming numbers into movies’, *Spatial Vision* **10**, 437–442.

Reutskaja, E., Nagel, R., Camerer, C. F. and Rangel, A. [2011], ‘Search dynamics in consumer choice under time pressure: An eye-tracking study’, *American Economic Review* **101**(2), 900–926.

Salmon, T. C. [2001], ‘An evaluation of econometric models of adaptive learning’, *Econometrica* **69**(6), 1597–1628. FLA 00129682 Econometric Society Copyright 2001 The Econometric Society.

Samuelson, P. A. [1938], ‘A note on the pure theory of consumer’s behaviour’, *Economica* **5**(17), 61–71.

Selten, R. [1991], ‘Properties of a measure of predictive success’, *Mathematical Social Sciences* **21**(2), 153–167.

Stahl, Dale, O. and Wilson, P. W. [1995], ‘On players’ models of other players: Theory and experimental evidence’, *Games and Economic Behavior* **10**(1), 218–254.

Vuong, Q. [1989], ‘Likelihood ratio tests for model selection and non-nested hypotheses’, *Econometrica* **57**(2), 307–333.

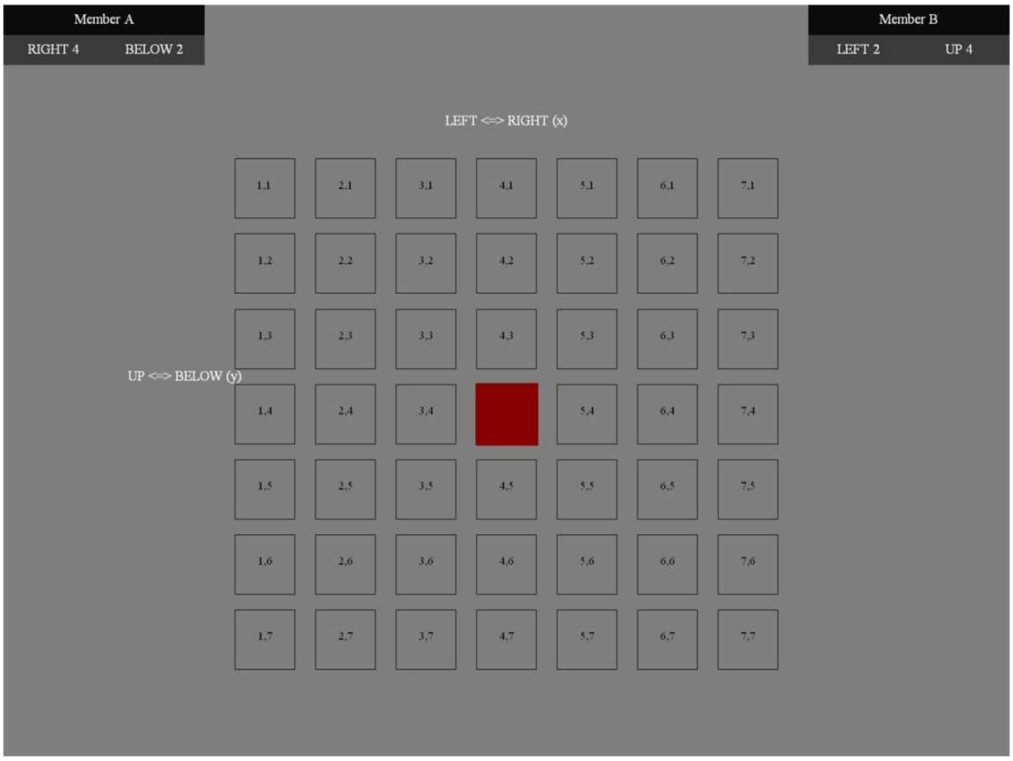
Wang, J. T.-y., Spezio, M. and Camerer, C. F. [2010], ‘Pinocchio’s pupil: Using eye-tracking and pupil dilation to understand truth telling and deception in sender-receiver games’, *American Economic Review* **100**(3), 1–26.

## Figures and Tables

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 3 | | $\mathbf{L1}_2$ | | $\mathbf{L3}_2$ | $\mathbf{E}_2$ | | |
| 2 | | | | | $\mathbf{L2}_2$ | | |
| 1 | | | | | | $\mathbf{L2}_1$ | $\mathbf{E}_1$ |
| 0 | | | | O | | | $\mathbf{L3}_1$ |
| -1 | | | | | | | |
| -2 | | | | | | | $\mathbf{L1}_1$ |
| -3 | | | | | | | |
| | -3 | -2 | -1 | 0 | 1 | 2 | 3 |

Figure I: Equilibrium and Level-*k* Predictions of a 7x7 Spatial Beauty Contest Game with Targets (4, -2) and (-2, 4) (Game 16). Predictions specifically for player 1 with Target (4,-2) are $\mathbf{L1}_1 \sim \mathbf{E}_1$, and predictions for player 2 with Target (-2,4) are $\mathbf{L1}_2 \sim \mathbf{E}_2$. O stands for the prediction of *L0* for both players. Note that $\mathbf{Lk}_1$ and $\mathbf{Lk}_2$ are the best responses to $\mathbf{L(k\text{-}1)}_2$ and $\mathbf{L(k\text{-}1)}_1$, respectively. For example, $\mathbf{L2}_2$'s choice (1,2) is the best response to $\mathbf{L1}_1$ since (3,-2) + (-2, 4) = (1, 2).

Figure II: Screen Shot of the GRAPH Presentation



| Member A | |
|---|---|
| RIGHT 4 | BELOW 2 |

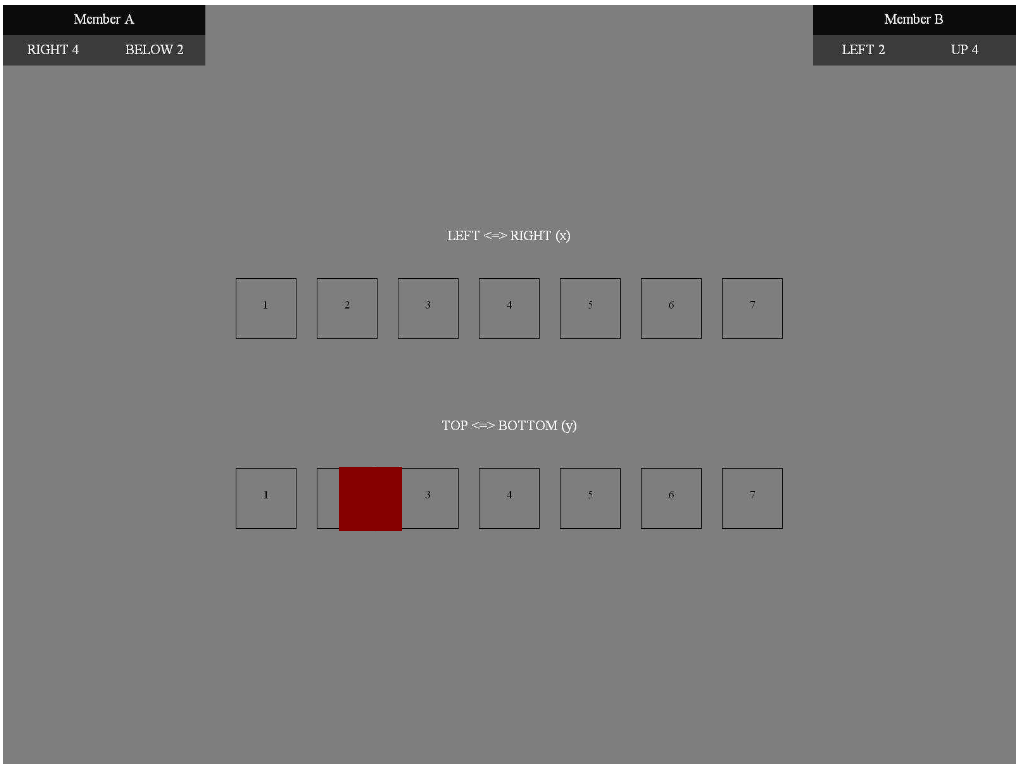
| Member B | |
|---|---|
| LEFT 2 | UP 4 |

LEFT <=> RIGHT (x)

UP <=> BELOW (y)

| | | | | | | |
|---|---|---|---|---|---|---|
| 1,1 | 2,1 | 3,1 | 4,1 | 5,1 | 6,1 | 7,1 |
| 1,2 | 2,2 | 3,2 | 4,2 | 5,2 | 6,2 | 7,2 |
| 1,3 | 2,3 | 3,3 | 4,3 | 5,3 | 6,3 | 7,3 |
| 1,4 | 2,4 | 3,4 | 🟥 | 5,4 | 6,4 | 7,4 |
| 1,5 | 2,5 | 3,5 | 4,5 | 5,5 | 6,5 | 7,5 |
| 1,6 | 2,6 | 3,6 | 4,6 | 5,6 | 6,6 | 7,6 |
| 1,7 | 2,7 | 3,7 | 4,7 | 5,7 | 6,7 | 7,7 |

Figure III: Screen Shot of the SEPARATE Presentation



| Member A | |
|---|---|
| RIGHT 4 | BELOW 2 |

| Member B | |
|---|---|
| LEFT 2 | UP 4 |

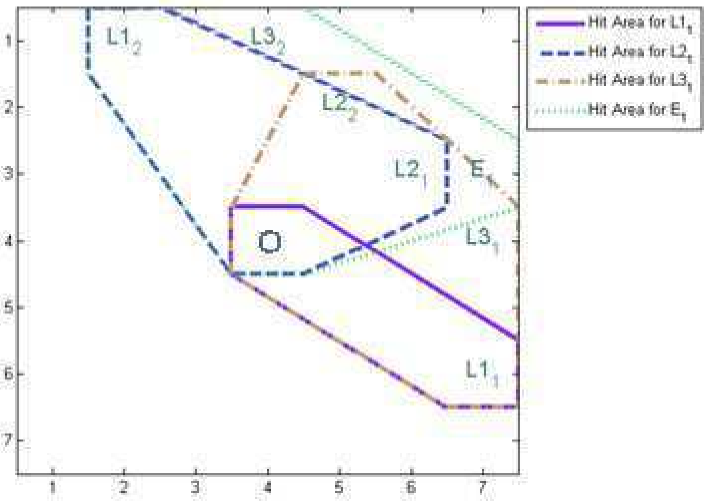
LEFT <=> RIGHT (x)

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |

TOP <=> BOTTOM (y)

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 🟥 | 3 | 4 | 5 | 6 | 7 |

Figure IV: *Hit* Areas for Various Level-*k* Types in Game 16 (7x7 with Target (4, -2) and the Opponent Target (-2, 4). *Hit* area is the minimal convex set enveloping the locations predicted by each level-*k* type's best response hierarchy.



Note: If we refer to Figure 1, for player 1, the *Hit* Area for level-*1* is the minimal convex set enveloping the locations (**O, L1**$_1$). The *Hit* Area for level-*2* is the minimal convex set enveloping the locations (**O, L1**$_2$**, L2**$_1$), and so on.

Figure V: Aggregate Empirical Percentage of Time Spent on the Union of *Hit* Areas ("Hit Time") in Each Game

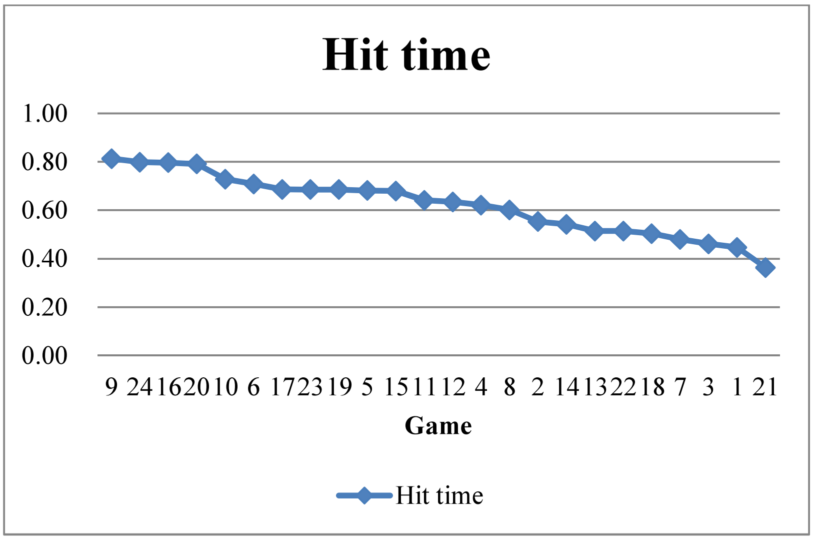


Figure VI: Aggregate Linear Difference Measure of Predicted Success in Each Game. It measures the difference between hit time and the hit area size.

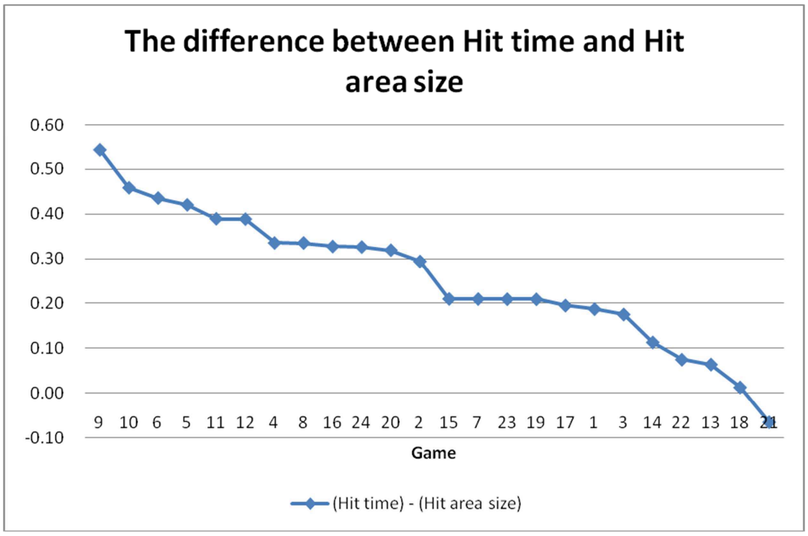


Figure VII: Subject 2's Eye Lookups in Trial 17 (as a Member B). The radius of the circle is proportional to the length of that lookup, so bigger circles indicate longer time spent.

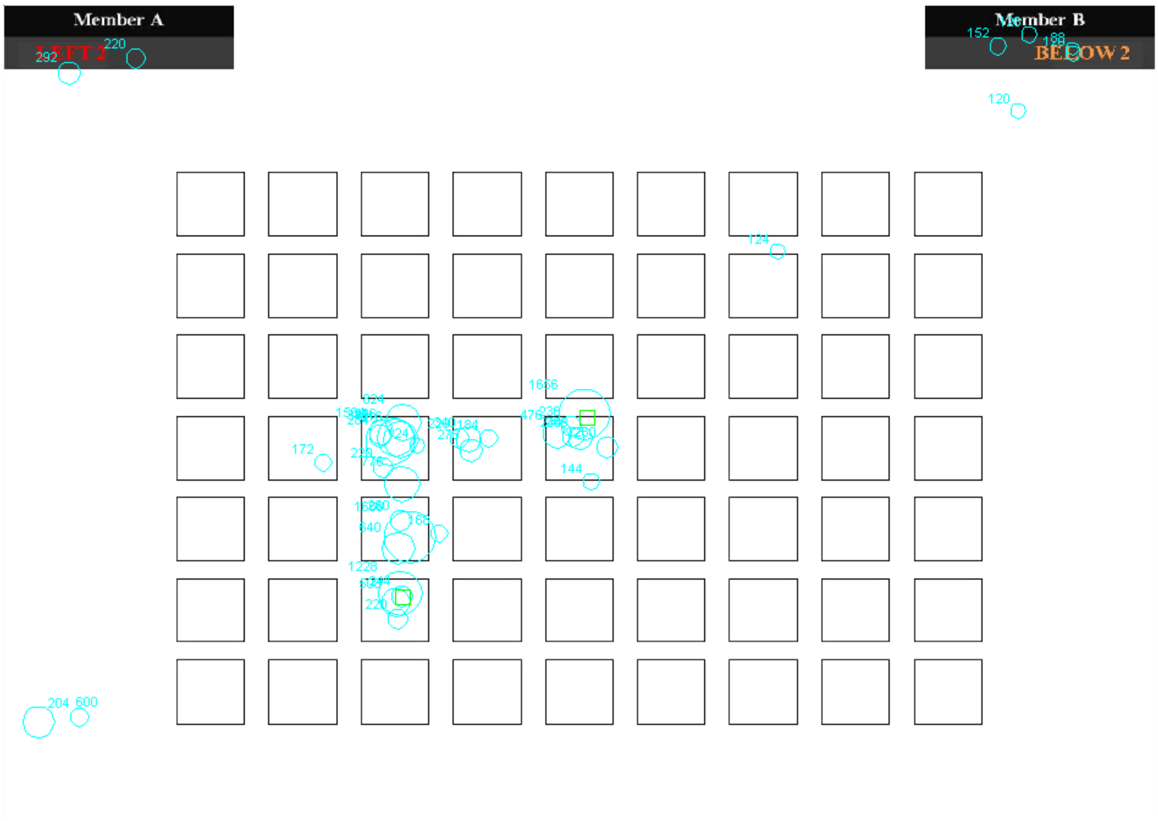


Table I: Level-$k$, Equilibrium Predictions and Minimum $\overline{k}$ 's in All Games

| *Game* | *Map size* | *Player 1 target* | *Player 2 target* | *L0* | *L1* | *L2* | *L3* | *EQ* | $\overline{k}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 9×9 | -2, 0 | 0,-4 | 0,0 | -2,0 | -2, -4 | $\underline{-4}, \underline{-4}$ | $\underline{-4}, \underline{-4}$ | 3 |
| 2 | 9×9 | 0, -4 | -2,0 | 0,0 | 0,-4 | $\underline{-2}, \underline{-4}$ | $\underline{-2}, \underline{-4}$ | -4,-4 | 4 |
| 3 | 7×7 | 2, 0 | 0,-2 | 0,0 | 2,0 | 2, -2 | 3, -2 | 3,-3 | 4 |
| 4 | 7×7 | 0, -2 | 2,0 | 0,0 | 0,-2 | 2, -2 | 2, -3 | 3,-3 | 4 |
| 5 | 11×5 | 2, 0 | 0,2 | 0,0 | 2,0 | 2, 2 | 4, 2 | 5,2 | 5 |
| 6 | 11×5 | 0, 2 | 2,0 | 0,0 | 0,2 | $\underline{2}, \underline{2}$ | $\underline{2}, \underline{2}$ | 5,2 | 6 |
| 7 | 9×7 | -2, 0 | 0,-2 | 0,0 | -2,0 | -2, -2 | -4, -2 | -4,-3 | 4 |
| 8 | 9×7 | 0, -2 | -2,0 | 0,0 | 0,-2 | -2, -2 | -2, -3 | -4,-3 | 4 |
| 9 | 7×9 | -4, 0 | 0,2 | 0,0 | -3,0 | $\underline{-3}, \underline{2}$ | $\underline{-3}, \underline{2}$ | -3,4 | 4 |
| 10 | 7×9 | 0, 2 | -4,0 | 0,0 | 0,2 | -3, 2 | $\underline{-3}, \underline{4}$ | $\underline{-3}, \underline{4}$ | 3 |
| 11 | 7×9 | 2, 0 | 0,2 | 0,0 | 2,0 | 2, 2 | 3, 2 | 3,4 | 5 |
| 12 | 7×9 | 0, 2 | 2,0 | 0,0 | 0,2 | 2, 2 | 2, 4 | 3,4 | 5 |
| 13 | 9×9 | -2, -6 | 4,4 | 0,0 | -2,-4 | 2, -2 | 0, -4 | 2,-4 | 4 |
| 14 | 9×9 | 4, 4 | -2,-6 | 0,0 | 4,4 | 2, 0 | 4, 2 | 4,0 | 4 |
| 15 | 7×7 | -2, 4 | 4,-2 | 0,0 | -2,3 | 1, 2 | 0, 3 | 1,3 | 4 |
| 16 | 7×7 | 4, -2 | -2,4 | 0,0 | 3,-2 | 2, 1 | 3, 0 | 3,1 | 4 |
| 17 | 11×5 | 6, 2 | -2,-4 | 0,0 | 5,2 | 4, 0 | $\underline{5}, \underline{0}$ | $\underline{5}, \underline{0}$ | 3 |
| 18 | 11×5 | -2, -4 | 6,2 | 0,0 | -2,-2 | $\underline{3}, \underline{-2}$ | 2, -2 | $\underline{3}, \underline{-2}$ | 4 |
| 19 | 9×7 | -6, -2 | 4,4 | 0,0 | -4,-2 | -2, 1 | -4, 0 | -4,1 | 4 |
| 20 | 9×7 | 4, 4 | -6,-2 | 0,0 | 4,3 | 0, 2 | 2, 3 | 0,3 | 4 |
| 21 | 7×9 | -2, -4 | 4,2 | 0,0 | -2,-4 | 1, -2 | 0, -4 | 1,-4 | 4 |
| 22 | 7×9 | 4, 2 | -2,-4 | 0,0 | 3,2 | 2, -2 | 3, 0 | 3,-2 | 4 |
| 23 | 7×9 | -2, 6 | 4,-4 | 0,0 | -2,4 | 1, 2 | 0, 4 | 1,4 | 4 |
| 24 | 7×9 | 4, -4 | -2,6 | 0,0 | 3,-4 | 2, 0 | 3, -2 | 3,0 | 4 |

Note: Each row corresponds to a game and contains the following information in order: (1) the game number, (2) the size of the grid map for that game, (3) the target of player 1, (4) the target of player 2, (5) the theoretic prediction of *L0* for player 1, (6) the theoretic prediction of *L1* for player 1, (7) the theoretic prediction of *L2* for player 1, (8) the theoretic prediction of *L3* for player 1, (9) the theoretic prediction of *EQ* for player 1, and (10) the minimum $\overline{k}$ for player 1 such that as long as the level is weakly higher, the choice of that type is the same as the choice of *EQ*. Non-separating types are $\underline{\text{underlined}}$.

Table II: Level-$k$ Types Based on Lookup Data (and Final Choice Data)

| (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|
| subject | $Lk^*$ | $Lk^a$ | Vuong's $V$ | $\boldsymbol{Lk^l}$ | $\boldsymbol{Lk^c}$ | bootstrap miss rate |
| 1 | L3 | L2 | 4.425+ | **L3** | **L3** | 0.000* |
| 2 | L3 | L2 | 0.689 | **L2** | **L2** | 0.076 |
| 3 | L3 | L1 | 1.577 | L1 | L3 | 0.244 |
| 4 | L3 | L1 | 1.597 | **L1** | **L1** | 0.110 |
| 5 | EQ | L2 | 2.977+ | **EQ** | **EQ** | 0.012* |
| 6 | EQ | L2 | 2.400+ | EQ | L2 | 0.236 |
| 7 | L2 | L0 | 1.582 | **L0** | **L0** | 0.034* |
| 8 | L3 | L1 | 2.812+ | L3 | EQ | 0.000* |
| 9 | EQ | L2 | 1.001 | L2 | L0 | 0.472 |
| 10 | L3 | L1 | 1.226 | **L1** | **L1** | 0.000* |
| 11 | L3 | L2 | 2.087+ | L3 | L2 | 0.365 |
| 12 | L3 | L1 | 0.853 | **L1** | **L1** | 0.010* |
| 13 | L3 | L1 | 3.939+ | **L3** | **L3** | 0.004* |
| 14 | L3 | L1 | 1.692 | L1 | EQ | 0.413 |
| 15 | L3 | L2 | 1.470 | L2 | L3 | 0.184 |
| 16 | L3 | L2 | 1.342 | **L2** | **L2** | 0.000* |
| 17 | L3 | L1 | 1.778 | **L1** | **L1** | 0.232 |

Note: + indicates Vuong's statistic $V$ is significant or $|V|>1.96$. ($Lk^*$ denotes the type with the largest likelihood; $Lk^a$ denotes the alternative lower level type which has the next-largest likelihood; $\boldsymbol{Lk^l}$ denotes the classified type based on Vuong's test; $Lk^c$ denotes level-$k$ type based on final choices alone.)

* indicates misclassification rate less than 0.05. 10 pairs of **boldfaced** level-$k$ types in columns (5)-(6) indicate agreement between the two.

Each row corresponds to a subject and contains the following information in order: (1) the subject number, (2) based on her lookups, the type with the largest likelihood, (3) based on her lookups, the alternative lower level type which has the next-largest likelihood, (4) Vuong's statistic in testing whether $Lk^*$and $Lk^a$ are equally good models, (5) subject's lookup type based on Vuong's test result, (Notice that in (5) we classify a subject as her $Lk^*$ type if according to Vuong's test, $Lk^*$ is a better model than $Lk^a$. If $Lk^*$and $Lk^a$ are equally good, since $Lk^a$ has fewer parameters, to avoid overfitting, we classify a subject as her $Lk^a$ type. The result in (5) is summarized in column (A) of Table 3.) (6) her choice-based level-$k$ type denoted by $\boldsymbol{Lk^c}$, (7) the bootstrap misclassification rate, i.e., the ratio that she is not classified as her original choice-based type ($\boldsymbol{Lk^c}$).

Table III: Distribution of Types under Various Specifications

| | (A) Lookup-based with Vuong's test | (B) Choice-based without Pseudotypes | (C) Choice-based with Pseudotypes |
|---|---|---|---|
| *L0* | 1 | 2 | 2 |
| *L1* | 6 | 4 | 3 |
| *L2* | 4 | 4 | 4 |
| *L3* | 4 | 4 | 3 |
| *Equilibrium* | 2 | 3 | 3 |
| *Pseudo-17* | - | - | 2 |
| *Aver. step* | 2.00 | 2.12 | 2.13 |

Note: In each row we list the number of subjects of that particular type based on various classifications. In the bottom row we list the average of thinking steps. We consider three ways to classify subjects. The first classification, reported in column (A), is based on the lookup data and we classify subjects to the type with the largest likelihood if according to Vuong's test, this type is a better model than the type with the next largest likelihood among all lower level types (and to the type with the next largest likelihood among all lower level types otherwise). The second classification, reported in column (B), uses the choice data in which pseudotypes are not included. The third classification, reported in column (C), also uses the choice data but in addition, pseudotypes are included.

Table IV: Distribution of Types in 1000 Bootstraps of Final Choice Data

| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|
| subject | $\boldsymbol{Lk^l}$ | $\boldsymbol{Lk^c}$ | L0 | L1 | L2 | L3 | EQ | bootstrap miss rate |
| 1 | **L3** | **L3** | 0 | 0 | 0 | 1000 | 0 | 0.000* |
| 2 | **L2** | **L2** | 1 | 0 | 924 | 75 | 0 | 0.076 |
| 3 | L1 | L3 | 0 | 233 | 1 | 756 | 10 | 0.244 |
| 4 | **L1** | **L1** | 63 | 890 | 11 | 36 | 0 | 0.110 |
| 5 | **EQ** | **EQ** | 0 | 0 | 1 | 11 | 988 | 0.012* |
| 6 | EQ | L2 | 0 | 3 | 764 | 5 | 228 | 0.236 |
| 7 | **L0** | **L0** | 966 | 0 | 12 | 17 | 5 | 0.034* |
| 8 | L3 | EQ | 0 | 0 | 0 | 0 | 1000 | 0.000* |
| 9 | L2 | L0 | 528 | 3 | 440 | 4 | 25 | 0.472 |
| 10 | **L1** | **L1** | 0 | 1000 | 0 | 0 | 0 | 0.000* |
| 11 | L3 | L2 | 0 | 0 | 635 | 363 | 2 | 0.365 |
| 12 | **L1** | **L1** | 0 | 990 | 6 | 4 | 0 | 0.010* |
| 13 | **L3** | **L3** | 0 | 1 | 3 | 996 | 0 | 0.004* |
| 14 | L1 | EQ | 0 | 185 | 0 | 228 | 587 | 0.413 |
| 15 | L2 | L3 | 0 | 9 | 165 | 816 | 10 | 0.184 |
| 16 | **L2** | **L2** | 0 | 0 | 1000 | 0 | 0 | 0.000* |
| 17 | **L1** | **L1** | 0 | 768 | 1 | 231 | 0 | 0.232 |

Note: * indicates misclassification rate less than 0.05. ($\boldsymbol{Lk^l}$ denotes the classified type based on lookup data; $\boldsymbol{Lk^c}$ denotes level-$k$ type based on final choices alone.) 10 pairs of **boldfaced** level-$k$ types in columns (2)-(3) indicate agreement between the two. Underlined numbers in columns (4)-(8) indicate each subject's lookup-based type. Notice that they are typically the second most frequent types subjects are classified into (if not the most frequent) if we resample their choices. The only exception is subject 14.

Each row corresponds to a subject and contains the following information in order: (1) the subject number, (2) subject's lookup type based on her lookups, (3) her choice-based level-$k$ type denoted by $\boldsymbol{Lk^c}$, (4)-(8) the number of times that she is classified as an *L0/L1/L2/L3/EQ* in 1000 times of bootstrapping her choice data, (9) the bootstrap misclassification rate, i.e., the ratio that she is not classified as her original choice-based type ($\boldsymbol{Lk^c}$).

Table V: Out-of-Sample Forecasting: Choice-based and Lookup-based Classifications

| Group | subject | Type | | Mean Square Error | | Economic Value | |
|---|---|---|---|---|---|---|---|
| | | $Lk_{16}^{l}$ | $Lk_{16}^{c}$ | $Lk_{16}^{l}$ | $Lk_{16}^{c}$ | $Lk_{16}^{l}$ | $Lk_{16}^{c}$ |
| Not Robust to Bootstrap | 2 | L2 | L3 | **0.750** | 4.875 | **0.790** | 0.422 |
| | 3 | L1 | L3 | **14.625** | 27.750 | **-0.243** | -0.450 |
| | 4 | L1 | L0 | **0.000** | 7.750 | **1.000** | -0.062 |
| | 6 | L2 | L2 | **1.500** | **1.500** | **0.813** | **0.813** |
| | 9 | L2 | L0 | **5.250** | 6.125 | 0.317 | **0.425** |
| | 11 | L3 | L3 | **3.000** | **3.000** | **0.415** | **0.415** |
| | 14 | L2 | EQ | 16.125 | **14.375** | -0.186 | **0.065** |
| | 15 | L2 | L3 | **1.875** | 4.000 | **0.669** | 0.501 |
| | 17 | L1 | L1 | **8.625** | **8.625** | **0.060** | **0.060** |
| Group Average | | | | 5.750 | 8.667 | 0.404 | 0.243 |
| (Std) | | | | (6.058) | (8.091) | (0.452) | (0.375) |
| Robust to Bootstrap | 1 | L3 | L3 | **1.875** | **1.875** | **0.735** | **0.735** |
| | 5 | EQ | EQ | **19.375** | **19.375** | **0.401** | **0.401** |
| | 7 | L3 | L0 | 27.375 | **7.875** | -0.490 | **0.249** |
| | 8 | L3 | EQ | **3.500** | 4.000 | 0.590 | **0.611** |
| | 10 | L1 | L1 | **0.625** | **0.625** | **0.793** | **0.793** |
| | 12 | L1 | L1 | **3.375** | **3.375** | **0.172** | **0.172** |
| | 13 | L3 | L3 | **2.750** | **2.750** | **0.624** | **0.624** |
| | 16 | L3 | L2 | 16.125 | **14.375** | 0.535 | **0.917** |
| Group Average | | | | 7.844 | 5.078 | 0.420 | 0.563 |
| (Std) | | | | (9.876) | (6.217) | (0.416) | (0.265) |
| Whole-sample Average | | | | 6.735 | 6.978 | 0.411 | 0.393 |
| (Std) | | | | (7.886) | (7.283) | (0.422) | (0.358) |

Note: $Lk_{16}^{l}$ denotes a subject's lookup-based type if we use the first 16 trials. $Lk_{16}^{c}$ denotes a subject's choice-based type if we use the first 16 trials.

We use the first 16 trials to estimate each subject's level-*k* type, and predict their final choices in the remaining 8 trials. The top panel is for the nine subjects whose choice data is not robust (i.e. those with misclassification rate greater than 0.05 as reported in the last column of Table 2). The bottom panel is for the eight subjects whose choice data is robust. We list in order (1) the subject number, (2) her lookup-based type if we use the first 16 trials, (3) her choice-based type if we use the first 16 trials, (4) the mean square error of the predicted choices for the remaining 8 trials based on the lookup-based type, (5) the mean square error of the predicted choices for the remaining 8 trials based on the choice-based type, (6) the economic value for the lookup-based type, (7) the economic value for the choice-based type. In the bottom row for each panel we list the mean and standard deviation of the mean square error and the economic value.

# Supplementary Appendix [For Online Reference Only]

## A1 Alternative L0 Specification

*For player i's choice* $(x_i, y_i)$ *and player i's target* $(a_i, b_i)$, *let player i's payoff be* $p_i(x_i, y_i; x_{-i}, y_{-i}; a_i, b_i) = \overline{s} - (|x_i - (x_{-i} + a_i)| + |y_i - (y_{-i} + b_i)|)$ *where* $\overline{s}$ *is a constant. Suppose player i is level-1 with a continuous von Neumann-Morgenstern utility function* $u(\cdot)$ *that values only monetary payoffs. Then, choosing location* $(a_i, b_i)$ *is the best response to a level-0 opponent -i who chooses randomly over the entire map,* $\{(x, y), x \in \{-X, -X+1, ..., X\}, y \in \{-Y, -Y+1, ..., Y\}\}$ .

***Proof.*** To best respond to the choice of player *-i*, player *i* should find $(x_i, y_i)$ that solves the maximization

$$(x_i, y_i) = \arg\max_{x,y} \sum_{y_{-i}=-Y}^{Y} \sum_{x_{-i}=-X}^{X} \frac{1}{(2X+1)(2Y+1)} u\left(\overline{s} - \left[|x - (x_{-i} + a_i)| + |y - (y_{-i} + b_i)|\right]\right).$$

To show that $(x_i, y_i) = (a_i, b_i)$ is the arg max, it suffices to show that $(x', y') = (0, 0)$ solves the maximization

$$\arg\max_{x',y'} \sum_{y_{-i}=-Y}^{Y} \sum_{x_{-i}=-X}^{X} u\left(\overline{s} - \left[|x' - x_{-i}| + |y' - y_{-i}|\right]\right). \tag{1}$$

For any given $y_{-i}$ , $y'$, let $Y_{-i} = y' - y_{-i}$ . Then the summation over *x*, given $y_{-i}$ , $y'$
, can be expressed as

$$\sum_{x_{-i}=-X}^{X} u\left(\overline{s} - Y_{-i} - |x' - x_{-i}|\right), \tag{2}$$

which is symmetric by $x' = 0$ .

Without loss of generality, consider two choices of $x'$, $X \geq x' = t > 0$ and $x' = 0$ . Player *i*'s payoff when choosing $x' = t$ differs from that when choosing $x' = 0$ by

$$\begin{aligned} &\sum_{x_{-i}=-X}^{X} u\left(\overline{s} - Y_{-i} - |t - x_{-i}|\right) - \sum_{x_{-i}=-X}^{X} u\left(\overline{s} - Y_{-i} - |0 - x_{-i}|\right) \\ &= \sum_{k=-t}^{2X-t} u\left(\overline{s} - Y_{-i} - |X - k|\right) - \sum_{k=0}^{2X} u\left(\overline{s} - Y_{-i} - |X - k|\right) \\ &= \sum_{k=-t}^{-1} u\left(\overline{s} - Y_{-i} - |X - k|\right) - \sum_{k=2X-t+1}^{2X} u\left(\overline{s} - Y_{-i} - |X - k|\right) \\ &= \sum_{k=X+1-t}^{X} \left[u\left(\overline{s} - Y_{-i} - |t + k|\right) - u\left(\overline{s} - Y_{-i} - |k|\right)\right] < 0, \end{aligned} \tag{3}$$

where the equalities follow because of simple algebra and the last inequality holds since $|t+k| > |k|$ for all $X + 1 - t \leq k \leq X$ (notice that $X \geq t$ implies that

$X+1-t\geq 1$), and $u(\cdot)$ is increasing. Hence, choosing $x'=t$ is worse than choosing $x'=0$. Since (2) is symmetric by $x'=0$, the same argument applies to show that choosing $x'=-t$ is worse than choosing $x'=0$. Thus $x'=0$ maximizes the summation of (2) for any given $y_{-i}$ ,$y'$.

Similarly, $y'=0$ maximizes $\sum_{y_{-i}=-Y}^{Y}\sum_{x_{-i}=-X}^{X} u(\overline{s}-(|0-x_{-i}|+|y'-y_{-i}|))$. Therefore, $(x_i,y_i)=(a_i,b_i)$ is optimal if the (level-*0*) opponent chooses uniformly on the map.

## A2 Proof of Proposition 2

*(1)* $(x_i^k,y_i^k)=R(X,Y;(a_i+x_{-i}^{k-1},b_i+y_{-i}^{k-1}))$ *for* $k\in\{1,2,\ldots\}$ *and* $(x_1^0,y_1^0)=(x_2^0,y_2^0)\equiv(0,0)$.

***Proof.*** Following the notations defined above in A1, to find $(x_i^k,y_i^k)$ that solves $\max_{x,y} u\left(\overline{s}-(|x-(x_{-i}^{k-1}+a_i)|+|y-(y_{-i}^{k-1}+b_i)|)\right)$, we may solve $x_i^k$ and $y_i^k$ separately since there is no interaction between the choice of $x_i^k$ and $y_i^k$. Hence, by symmetry we only need to show that $x_i^k=\min\left\{X,\max\left\{-X,a_i+x_{-i}^{k-1}\right\}\right\}$. Notice that

$$\min\left\{X,\max\left\{-X,x_{-i}^{k-1}+a_i\right\}\right\}=\begin{cases}-X, & x_{-i}^{k-1}+a_i<-X\\ x_{-i}^{k-1}+a_i, & x_{-i}^{k-1}+a_i\in\{-X,-X+1,\ldots X\}.\\ X, & x_{-i}^{k-1}+a_i>X\end{cases}$$

In other words, when the unadjusted best response $x_{-i}^{k-1}+a_i$ is lower than the lowest possible choice of $x_i^k$ on the grid map, the adjusted best response is the lower bound $-X$ . When it is higher than the highest possible choice of $x_i^k$ on the grid map, the adjusted best response is the upper bound $X$ . When the unadjusted best response $x_{-i}^{k-1}+a_i$ is within the possible range of $x_i^k$ on the grid map, the adjusted best response coincides with the unadjusted best response. Notice that:

1. If $x_{-i}^{k-1}+a_i\in\{-X,-X+1,\ldots,X\}$ , $\min_{x\in\{-X,-X+1,\ldots,X\}}\left|x-(x_{-i}^{k-1}+a_i)\right|=0$ at $x=x_{-i}^{k-1}+a_i$ ;
2. If $x_{-i}^{k-1}+a_i>X$ , $\min_{x\in\{-X,-X+1,\ldots,X\}}\left|x-\left(x_{-i}^{k-1}+a_i\right)\right|=-X+\left(x_{-i}^{k-1}+a_i\right)$ at $x=X$ ;
3. If $x_{-i}^{k-1}+a_i<-X$, $\min_{x\in\{-X,-X+1,\ldots,X\}}\left|x-\left(x_{-i}^{k-1}+a_i\right)\right|=-X-\left(x_{-i}^{k-1}+a_i\right)$ at $x=-X$ .

Thus, $x_i^k=\min\left\{X,\max\left\{-X,x_{-i}^{k-1}+a_i\right\}\right\}$ indeed maximizes player *i*'s utility (which is decreasing in the distance between the target $x_{-i}^{k-1}+a_i$ and the choice).

*(2) there exists a smallest integer* $\overline{k}$ *such that for all* $k \geq \overline{k}$ *,* $(x_i^k, y_i^k) = (x_i^e, y_i^e)$ *and* $(x_{-i}^k, y_{-i}^k) = (x_{-i}^e, y_{-i}^e)$.

***Proof.*** It suffices to show that there exists a smallest positive integer $\overline{k}$ such that $(x_i^k, y_i^k) = (x_i^e, y_i^e)$ for all $k \geq \overline{k}$ when $a_1 + a_2 < 0$. All other possibilities can be argued analogously.

There are 2 cases to consider: $a_i < 0 \leq a_{-i}$ and $a_1, a_2 < 0$.

**Case 1**: $a_i < 0 \leq a_{-i}$ .

We show that when $x_i^k > -X$ , $x_i^{k+2}$ is strictly less than $x_i^k$, and when $x_i^k = -X$, $x_i^{k+2} = -X$ . Then all subsequences taking the form of $\{x_i^k, x_i^{k+2}, x_i^{k+4}, ...\}$ will eventually converge to $x_i^e = -X$ , implying the sequence $\{x_i^0, x_i^1, x_i^2, ...\}$ also converges to $x_i^e = -X$ .

For any nonnegative integer $k$ ,

$$x_i^{k+2} - x_i^k = \min\left\{X, \max\left\{-X, x_{-i}^{k+1} + a_i\right\}\right\} - x_i^k$$

where

$$\begin{aligned} x_{-i}^{k+1} &= \min\{X, \max\{-X, \underbrace{x_i^k}_{\geq -X} + \underbrace{a_{-i}}_{\geq 0}\}\} \\ &= \min\left\{X, x_i^k + a_{-i}\right\}. \end{aligned}$$

If $x_i^k > -X$,

$$\begin{aligned} &x_i^{k+2} - x_i^k \\ &= \min\left\{X, \max\left\{-X, x_{-i}^{k+1} + a_i\right\}\right\} - x_i^k \\ &= \min\left\{X, \max\left\{-X, \min\left\{X, x_i^k + a_{-i}\right\} + a_i\right\}\right\} - x_i^k \\ &= \min\{X, \underbrace{\max\{-X, \underbrace{\min\{\underbrace{X + a_i}_{<X}, x_i^k + a_{-i} + a_i\}}_{<X}\}}_{<X}\} - x_i^k \\ &= \max\left\{-X, \min\left\{X + a_i, x_i^k + a_{-i} + a_i\right\}\right\} - x_i^k \\ &= \max\{\underbrace{-X - x_i^k}_{<0}, \underbrace{\min\{X + a_i, \underbrace{x_i^k + a_{-i} + a_i}_{<x_i^k}\}}_{<x_i^k} - x_i^k\} < 0. \end{aligned}$$

If $x_i^k = -X$,

$$
\begin{aligned}
&x_i^{k+2}-x_i^k\\
&=\min\left\{X,\max\left\{-X,x_{-i}^{k+1}+a_i\right\}\right\}-x_i^k\\
&=\min\left\{X,\max\left\{-X,\min\left\{X,x_i^k+a_{-i}\right\}+a_i\right\}\right\}-x_i^k\\
&=\min\left\{X,\max\left\{-X,\min\left\{X,-X+a_{-i}\right\}+a_i\right\}\right\}-(-X)\\
&=\min\Big\{X,\max\Big\{-X,\underbrace{\min\{X+a_i,\underbrace{-X+\underbrace{a_{-i}+a_i}_{<0}}_{<-X}\}}_{<-X}\Big\}\Big\}-(-X)\\
&=\min\left\{X,-X\right\}-(-X)\\
&=-X-(-X)=0.
\end{aligned}
$$

For player *-i*, we know from Case 1 that there exists a positive integer $\overline{k}_i$ where the opponent chooses $x_i^k=x_i^e=-X$ for all $k\geq\overline{k}_i$ . This implies $x_{-i}^{k+1}=x_{-i}^e=-X+a_{-i}$ for all $k\geq\overline{k}_i$, since $x_{-i}^{k+1}=\min\left\{X,\max\left\{-X,x_i^k+a_{-i}\right\}\right\}$.

**Case 2**: $a_1,a_2<0$.

As in Case 1, again we show that when $x_i^k>-X$ , $x_i^{k+2}$ is strictly less than $x_i^k$ , and when $x_i^k=-X$ , $x_i^{k+2}=-X$ . Then all subsequences taking the form of $\{x_i^k,x_i^{k+2},x_i^{k+4},\ldots\}$ will eventually converge to $x_i^e=-X$ , implying the sequence $\{x_i^0,x_i^1,x_i^2,\ldots\}$ also converges to $x_i^e=-X$ . Since $x_{-i}^{k+1}=\min\Big\{X,\max\{-X,\underbrace{\underbrace{x_i^k}_{\leq m}+\underbrace{a_{-i}}_{<0}}_{<m}\}\Big\}=\max\left\{-X,x_i^k+a_{-i}\right\}$,

$$
\begin{aligned}
x_i^{k+2}-x_i^k&=\min\left\{X,\max\left\{-X,x_{-i}^{k+1}+a_i\right\}\right\}-x_i^k\\
&=\min\left\{X,\max\left\{-X,\max\left\{-X,x_i^k+a_{-i}\right\}+a_i\right\}\right\}-x_i^k\\
&=\min\Big\{X,\max\Big\{-X,\max\{\underbrace{-X+a_i}_{<-X},x_i^k+a_{-i}+a_i\}\Big\}\Big\}-x_i^k\\
&=\min\left\{X,\max\left\{-X,x_i^k+a_{-i}+a_i\right\}\right\}-x_i^k.
\end{aligned}
$$

If $x_i^k>-X$,

$$x_i^{k+2}-x_i^k=\min\Big\{X-x_i^k,\max\{\underbrace{-X-x_i^k}_{<0},\underbrace{a_{-i}+a_i}_{<0}\}\Big\}<0.$$

If $x_i^k=-X$,

$$\begin{aligned} x_i^{k+2} - x_i^k &= \min\{X, \max\{-X, \underbrace{-X + \underbrace{\underline{a}_{-i} + \underline{a}_i}_{<0}}_{<-X}\}\} - (-X) \\ &= \min\{X, -X\} - (-X) \\ &= 0. \end{aligned}$$

Then we can argue as in Case 1 that player –*i* will eventually choose $x_{-i}^k = x_{-i}^e = X$.

## A3 Initial Distribution of States

Formally, we start with the assumption that $\Pr(\mathbf{S}_0 = s_0) = 1$ when the initial state $s_0$ is ***0*** and zero otherwise. Then we derive the following step by step. First, for $\Pr(s_0)$ given by the initial distribution of states and $\Pr(s_1 \mid s_0)$ given by the Markov transition matrix, $\Pr(s_1) = \sum_{s_0 \in \Omega_k} \left[\Pr(s_0)\Pr(s_1 \mid s_0)\right]$ .

Second, for $\Pr(s_1)$ given by the first step and $\Pr(r_n^1 \mid s_1)$ given by the logit error, $\Pr(r_n^1) = \sum_{s_1 \in \Omega_k} \left[\Pr(s_1)\Pr(r_n^1 \mid s_1)\right]$ . Third, we update the state by the current lookup or $\Pr(s_1 \mid r_n^1) = \left[\Pr(s_1)\Pr(r_n^1 \mid s_1)\right] / \Pr(r_n^1)$ where terms in the numerator and the denominator are both derived in the second step. Fourth, for $\Pr(s_1 \mid r_n^1)$ derived in the third step and $\Pr(s_2 \mid s_1)$ given by the Markov transition matrix, we derive the next state from the current lookup, or

$$\Pr(s_2 \mid r_n^1) = \sum_{s_1 \in \Omega_k} \left[\Pr(s_1 \mid r_n^1)\Pr(s_2 \mid r_n^1, s_1)\right] = \sum_{s_1 \in \Omega_k} \left[\Pr(s_1 \mid r_n^1)\Pr(s_2 \mid s_1)\right]$$

where the second equality follows because by Markov, the transition to the next step only depends on the current state.

Fifth, for $\Pr(s_2 \mid r_n^1)$ given by the fourth step and $\Pr(r_n^2 \mid r_n^1, s_2) = \Pr(r_n^2 \mid s_2)$ given by the logit error, we derive the next lookup from the current lookup or $\Pr(r_n^2 \mid r_n^1) = \sum_{s_2 \in \Omega_k} \left[\Pr(s_2 \mid r_n^1)\Pr(r_n^2 \mid r_n^1, s_2)\right]$. Sixth, as in the third step, we update the state by the lookups up to now or

$$\Pr(s_2 \mid r_n^1, r_n^2) = \frac{\Pr(s_2 \mid r_n^1)\Pr(r_n^2 \mid r_n^1, s_2)}{\Pr(r_n^2 \mid r_n^1)}$$

where terms in the numerator and the denominator are both derived in the fifth step.

Seventh, as in the fourth step, for $\Pr(s_2 \mid r_n^1, r_n^2)$ derived in the sixth step and $\Pr(s_3 \mid s_2)$ given by the Markov transition matrix, we derive the next state from the lookups up to now or

$$\Pr(s_3 \mid r_n^1, r_n^2) = \sum_{s_2 \in \Omega_k} \left[ \Pr(s_2 \mid r_n^1, r_n^2) \Pr(s_3 \mid r_n^1, r_n^2, s_2) \right] = \sum_{s_2 \in \Omega_k} \left[ \Pr(s_2 \mid r_n^1, r_n^2) \Pr(s_3 \mid s_2) \right].$$

Eighth, as in the fifth step, for $\Pr(s_3 \mid r_n^1, r_n^2)$ given by the seventh step and $\Pr(r_n^3 \mid r_n^1, r_n^2, s_3) = \Pr(r_n^3 \mid s_3)$ given by the logit error, we derive the next lookup from the lookups up to now, or $\Pr(r_n^3 \mid r_n^1, r_n^2) = \sum_{s_3 \in \Omega_k} \left[ \Pr(s_3 \mid r_n^1, r_n^2) \Pr(r_n^3 \mid r_n^1, r_n^2, s_3) \right]$.

Continuing in this fashion and multiplying altogether the second step, the fifth step, the eighth step, and so on, we derive $\Pr(r_n^1)\Pr(r_n^2 \mid r_n^1)\Pr(r_n^3 \mid r_n^1, r_n^2)\ldots\Pr(r_n^{T_n} \mid r_n^1, r_n^2, \ldots, r_n^{T_n - 1})$ or (5). Regarding the assumption on the initial state, alternatively, we could follow the tradition in the Markov literature and assume uniform priors, or $\Pr(\mathbf{S}_0 = s_0) = 1/(k+1)$ for all $s_0 \in \Omega_k$. But this raises the question how subjects could figure out locations of higher states without even actually going through the best response hierarchy. This is the reason why we employ the current assumption that $\Pr(\mathbf{S}_0 = s_0) = 1$ when the initial state $s_0$ is ***0*** and zero otherwise.

## A4 Vuong's Test for Non-Nested But Overlapping Models

Let $Lk^*$ be the type which has the largest likelihood with corresponding parameters $(\lambda_{k^*}, \theta_{k^*})$. Let $Lk^a$ be an alternative type with corresponding parameters $(\lambda_{k^a}, \theta_{k^a})$. To test if these two competing types, $Lk^*$ and $Lk^a$, are equally good at explaining the true data, or it is the case that one of them is a better model, we choose a critical value from the standardized normal distribution. If the absolute value of the test statistic is no larger than the critical value, then we conclude that $Lk^*$ and $Lk^a$ are equally good at explaining the true data. If the test statistic is higher than the critical value, then we conclude that $Lk^*$ is a better model than $Lk^a$. Lastly, if the test statistic is less than the negative of the critical value, then we conclude that $Lk^a$ is a better model than $Lk^*$.

Equation (6) can be rearranged as

$$L(\lambda_k, \theta_k) = \sum_{n=1}^{24} lr_n(\lambda_k, \theta_k),$$

where $lr_n(\lambda_k, \theta_k) \equiv \ln f_n^k(r_n^1, \ldots, r_n^{T_n - 1}, r_n^{T_n})$. This indicates that we assume subject's lookups are independent across trials and follow the same Markov switching process, although each trial's lookups sequence may be serially-correlated.

To perform Vuong's test, we construct the log-likelihood ratio trial by trial:

$$m_n = lr_n(\lambda_{k^*}, \theta_{k^*}) - lr_n(\lambda_{k^a}, \theta_{k^a})$$ for trial *n*=1,..., *24*.

Let $\bar{m} = \frac{1}{N}\sum_{n=1}^{N} m_n$ , *N*=24. Vuong (1989) proposes a sequential procedure (p.321) for overlapping models. Its general result describes the behavior of

$$V = \frac{\sqrt{N}\left[\frac{1}{N}\sum_{n=1}^{N} m_n\right]}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(m_n - \bar{m})^2}},$$

when the sample variance $\omega_N^2 = \frac{1}{N}\sum_{n=1}^{N}(m_n - \bar{m})^2$ is significantly different from zero (the variance test). If the variance test is passed (which is the case for all of our subjects), $V$ has the property that (under standard assumptions):

(V1) If $Lk^*$ and $Lk^a$ are equivalently good at fitting the data,

$$V \xrightarrow{D} N(0,1);$$

(V2) if $Lk^*$ is better than $Lk^a$ at fitting the data,

$$V \xrightarrow{A.S.} \infty;$$

(V3) if $Lk^a$ is better than $Lk^*$ at fitting the data,

$$V \xrightarrow{A.S.} -\infty.$$

Hence, Vuong's test is performed by first conducting the variance test, then calculating $V$ and applying the above three cases depending on whether $V < -c$ , $|V| < c$ , or $V > c$ . (*c*=1.96 for *p*-value = 0.05.) Notice that this is the generalized version of the well-known "nested" Vuong's test, which does not require the variance test prior to calculating $V$.

Note that in our case $Lk^*$ is the type with the largest likelihood based on lookups, and the alternative type $Lk^a$ be the type having the next largest likelihood among all lower level types. Hence, either (V2) applies so that $Lk^*$ is a better model than $Lk^a$, or (V1) applies so that $Lk^*$ and $Lk^a$ are equally good (and we conservatively classify the subject as the second largest lower type $Lk^a$). (V3) does not apply since $V > 0$ by construct.

Two points are worth noting regarding Vuong's test in our setting. First, one might worry about non-identification issues caused by nuisance parameters when the two competing types are strictly nested and if the subject were truly *Lk^a^*. Hence, as an alternative, we perform Hansen's test (Hansen, 1992). Columns 7 and 8 of Supplementary Table 3 report the nearly identical results (to those based on Vuong's test). Secondly, we only perform Vuong's test once, and if we find

$Lk^*$ and $Lk^a$ explain the data equally well, we classify subjects as $Lk^a$, the lower level type that has the next largest likelihood. It is possible that the lower level type with the next-largest likelihood is still not different from the even lower level type with the even-next-largest likelihood (and so on). Hence, one might wonder whether we should stop here. Thus, we employ an iterative Vuong's test and classify subjects as the type that is, for the first time, significantly different from a lower level type of which the likelihood is immediate lower. We re-classify only two L2 subjects as L1, one L2 subject as L0 and two L1 subjects as L0, making the average number of thinking steps drop to 1.65. This provides a lower bound to the possible type distribution. The iterative Vuong's test result is reported in the sixth column of Supplementary Table 3.

## A5 Level-*k* Classification Based on Final Choices

We classify subjects into various (level-*k*) behavioral types based on their final choices using maximum likelihood estimation. In addition, a bootstrap procedure is employed to evaluate the robustness of the classification. In particular, similar to Costa-Gomes and Crawford (2006), we perform a maximum likelihood estimation to classify each individual subject into a particular behavioral level-*k* type. In particular, we model subjects following a particular level-*k* type but playing quantal response using the following logit error structure.[1] Let all possible level-*k* types be $k = 1,...,K$ and each subject goes through trial $n = 1,...,24$. For a given trial $n$, according to Hypothesis 1, a level-*k* subject *i*'s final choice is denoted as $c_n^k = \left(x_{i,n}^k, y_{i,n}^k\right) \in G_n$ where

$$G_n = \left\{(x, y) \middle| x \in \{-X_n, -X_n + 1, ..., X_n\}, y \in \{-Y_n, -Y_n + 1, ..., Y_n\}\right\}$$

is the finite countable choice set for trial $n$. $|G_n| = (2X+1)(2Y+1)$ is the number of elements in $G_n$, which depends on the map size (*X, Y)* of the game in that particular trial.[2] For any two elements of $G_n$, $g_1 = (x_1, y_1), g_2 = (x_2, y_2)$, their distance is defined as $\|g_1 - g_2\| = |x_1 - x_2| + |y_1 - y_2|$, i.e. the "steps" on the map (the sum of vertical and horizontal distance) between $g_1, g_2$. Then, if a

[1] Since we do not have a large choice set as in Costa-Gomes and Crawford (2006), we employ a "logit" specification instead of a "spike-logit" specification to describe the error structure of subjects' choices.

[2] For instance, as shown in Figure 1, the grid map of Game 16 (as listed in Table 1) has a choice set of $G_n = \{(x, y) \mid x \in \{-3, -2, ..., 3\}, y \in \{-3, -2, ..., 3\}\}$ consisting of the 7x7 = 49 locations.

subject chooses a location $g_n = (x_{i,n}, y_{i,n})$ in trial *n*, the distance between her choice $g_n$ and the choice of a level-*k* subject $c_n^k$ is $\left\| g_n - c_n^k \right\| = \left| x_{i,n} - x_{i,n}^k \right| + \left| y_{i,n} - y_{i,n}^k \right|$. In a logit error model with precision $\lambda_k$, the probability of observing $g_n$ is

$$d^k(g_n) = \frac{\exp\left(-\lambda_k \times \| g_n - c_n^k \|\right)}{\sum_{g \in G_n} \exp\left(-\lambda_k \times \| g - c_n^k \|\right)}$$

When $\lambda_k \to 0$, $d^k(g_n) = \dfrac{1}{|G_n|}$ and the subject randomly chooses from the choice set $G_n$. As $\lambda_k \to \infty$, $d^k(g_n) = \begin{cases} 1, \text{ if } g_n = c_n^k \\ 0, \text{ if } g_n \neq c_n^k \end{cases}$ and the choice of the subject approaches to the level-*k* choice $c_n^k$. The log likelihood over all trials with choices $(g_1, g_2, ..., g_{24})$ trial-by-trial can then be expressed as

$$\ln \prod_{n=1}^{24} d^k\left(g_n\right). \tag{1}$$

For each *k*, we estimate the precision parameter $\lambda_k$ by fitting the data with the logit error model to maximize empirical likelihood. Then we choose the *k* which maximizes the empirical likelihood and classify the subject into this particular level-*k* type. We consider all the level-*k* types separable in our games: *L0*, *L1*, *L2*, *L3*, and *EQ*. Results are reported in column (B) of Table 3. Among the 17 subjects, there are two *L0*, four *L1*, four *L2*, four *L3*, and three *EQ*. The average number of thinking steps is 2.12, similar to the lookup-based classifications.[3]

One possible concern is whether some subjects do not follow any of the pre-specified level-*k* types, and hence, the model is misspecified. To incorporate all empirically possible behavioral types, we follow Costa-Gomes and Crawford (2006) and perform the pseudotype test by including 17 pseudotypes, each constructed from one of our subject's choices in 24 trials. This is to see whether there are clusters of subjects whose choices resemble each other's and thus predict other's choices in the cluster better than the pre-specified level-*k* types. We report results of the pseudotype test in Supplementary Table 1 where pseudo-*i* is the pseudotype constructed from subject-*i*,. We find that two subjects (subject 3 and subject 17) have likelihoods for each other's pseudotype higher than all other types. So, based on the same criteria of Costa-Gomes and

[3] We treat the *EQ* type as having a thinking step of 4 in calculating the average number of thinking steps.

Crawford (2006), these two subjects could be classified as a cluster (pseudo-17). In other words, there may be a cluster of pseudo-17 type subjects (subjects 3 and 17) whose behaviors are not explained well by the predefined level-$k$ types. Despite of this, there are still 15 subjects out of 17 who can be classified into level-$k$ types, comparable to Costa-Gomes and Crawford (2006), who find 12.5% (11/88) of their subjects fail the pseudotype test and could be classified as 5 different clusters. Table 3 lists the classification with and without pseudotypes in columns (C) and (B) respectively. The distribution of level-$k$ types in column (C) of Table 3 does not change much even if we include pseudotypes, having two *L0*, three *L1*, four *L2*, three *L3*, and three *EQ*. The average of thinking steps is 2.13, nearly identical to that without pseudotypes.[4] This suggests that in our games, the level-$k$ classification is quite robust to empirically omitted types that explain more than one subject. In other words, Hypothesis 1 is confirmed is the sense that most subjects indeed follow the prediction of a particular level-$k$ type for choices, and few alternative models can explain the behavior of more than one subject.[5]

[4] In calculating the average number of thinking steps, we ignore the two pseudo-17 subjects. For these two pseudo-17 subjects, one is re-classified as *L1*, and the other *L3* when pseudotypes are not included.

[5] Given that we have only seventeen subjects, it is true that we cannot rule out the possibility that our small pool of subjects did not capture all possible behavioral types. However, Costa-Gomes and Crawford (2006) also find few omitted types in their pool of 88 subjects.

Supplementary Table 1: Subject's Maximized Likelihood for Various Level-$k$ Types and Pseudotypes Based on Final Choices

| *Subject* | *Lk type* | | | | | *Pseudotype* | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *L0* | *L1* | *L2* | *L3* | *EQ* | *pseudo-1* | *pseudo-2* | *pseudo-3* | *pseudo-4* | *pseudo-5* | *pseudo-6* | *pseudo-7* | *pseudo-8* | *pseudo-9* | *pseudo-10* | *pseudo-11* | *pseudo-12* | *pseudo-13* | *pseudo-14* | *pseudo-15* | *pseudo-16* | *pseudo-17* | |
| 1 | -98.89 | -88.92 | -79.24 | -52.32 | -67.21 | . | -98.27 | -98.89 | -97.34 | -81.26 | -85.83 | -98.89 | -84.57 | -98.57 | -92.07 | -77.82 | -91.05 | -97.86 | -97.81 | -85.92 | -86.76 | -98.89 | L3 |
| 2 | -97.16 | -93.70 | -81.14 | -85.76 | -90.68 | -96.12 | . | -97.47 | -94.69 | -97.09 | -91.68 | -98.47 | -95.52 | -96.63 | -98.55 | -92.85 | -97.14 | -86.30 | -93.32 | -97.06 | -89.75 | -97.60 | L2 |
| 3 | -98.89 | -93.70 | -97.84 | -93.13 | -95.03 | -98.89 | -98.76 | . | -98.89 | -98.89 | -98.89 | -96.50 | -98.89 | -98.74 | -98.89 | -98.89 | -98.89 | -95.56 | -95.28 | -98.89 | -98.87 | **-61.67** | pseudo-17 |
| 4 | -96.92 | -90.37 | -96.82 | -95.96 | -97.36 | -95.15 | -95.24 | -98.89 | . | -95.27 | -96.40 | -98.55 | -96.69 | -97.76 | -91.42 | -93.75 | -93.73 | -98.82 | -98.89 | -95.38 | -96.31 | -98.85 | L1 |
| 5 | -98.89 | -97.28 | -86.53 | -82.12 | -70.69 | -83.19 | -98.89 | -98.89 | -98.21 | . | -87.17 | -98.89 | -90.70 | -98.89 | -98.04 | -91.52 | -97.98 | -97.25 | -95.28 | -94.69 | -93.74 | -98.89 | EQ |
| 6 | -98.89 | -90.71 | -69.22 | -78.98 | -73.24 | -84.16 | -93.91 | -98.89 | -97.49 | -83.63 | . | -98.89 | -82.34 | -98.32 | -91.09 | -79.83 | -90.67 | -94.70 | -97.29 | -89.75 | -80.90 | -98.89 | L2 |
| 7 | -94.15 | -98.89 | -98.69 | -98.88 | -98.87 | -98.40 | -98.17 | **-93.04** | -98.00 | -98.89 | -98.89 | . | -98.89 | -96.99 | -98.89 | -98.64 | -98.67 | -97.75 | -98.59 | -98.69 | -98.64 | **-91.95** | L0 |
| 8 | -98.89 | -93.43 | -89.58 | -81.62 | -70.02 | -87.23 | -98.71 | -98.89 | -98.86 | -91.38 | -86.73 | -98.89 | . | -98.89 | -96.49 | -91.17 | -98.09 | -98.25 | -96.93 | -94.93 | -90.87 | -98.89 | EQ |
| 9 | -92.52 | -97.28 | -93.37 | -95.39 | -95.76 | -96.80 | -96.74 | -97.47 | -97.49 | -97.35 | -97.21 | -97.55 | -98.34 | . | -98.66 | -95.98 | -98.38 | -95.95 | -97.51 | -95.60 | -94.78 | -98.25 | L0 |
| 10 | -98.89 | -39.73 | -93.07 | -86.98 | -94.00 | -90.22 | -98.89 | -98.89 | -92.67 | -96.01 | -90.68 | -98.89 | -93.16 | -98.89 | . | -87.94 | -75.40 | -98.89 | -98.89 | -89.36 | -93.45 | -98.89 | L1 |
| 11 | -97.77 | -85.59 | -68.30 | -70.72 | -77.28 | -74.45 | -93.62 | -98.89 | -93.88 | -86.59 | -78.18 | -98.88 | -85.40 | -96.43 | -86.75 | . | -82.33 | -95.75 | -98.54 | -80.56 | -74.79 | -98.89 | L2 |
| 12 | -98.24 | -72.01 | -86.53 | -84.90 | -92.83 | -86.82 | -97.09 | -98.89 | -92.99 | -94.22 | -88.02 | -98.82 | -93.87 | -98.32 | -73.09 | -81.09 | . | -98.89 | -98.89 | -84.44 | -89.75 | -98.89 | L1 |
| 13 | -98.89 | -85.59 | -81.74 | -72.14 | -79.36 | -97.74 | -90.12 | -95.65 | -98.89 | -96.01 | -95.63 | -98.89 | -96.97 | -98.01 | -98.89 | -97.55 | -98.89 | . | -75.94 | -98.51 | -92.86 | -94.94 | L3 |
| 14 | -98.89 | -84.17 | -96.99 | -76.67 | -74.45 | -98.89 | -98.83 | -98.60 | -98.89 | -97.81 | -98.89 | -98.89 | -98.47 | -98.89 | -98.89 | -98.89 | -98.89 | -83.53 | . | -98.89 | -98.89 | -98.56 | EQ |
| 15 | -98.89 | -90.71 | -84.53 | -82.12 | -86.72 | -83.19 | -97.94 | -98.89 | -96.09 | -91.08 | -88.82 | -98.89 | -90.70 | -96.63 | -88.87 | **-81.09** | -86.22 | -97.75 | -98.81 | . | -88.95 | -98.89 | L3 |
| 16 | -98.89 | -92.88 | -57.64 | -80.59 | -80.35 | -83.68 | -90.91 | -98.39 | -96.67 | -89.50 | -79.42 | -98.89 | -85.40 | -95.56 | -92.69 | -74.91 | -91.05 | -90.50 | -95.78 | -88.56 | . | -98.85 | L2 |
| 17 | -98.89 | -92.88 | -98.17 | -95.19 | -97.36 | -98.89 | -98.59 | **-60.42** | -98.89 | -98.89 | -98.89 | -94.87 | -98.89 | -98.85 | -98.89 | -98.89 | -98.89 | -93.99 | -94.35 | -98.89 | -98.89 | . | pseudo-17 |

Note: Each row corresponds to a subject and contains the following information in order: subject number, the likelihood of various level-$k$ types, the likelihood of various pseudotypes (excluding the pseudotype corresponding to the subject herself), and the type with the largest likelihood of level-$k$ types (listed in the last column) unless a pseudotype cluster is identified. Note that the likelihoods are based on choice data and the type with the largest likelihood among the various level-$k$ types is underlined, which the largest likelihood among the pseudotypes are in **bold** when it is higher than that of all level-$k$ types. The likelihood of a subject's lookup type, when different from her choice type, is double underlined.

Supplementary Table 2: Comparisons on Final Choices of GRAPH and SEPARATE Presentations

| *Subject* | *Average Difference* | | | | *Fitting GRAPH data (with logit)* | |
|---|---|---|---|---|---|---|
| | in X axis | (s.e.) | in Y axis | (s.e.) | lambda | (s.e.) |
| 1 | 0.032 | (0.209) | 0.127 | (0.315) | 0.542* | (0.108) |
| 2 | 0.032 | (0.406) | 0.196 | (0.363) | 0.085 | (0.074) |
| 3 | -0.048 | (0.112) | -0.024 | (0.140) | 1.287* | (0.195) |
| 4 | -0.066 | (0.243) | -0.012 | (0.271) | 0.426* | (0.103) |
| 5 | 0.141 | (0.268) | 0.152 | (0.329) | 0.272* | (0.091) |
| 6 | -0.033 | (0.255) | -0.075 | (0.154) | 0.583* | (0.110) |
| 7 | -0.017 | (0.293) | 0.073 | (0.260) | 0.256* | (0.095) |
| 8 | 0.029 | (0.233) | 0.056 | (0.311) | 0.374* | (0.096) |
| 9 | -0.109 | (0.213) | -0.103 | (0.420) | 0.177* | (0.081) |
| 10 | 0.044 | (0.210) | 0.023 | (0.131) | 0.857* | (0.149) |
| 11 | -0.071 | (0.142) | 0.000 | (0.235) | 0.680* | (0.126) |
| 12 | 0.029 | (0.186) | 0.066 | (0.140) | 0.696* | (0.129) |
| 13 | 0.122 | (0.218) | 0.077 | (0.183) | 0.449* | (0.109) |
| 14 | -0.006 | (0.029) | -0.004 | (0.113) | 2.061* | (0.354) |
| 15 | 0.014 | (0.176) | -0.039 | (0.242) | 0.579* | (0.119) |
| 16 | 0.064 | (0.193) | 0.351 | (0.286) | 0.150 | (0.087) |
| 17 | -0.003 | (0.175) | 0.035 | (0.145) | 0.765* | (0.138) |

Note: * denotes significance at the 0.05 level.

Each row corresponds to a subject and contains the following information in order: (1) subject number, (2) the average difference of choices in the X axis in the two presentations (standard errors in parentheses), (3) the average difference of choices in the Y axis in the two presentations (standard errors in parentheses), (4) the precision parameter of the logit error if we treat each subject's choice in the SEPARATE presentation as a pseudotype to fit her choice in the GRAPH presentation. Notice that if a subject's choices in the two presentations are similar, we should expect that her average difference of choices in either the X axis or the Y axis should not be significantly different from zero. Moreover, we should expect that her choices in the SEPARATE presentation as a pseudotype can predict her choices in the GRAPH presentation well and hence the precision parameter in the error structure should be significantly different from zero. This is indeed the case. For none of the subjects, the average difference in either axis is significantly different from zero. For fifteen of the seventeen subjects, the precision parameters are significantly different from zero.

Supplementary Table 3: Distribution of Types Based on Hansen's Test and Iterative Vuong's Test

| *subject* | $Lk^*$ | $Lk^a$ | Vuong's ($V$) | $Lk^l$ | *Iterative Vuong* | Hansen's $p$-value | $Lk^{Hansen}$ |
|---|---|---|---|---|---|---|---|
| 1 | L3 | L2 | 4.425+ | L3 | - | - | - |
| 2 | L3 | L2 | 0.689 | L2 | L1 | - | - |
| 3 | L3 | L1 | 1.577 | L1 | L0 | 0.084 | L1 |
| 4 | L3 | L1 | 1.597 | L1 | - | 0.095 | L1 |
| 5 | EQ | L2 | 2.977+ | EQ | - | 0.023* | EQ |
| 6 | EQ | L2 | 2.400+ | EQ | - | 0.053 | L2 |
| 7 | L2 | L0 | 1.582 | L0 | - | 0.769 | L0 |
| 8 | L3 | L1 | 2.812+ | L3 | - | 0.025* | L3 |
| 9 | EQ | L2 | 1.001 | L2 | - | 0.498 | L2 |
| 10 | L3 | L1 | 1.226 | L1 | - | 0.497 | L1 |
| 11 | L3 | L2 | 2.087+ | L3 | - | - | - |
| 12 | L3 | L1 | 0.853 | L1 | - | 0.500 | L1 |
| 13 | L3 | L1 | 3.939+ | L3 | - | 0.015* | L3 |
| 14 | L3 | L1 | 1.692 | L1 | L0 | 0.096 | L1 |
| 15 | L3 | L2 | 1.470 | L2 | L0 | - | - |
| 16 | L3 | L2 | 1.342 | L2 | L1 | - | - |
| 17 | L3 | L1 | 1.778 | L1 | - | 0.082 | L1 |

Note: + indicates the Vuong statistic $V$ is significant or $|V|>1.96$.

* indicates $p$-value less than 0.05.

Each row corresponds to a subject and contains the following information in order: (1) subject number, (2) based on her lookups, the type with the largest likelihood (denoted by $Lk^*$, also reported under the $Lk^*$ column of Table 2), (3) based on her lookups, the alternative lower level type which has the next largest likelihood (denoted by $Lk^a$, also reported under the $Lk^a$ column of Table 2), (4) Vuong's statistic in testing whether $Lk^*$ and $Lk^a$ are equally good models, (5) subject's type based on Vuong' test result, (6) types of those subjects who, based on the iterative Vuong's test, have different types from those based on Vuong's test, (7) p-value in Hansen's test if $Lk^*$ and $Lk^a$ are strictly nested, (8) types based on Hansen's test result when $Lk^*$ and $Lk^a$ are strictly nested. Note that among the twelve subjects whose $Lk^*$ and $Lk^a$ are strictly nested, Vuong's test results are almost identical to Hansen's test results. The only exception is that for subject 6 (her $Lk^{Hansen}$ type underlined), her Vuong's test statistic $V$ is 2.400>1.96 while the p-value of the Hansen's test is 0.053, at the margin of significance.

Supplementary Table 4: Distribution of Types Based on Lookup Data (Unconstrained Markov)

| subject | Unconstrained | Unconstrained | Vuong's | Unconstrained | Constrained |
|---|---|---|---|---|---|
| | $Lk^{*}$ | $Lk^{a}$ | statistic $V$ | $Lk^{j}$ | $Lk^{j}$ |
| 1 | L3 | L2 | 3.818 + | L3 | L3 |
| 2 | L3 | L2 | 0.704 | L2 | L2 |
| 3 | EQ | L3 | 0.824 | L3 | L1 |
| 4 | L3 | L1 | 1.740 | L1 | L1 |
| 5 | EQ | L2 | 4.482 + | EQ | EQ |
| 6 | EQ | L3 | 0.890 | L3 | EQ |
| 7 | L2 | L0 | 1.385 | L0 | L0 |
| 8 | EQ | L3 | 0.929 | L3 | L3 |
| 9 | EQ | L2 | 1.381 | L2 | L2 |
| 10 | L3 | L1 | 1.600 | L1 | L1 |
| 11 | L3 | L2 | 2.290 + | L3 | L3 |
| 12 | L3 | L1 | 1.480 | L1 | L1 |
| 13 | L3 | L1 | 4.224 + | L3 | L3 |
| 14 | L3 | L1 | 2.301 + | L3 | L1 |
| 15 | L3 | L2 | 2.149 + | L3 | L2 |
| 16 | L3 | L2 | 0.881 | L2 | L2 |
| 17 | L3 | L1 | 2.191 + | L3 | L1 |

Note: + indicates Vuong's statistic $V$ is significant or $|V|>1.96$.

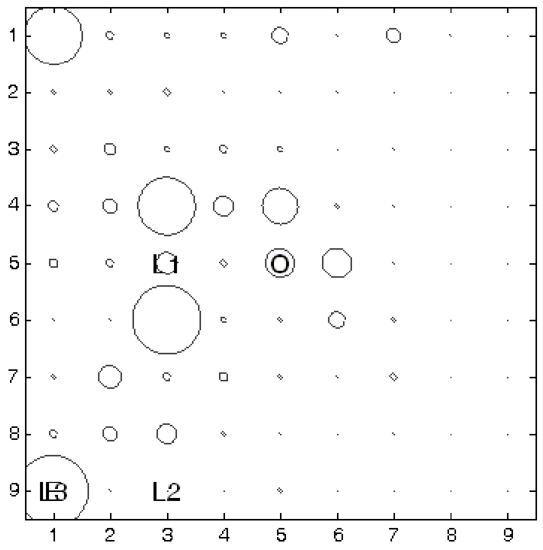
Unconstrained $Lk^{*}$ denotes the type with the largest likelihood.

Unconstrained $Lk^{a}$ denotes the alternative lower level type which has the next-largest likelihood.

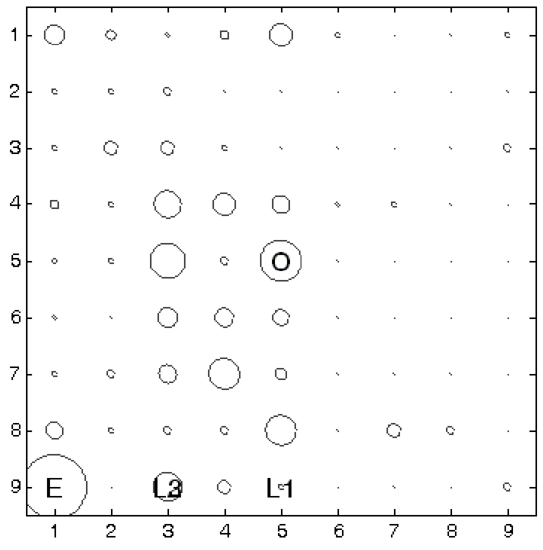
Unconstrained $Lk^{j}$ denotes the classified type based on Vuong's test result for the unconstrained Markov-switching model.

Constrained $Lk^{j}$ denotes the classified type based on the constrained Markov model.

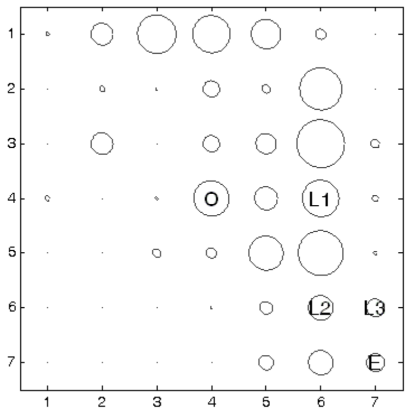
Each row corresponds to a subject and contains the following information in order: (1) the subject number, (2) based on her lookups, the type with the largest likelihood using the unconstrained Markov-switching model, (3) based on her lookups, the alternative lower level type which has the next-largest likelihood, (4) Vuong's statistic in testing whether $Lk^{*}$ and $Lk^{a}$ are equally good models, (5) subject's lookup type based on Vuong's test result, (6) subject's lookup type based on the constrained Markov-switching model (as reported in the fifth column of Table 2). Any difference between (5) and (6) are underlined. Notice that in (5) we classify a subject as her $Lk^{*}$ type if according to Vuong's test, $Lk^{*}$ is a better model than $Lk^{a}$. On the other hand, if $Lk^{*}$ and $Lk^{a}$ are equally good models, since $Lk^{a}$ has fewer parameters, to avoid overfitting, we classify a subject as her $Lk^{a}$ type.



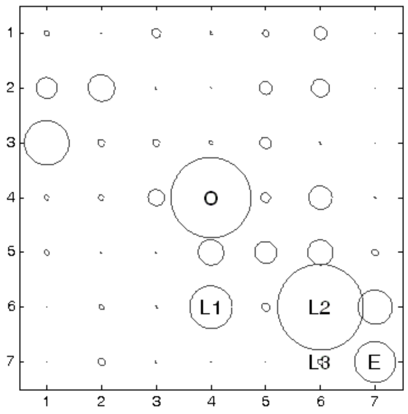
Supplementary Figure 1: Aggregate Empirical Percentage of Time Spent on Each Location for Game 1 with 1-dimensional Targets (-2,0) (own) and (0,-4) (opponent) on a 9x9 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-*k* types.



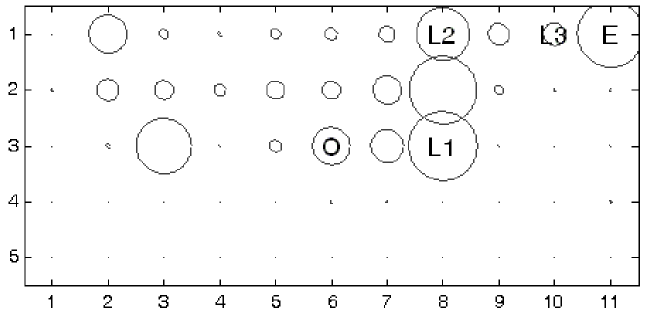
Supplementary Figure 2: Aggregate Empirical Percentage of Time Spent on Each Location for Game 2 with 1-dimensional Targets (0,-4) (own) and (-2,0) (opponent) on a 9x9 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-*k* types.



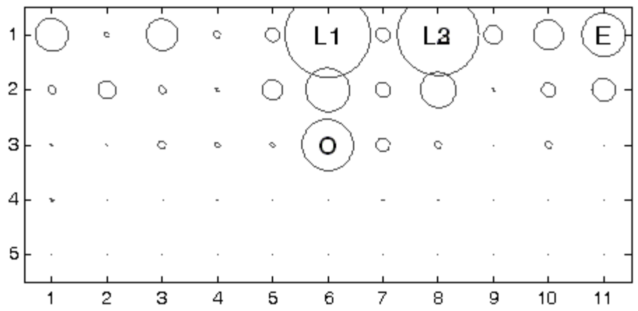
Supplementary Figure 3: Aggregate Empirical Percentage of Time Spent on Each Location for Game 3 with 1-dimensional Targets (2, 0) (own) and (0,-2) (opponent) on a 7x7 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-$k$ types.



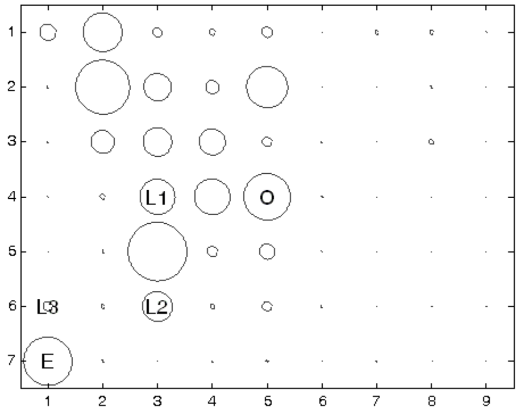
Supplementary Figure 4: Aggregate Empirical Percentage of Time Spent on Each Location for Game 4 with 1-dimensional Targets (0,-2) (own) and (2, 0) (opponent) on a 7x7 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-$k$ types.



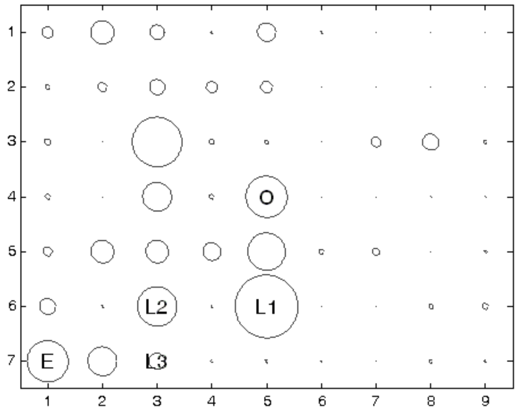
Supplementary Figure 5: Aggregate Empirical Percentage of Time Spent on Each Location for Game 5 with 1-dimensional Targets (2, 0) (own) and (0, 2) (opponent) on an 11x5 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-*k* types.



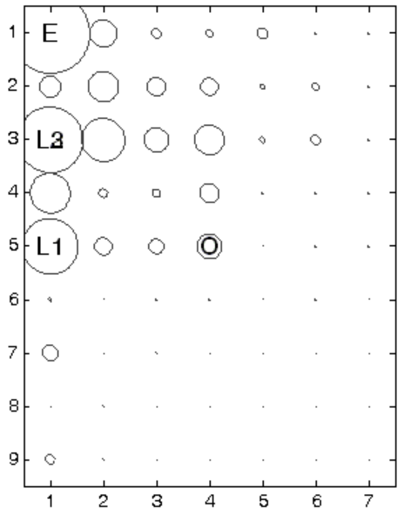
Supplementary Figure 6: Aggregate Empirical Percentage of Time Spent on Each Location for Game 6 with 1-dimensional Targets (0,2) (own) and (2, 0) (opponent) on an 11x5 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-*k* types.



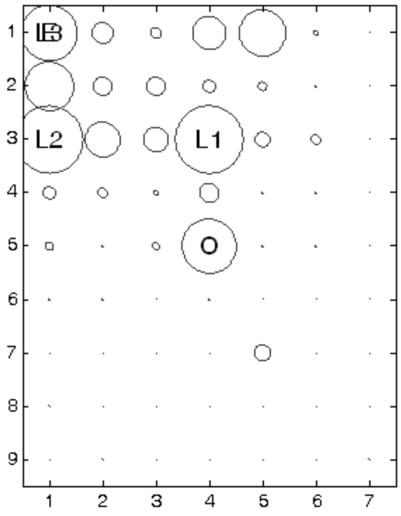
Supplementary Figure 7: Aggregate Empirical Percentage of Time Spent on Each Location for Game 7 with 1-dimensional Targets (-2, 0) (own) and (0, -2) (opponent) on a 9x7 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-*k* types.



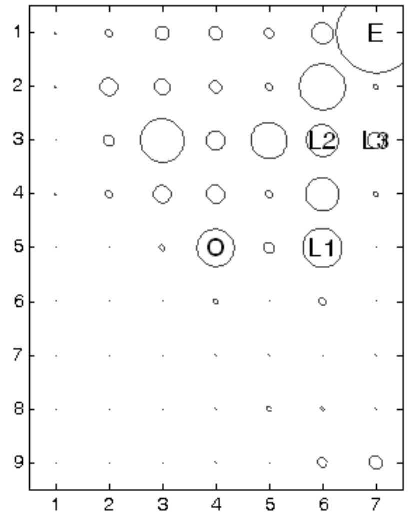
Supplementary Figure 8: Aggregate Empirical Percentage of Time Spent on Each Location for Game 8 with 1-dimensional Targets (0, -2) (own) and (-2, 0) (opponent) on a 9x7 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-*k* types.



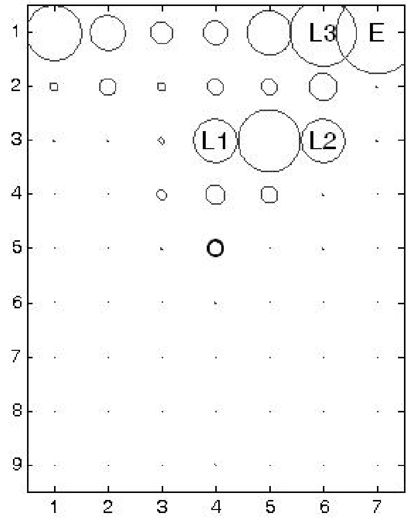
Supplementary Figure 9: Aggregate Empirical Percentage of Time Spent on Each Location for Game 9 with 1-dimensional Targets (-4, 0) (own) and (0, 2) (opponent) on a 7x9 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-*k* types.



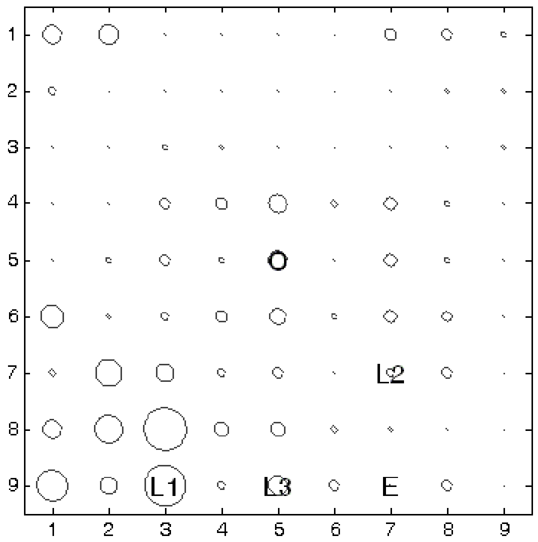
Supplementary Figure 10: Aggregate Empirical Percentage of Time Spent on Each Location for Game 10 with 1-dimensional Targets (0, 2) (own) and (-4, 0) (opponent) on a 7x9 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-*k* types.



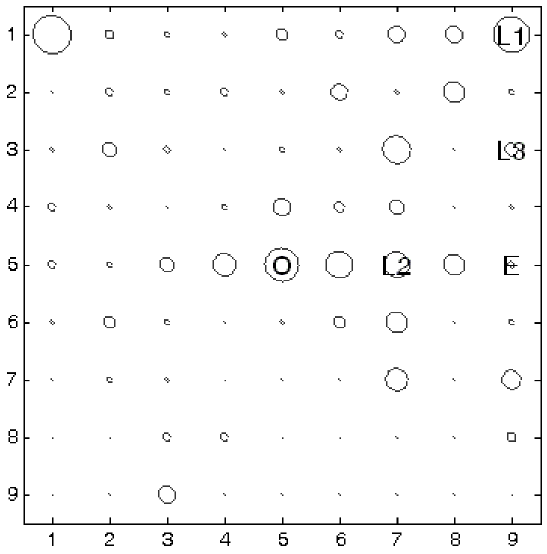
Supplementary Figure 11: Aggregate Empirical Percentage of Time Spent on Each Location for Game 11 with 1-dimensional Targets (2, 0) (own) and (0, 2) (opponent) on a 7x9 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-*k* types.



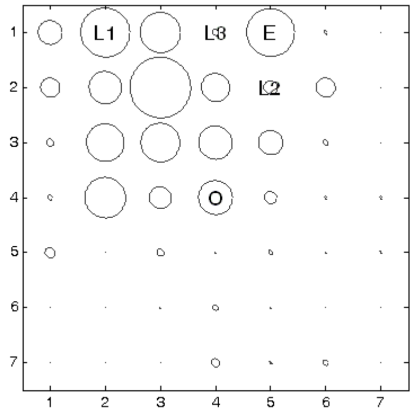
Supplementary Figure 12: Aggregate Empirical Percentage of Time Spent on Each Location for Game 12 with 1-dimensional Targets (0, 2) (own) and (2, 0) (opponent) on a 7x9 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-*k* types.



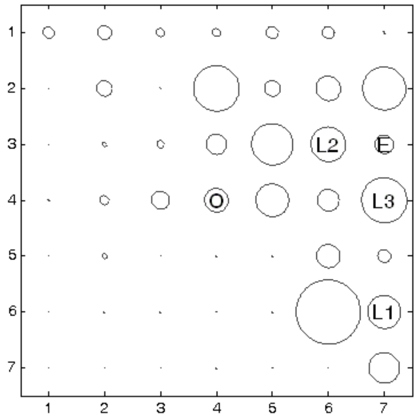
Supplementary Figure 13: Aggregate Empirical Percentage of Time Spent on Each Location for Game 13 with 2-dimensional Targets (-2, 6) (own) and (4, 4) (opponent) on a 9x9 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-*k* types.



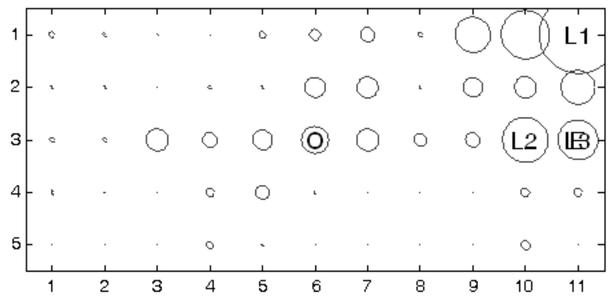
Supplementary Figure 14: Aggregate Empirical Percentage of Time Spent on Each Location for Game 14 with 2-dimensional Targets (4, 4) (own) and (-2, 6) (opponent) on a 9x9 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-*k* types.



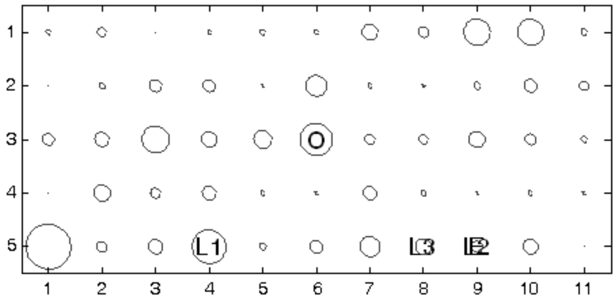
Supplementary Figure 15: Aggregate Empirical Percentage of Time Spent on Each Location for Game 15 with 2-dimensional Targets (-2, 4) (own) and (4, -2) (opponent) on a 7x7 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-$k$ types.



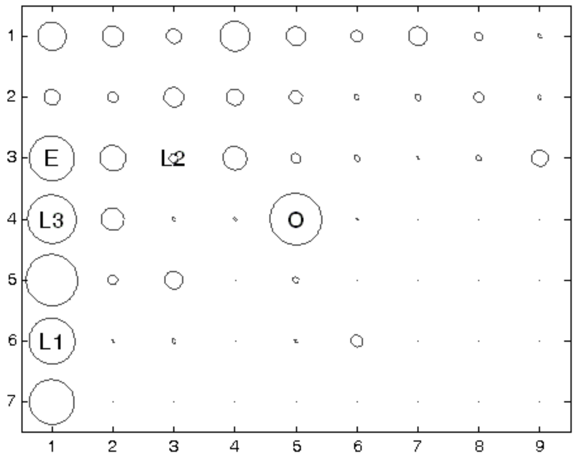
Supplementary Figure 16: Aggregate Empirical Percentage of Time Spent on Each Location for Game 16 with 2-dimensional Targets (4, -2) (own) and (-2, 4) (opponent) on a 7x7 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-$k$ types.



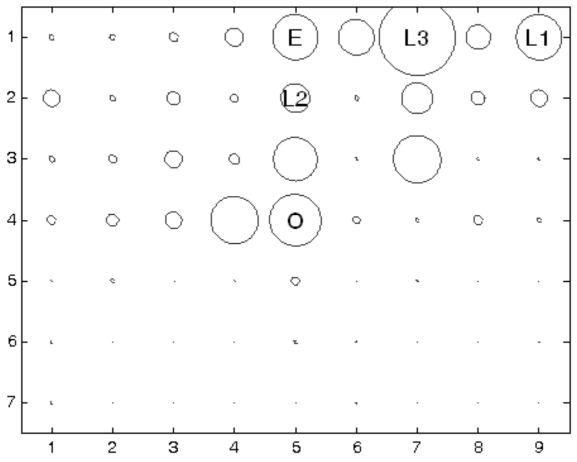
Supplementary Figure 17: Aggregate Empirical Percentage of Time Spent on Each Location for Game 17 with 2-dimensional Targets (6, 2) (own) and (-2, -4) (opponent) on an 11x5 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1,..., E are player *i*'s predicted choices of various level-*k* types.



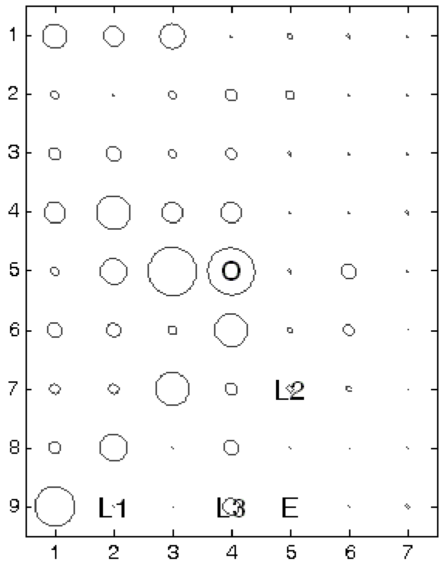
Supplementary Figure 18: Aggregate Empirical Percentage of Time Spent on Each Location for Game 18 with 2-dimensional Targets (6, 2) (own) and (-2, -4) (opponent) on a 11x5 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-*k* types.



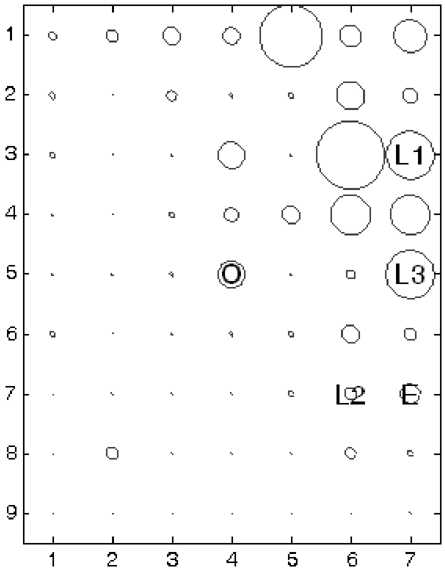
Supplementary Figure 19: Aggregate Empirical Percentage of Time Spent on Each Location for Game 19 with 2-dimensional Targets (-6, -2) (own) and (4, 4) (opponent) on a 9x7 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-*k* types.



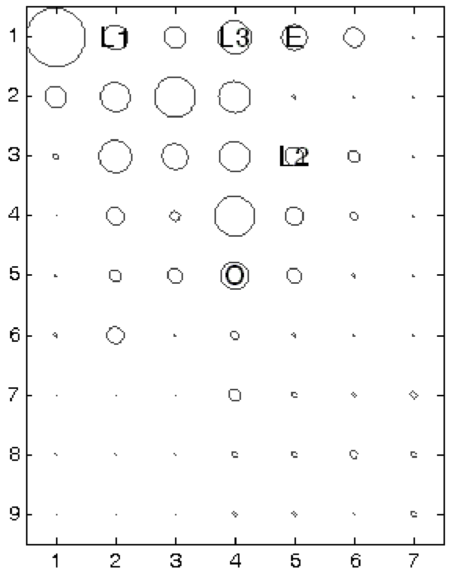
Supplementary Figure 20: Aggregate Empirical Percentage of Time Spent on Each Location for Game 20 with 2-dimensional Targets (4, 4) (own) and (-6, -2) (opponent) on a 9x7 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-*k* types.



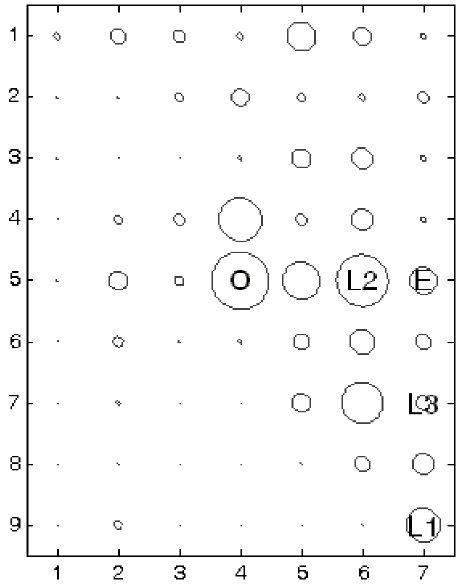
Supplementary Figure 21: Aggregate Empirical Percentage of Time Spent on Each Location for Game 21 with 2-dimensional Targets (-2, -4) (own) and (4, 2) (other) on a 7x9 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-*k* types.



Supplementary Figure 22: Aggregate Empirical Percentage of Time Spent on Each Location for Game 22 with 2-dimensional Targets (4, 2) (own) and (-2, -4) (other) on a 7x9 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-*k* types.



Supplementary Figure 23: Aggregate Empirical Percentage of Time Spent on Each Location for Game 23 with 2-dimensional Targets (-2, 6) (own) and (4, -4) (other) on a 7x9 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-$k$ types.



Supplementary Figure 24: Aggregate Empirical Percentage of Time Spent on Each Location for Game 24 with 2-dimensional Targets (4, -4) (own) and (-2, 6) (other) on a 7x9 map. The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, ..., E are player *i*'s predicted choices of various level-$k$ types.

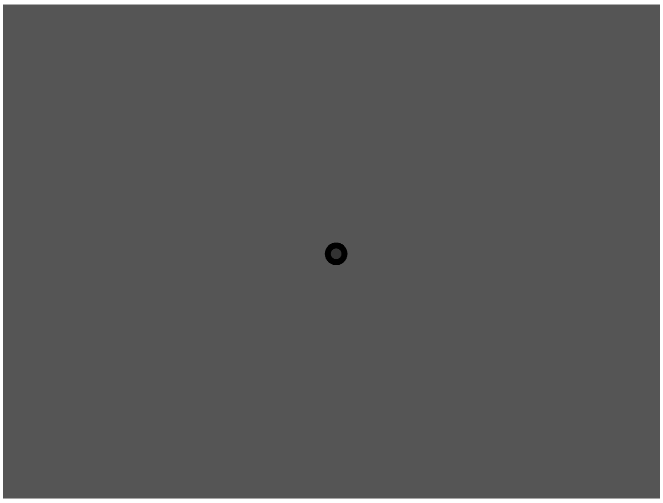
Sample Instructions:

## EXPERIMENT INSTRUCTIONS

The experiment you are participating in consists of 48 rounds. At the end, you will be paid the amount you have earned from THREE randomly drawn rounds, plus a $20 show-up fee. Everybody will be paid in private, and you are under no obligation to tell others how much you earned. During the experiment all the earnings are denominated in FRANCS. Your dollar earnings are determined by the FRANC/$ exchange rate: 3 FRANCS = $1.

You will wear an eye-tracking device which will track your eye movements. Please make sure you are not wearing contact lenses. You will be seated in front of the computer screen, showing the earnings tables, and make your choice by looking at the boxes on the screen. When looking at a box, it will light up, and will become your choice of action if you hit "space".

At the beginning of the session, the experimenter will adjust and calibrate the eye-tracker. To perform a calibration, look at the center of the screen (black dot) and hit space **once**. Then, the dot will disappear and move to a new location. Follow the black dot with your eyes and fixate at the new location until it disappears again. This procedure will be repeated until the dot returns to the center. (The same procedure will be repeated to validate the calibration.) At the start of each round, you will perform a self-correction by looking at the center of the screen (black dot) and hit the space bar.



### Roles

You and the other participant are paired to form a group, in which one participant will be member A, and the other member B. The roles of member A and B will be decided randomly by a die roll and you will maintain the same roles throughout the experiment.

### The Decision

There are 3 practice rounds and 48 real rounds. In each round, each of you simultaneously chooses a location (X, Y) on a given map, and your earnings are determined by your location and the other participant's location. In particular, each of you will have a "goal" which (together with the other participant's location) determines your "target location" for each round. Then, your earnings will be determined by how close you hit your target location.

For example, suppose the map consists of X=1~5 and Y=1~7, and your goals are:

| Member A | |
|---|---|
| LEFT 2 | |

| Member B | |
|---|---|
| | BELOW 4 |

(This means that member A's target location is to choose **two blocks to the LEFT** of member B's location, while member B's target location is to choose **four blocks below** member A's location.)

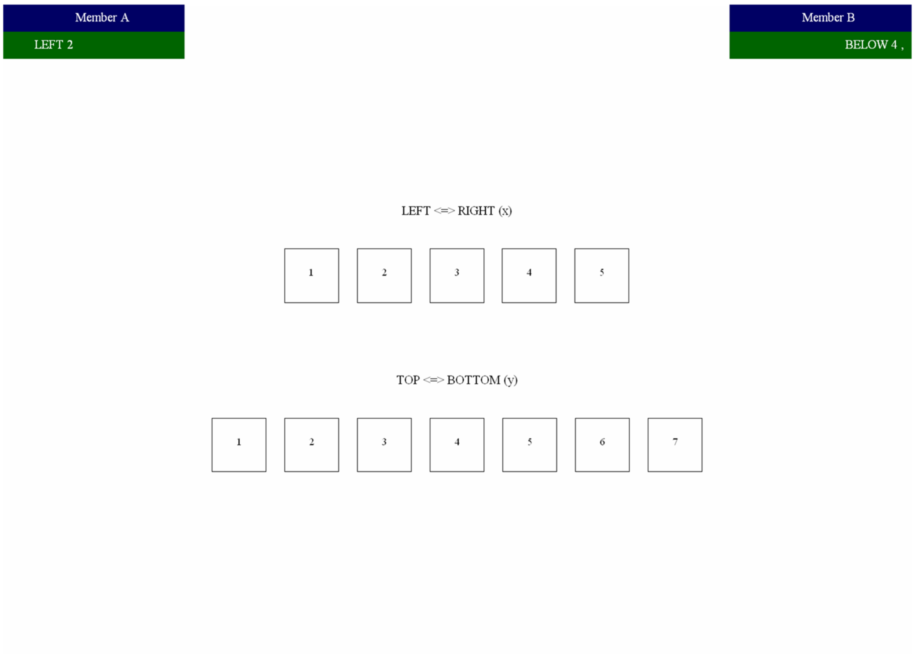
Suppose member A's location is ($X_a$, $Y_a$), and member B's location is ($X_b$, $Y_b$). The target location for member A is ($\mathbf{X_b - 2}$, $Y_b$), and the target location of member B is ($X_a$, $\mathbf{Y_a + 4}$). The earnings for member A is (in FRANCS):

$$20 - |\ X_a - (\mathbf{X_b - 2})\ | - |\ Y_a - (Y_b + 0)\ |$$

While the earnings for member B is (in FRANCS):

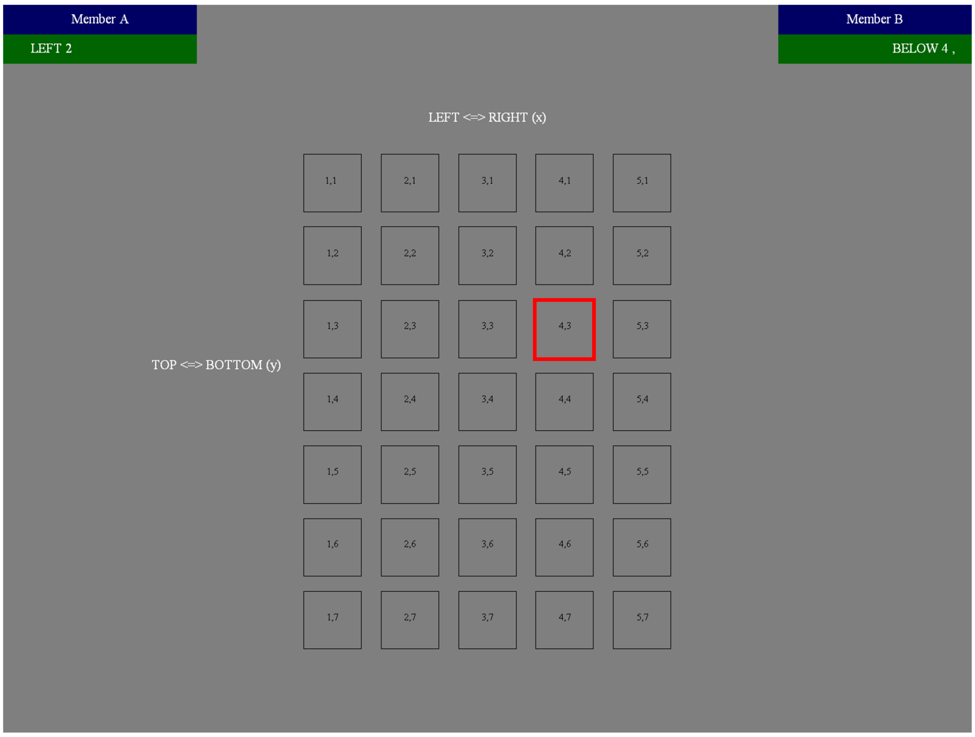
$$20 - |\ X_b - (X_a + 0)\ | - |\ Y_b - (\mathbf{Y_a + 4})\ |$$

Note that the target location may be outside the map so you might not achieve 20. Also, note that the X's increase from left to right, and the Y's increase from top to bottom.



In each round, you will make a similar decision with **different** goals on a **different** map, which is shown to both sides. **However, no feedback will be provided after each round.**

In each round, the goals of member A and B will be shown on the top-left and top-right corner. When you look at a location (X, Y), it will light up with a red frame.



Member A

LEFT 2

Member B

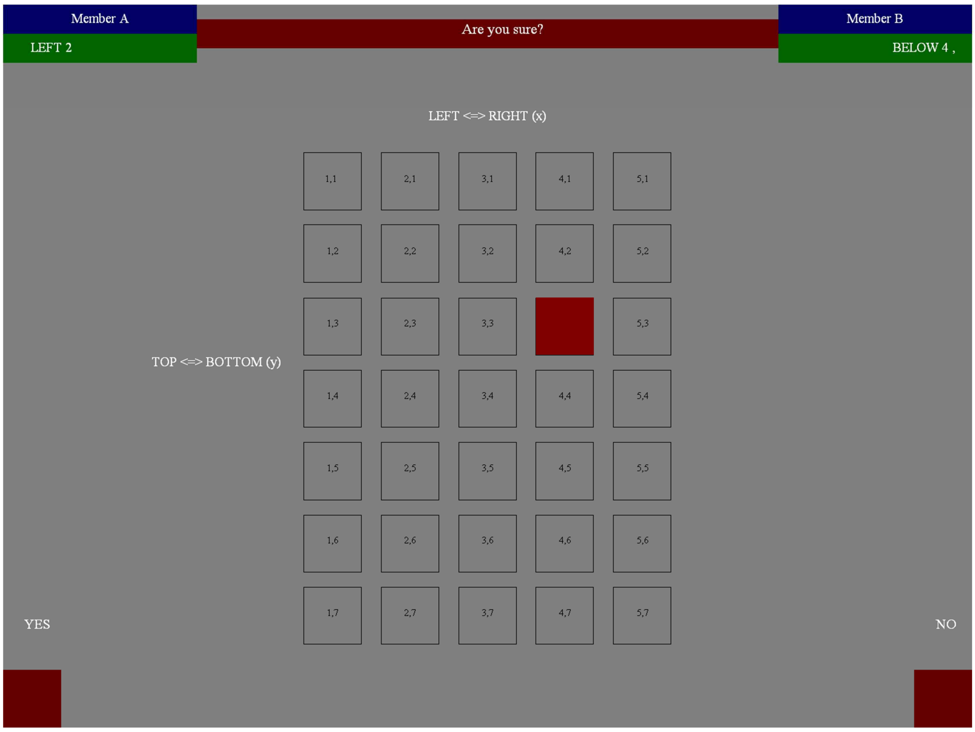
BELOW 4 ,

LEFT <=> RIGHT (x)

TOP <=> BOTTOM (y)

| | | | | |
|---|---|---|---|---|
| 1,1 | 2,1 | 3,1 | 4,1 | 5,1 |
| 1,2 | 2,2 | 3,2 | 4,2 | 5,2 |
| 1,3 | 2,3 | 3,3 | 4,3 | 5,3 |
| 1,4 | 2,4 | 3,4 | 4,4 | 5,4 |
| 1,5 | 2,5 | 3,5 | 4,5 | 5,5 |
| 1,6 | 2,6 | 3,6 | 4,6 | 5,6 |
| 1,7 | 2,7 | 3,7 | 4,7 | 5,7 |

When looking at the box you want to choose, press "space" to make your choice. Then, the box will become red, and you will be asked "Are you sure?" Look at the bottom-left (YES) to confirm, or the bottom-right (NO) to start over again.



Member A

LEFT 2

Are you sure?

Member B

BELOW 4 ,

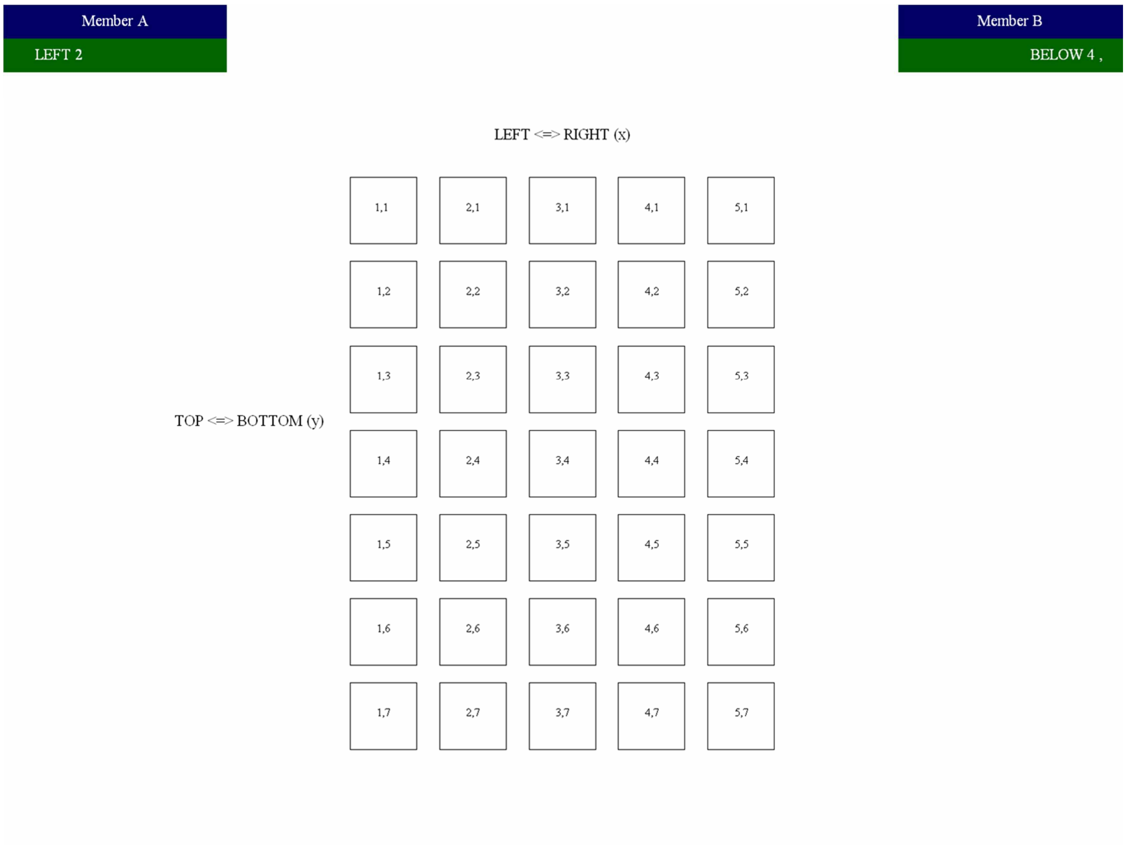
LEFT <=> RIGHT (x)

TOP <=> BOTTOM (y)

| | | | | |
|---|---|---|---|---|
| 1,1 | 2,1 | 3,1 | 4,1 | 5,1 |
| 1,2 | 2,2 | 3,2 | 4,2 | 5,2 |
| 1,3 | 2,3 | 3,3 | 🟥 | 5,3 |
| 1,4 | 2,4 | 3,4 | 4,4 | 5,4 |
| 1,5 | 2,5 | 3,5 | 4,5 | 5,5 |
| 1,6 | 2,6 | 3,6 | 4,6 | 5,6 |
| 1,7 | 2,7 | 3,7 | 4,7 | 5,7 |

YES

NO



| Member A |
|---|
| LEFT 2 |

| Member B |
|---|
| BELOW 4 , |

LEFT <=> RIGHT (x)

TOP <=> BOTTOM (y)

| | | | | |
|---|---|---|---|---|
| 1,1 | 2,1 | 3,1 | 4,1 | 5,1 |
| 1,2 | 2,2 | 3,2 | 4,2 | 5,2 |
| 1,3 | 2,3 | 3,3 | 4,3 | 5,3 |
| 1,4 | 2,4 | 3,4 | 4,4 | 5,4 |
| 1,5 | 2,5 | 3,5 | 4,5 | 5,5 |
| 1,6 | 2,6 | 3,6 | 4,6 | 5,6 |
| 1,7 | 2,7 | 3,7 | 4,7 | 5,7 |

## QUIZ

In order to make sure you understand how your earnings are determined, we will now preform a quiz. Suppose you are member B, and the range of locations are X=**1~5** and Y=**1~7**. Please write down your location choice. Then, the experimenter will tell you the (hypothetical) other's location choice, so you may calculate earnings for each member.

| Member A | |
|---|---|
| LEFT 2 | |

| Member B | |
|---|---|
| | BELOW 4 |

Member B's location choice: X=______________________, Y=____________________

Member A's location choice: X=_____________________, Y=____________________

Member B's target location: X=_____________________, Y=____________________

Member A's target location: X=_____________________, Y=____________________

Member B's earning: 20 – __________- __________=__________

Member A's earning: 20 – __________ - __________=__________

Please tell the experimenter if you have any concerns. **Your payments will be rounded up to the next dollar.** Thank you for your participation!