A Window of Cognition: Eyetracking the Reasoning Process in Spatial Beauty Contest Games

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

National Taiwan University*

Abstract

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. Analyzing both final choices and lookups, we classify subjects into various types based on a level-*k* model. More than a half of the subjects are classified into the same type for both final choices and lookups, supporting the level-*k* model as a complete model of choice and reasoning altogether. When classifications disagree, lookup data could provide additional separation of types.

Keywords: guessing game, initial response, cognitive hierarchy, neuroeconomics
JEL: C91, C72, D87

* August 28, 2009. Department of Economics, National Taiwan University, 21 Hsu-Chow Road, Taipei 100, Taiwan. Chen: r94323016@ntu.edu.tw; Huang: chenying@ntu.edu.tw; Wang: josephw@ntu.edu.tw (corresponding author).

A Window of Cognition: Eyetracking the Reasoning Process in Spatial Beauty Contest Games

I. Introduction

Most economic theories are tested by their predictions about people's choices, since economists usually take the revealed preference approach when interpreting these choices. Moreover, empirical data on choices are relatively easier to obtain, either from the field or from laboratory experiments. Nonetheless, in many cases, the economic theory employed also predicts how people evaluate various situations to arrive at their final choices. For example, in extensive form games, subgame perfect equilibrium is typically solved by backward induction, a procedure that can be carried out step-by-step by players of the game. Since these theories provide clear predictions on people's decision-making process, it is natural to ask whether one could test these predictions using some form of empirical data.

One possible obstacle to this kind of test is the availability of data, since the decision-making process is usually unobservable. However, some experimental studies do attempt to investigate "information search" patterns in games, in order to capture part of the reasoning process. For example, Camerer, Johnson, Rymon, and Sen (1993) and Johnson, Camerer, Sen, and Rymon (2002) employ a mouse-tracking technology called "mouselab" (as a proxy of eyetracking) to study backward induction in three-stage bargaining games by requiring subjects to click on the box to see the pie size in different stages. Costa-Gomes, Crawford and Broseta (2001) and Costa-Gomes and Crawford (2006) employ a similar technology to study payoff lookups in normal form games and information search in two-person guessing (*p*-beauty contest) 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 (2009) employ advanced video-based eyetracking technology to observe the decision-making process of a deceptive sender in sender-receiver games.

All the studies mentioned above take advantage of the additional information acquisition data, gathered by some form of eyetracking technology, to test competing theories of behaviors. One crucial feature in these studies is that 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. This begs the question whether decision-making processes can be observed when there is no hidden information and if such observation is possible, whether economic theory can predict them.

This paper is the first attempt, to our knowledge, that analyzes the reasoning process when there is no explicit hidden information. We design a new set of games, called (two-person) spatial beauty contest games, similar to the *p*-beauty contests (aka "guessing games") studied by Nagel (1995), Ho, Camerer and Weigelt (1998), and Costa-Gomes and Crawford (2006). 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 well known that initial responses in the standard *p*-beauty contest games can be explained by models of heterogeneous levels of rationality such as the level-*k* model (Stahl and Wilson (1995), Nagel (1995), and Costa-Gomes and Crawford (2006)) and the cognitive hierarchy model (Camerer, Ho and Chong (2004)). A key in these models of heterogeneous levels of rationality 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 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.[2] In other words, to figure out which choice to make, a subject has to go through an entire best response hierarchy. The graphical representation of the spatial beauty contest games induces subjects to go through this hierarchy of best responses by counting on the computer screen (instead of reasoning in their minds), leaving footprints that the experimenter can trace.

[1] Two-person guessing games in which players are asymmetric are first proposed by Costa-Gomes and Crawford (2006). Grosskoph and Nagel (2008) also study two-person beauty contest games.
[2] To avoid confusion, the subject is denoted by her while her opponent is denoted by him.

We eyetrack each subject's reasoning process by recording the entire sequence of locations she looks at or fixates at. In other words, we record not only her final choice, but also every 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. By wedding a level-*k* (*Lk*) constrained Markov-switching model (to describe changes between a subject's thinking states of the best response hierarchy) and a logit error model (to describe eye fixations conditional on each thinking state), we construct a model for the lookup data to characterize how subjects think through various best response hierarchies as predicted by the level-*k* model, and classify them into various level-*k* types based on maximum likelihood estimation using individual lookup data. Moreover, we adopt an empirical likelihood ratio test proposed by Vuong (1989) to ensure the estimated type with the largest likelihood is distinctively separated from other competing types. Results show that among the seventeen subjects we tracked, one is level-0 (*L0*), six are level-1 (*L1*), four are level-2 (*L2*), four are level-3 (*L3*), and the remaining two are the equilibrium type (*EQ*) which coincides with level-4 (*L4*) or above in most games of our experiment.

To check whether the estimation on the lookup data is robust, we further classify subjects by using their final choice data only. Following the literature, we adopt a procedure similar to Costa-Gomes and Crawford (2006) to classify subjects using the choice data, and compare the results with our lookup-based classification results. We find that choice-based and lookup-based estimations are pretty consistent, classifying ten of the seventeen subjects as the same type. Furthermore, among the seven subjects where the two classifications differ, for four subjects, results from applying Vuong's test to lookup data indicate that the lookup-based classification types are better than the choice-based ones, while the remaining three subjects have indistinguishable likelihoods (but the lookup-based types all have fewer parameters and may thus be argued to be better classifications should we wish to act conservatively to avoid overfitting). To the contrary, using choice data, for all but one subject, the choice-based classification types are not robust according to a resampling test, having a misclassification rate of at least 18% if one re-samples the choice data and performs the same estimation.

Consistency between choice-based and lookup-based estimations suggests that for a high percentage of subjects, if their final choices are classified as a particular level-*k* type, their lookups follow the best response hierarchy of that level-*k* type as well. This is a strong support to the level-*k* model. It means that the level-*k* model is not just a model on final choices. The best response hierarchy implied by the level-*k* model can also predict the reasoning process of subjects very well. In other words, the level-*k* model is quite complete in that it is a model of choice and reasoning process altogether. On the other hand, when estimation of types based on lookup data differs from that based on choice data, results suggest that lookup data may provide additional separation between competing level-*k* types even when choice data is not enough to distinguish between types. In other words, looking into players' reasoning process gives us valuable information if we are to classify them properly.

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

II. The Experiment

II.A The Spatial Beauty Contest Game

We introduce a two-person guessing game similar to the "*p*-beauty contest," in which players choose locations (instead of numbers) simultaneously on a 2-dimensional plane. Each player has a target location. The target location is defined as a relative location to the other player's choice of location (just like *p* in *p*-beauty contest games) 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." Payoffs are determined by how many steps (the sum of horizontal distance and vertical

distance) a player is away from the target. The larger this number is, the lower her payoff is. Players can only choose locations on a given grid map, though one's target may fall outside. For example, consider the 7x7 grid map in Figure 1. 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 since location (-4, 0) is not on the map, choosing location (-3, 0) is optimal among all 49 feasible choices because location (-3, 0) is only one step from location (-4, 0). To go from any of all other possible 48 locations on the map to location (-4, 0) takes at least two steps. 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 sum of the number of steps is 2.

The equilibrium prediction of this spatial beauty contest game is determined by the targets of both players (as is the case of the *p*-beauty contest games). 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 both choosing zero (hitting the lower bound) in the *p*-beauty contest game where *p* is less than 1. Note that in general the equilibrium needs not be at the corner since targets themselves could be 2-dimensional. For example, when the targets are (4, -2) and (-2, 4) played on a 7x7 grid map, the equilibrium locations for the two players are both two steps away from the corner (labeled as **E** in **bold** and *E* in *italic and underlined* for the two players respectively in Figure 1).

We derive the equilibrium predictions for the general case as follows. Formally, consider a spatial beauty contest game with targets (a_i, b_i) and (a_j, b_j). With some abuse of notation, suppose player *i* chooses location (x_i, y_i) on a map satisfying $x_i \in \{-m, ..., m\}$, $y_i \in \{-n, ..., n\}$ where (0, 0) is the center of the map. For instance, $(x_i, y_i) = (m, n)$ means player *i* chooses the Top-Right corner of the map. The other player *j* also chooses a location (x_j, y_j) on the same map: $x_j \in \{-m, ..., m\}$, $y_j \in \{-n, ..., n\}$. The payoff to player *i* in this game is (the payoff to player *j* is defined similarly)

$$p_i(x_i, y_i; x_j, y_j; a_i, b_i) = \bar{s} - \left(\left| x_i - (x_j + a_i) \right| + \left| y_i - (y_j + b_i) \right| \right) \text{ where } \bar{s} \text{ is a constant.}$$

Notice that payoffs are decreasing in the number of steps a player is away from the target, which in turn depends on the choice of the other player. There is no 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.

To ensure uniqueness, in all our experimental trials, $a_i + a_j \neq 0$. (Otherwise, if $a_i = -a_j = a > 0$, any feasible x_i, x_j satisfying $x_i - x_j = a$ constitutes an equilibrium.[3]) Without loss of generality, we assume that $a_i + a_j < 0$ so that the overall trend is to move leftward.[4] Suppose $a_i < 0$. If $a_i a_j \leq 0$, implying player *i* would like to move leftward but player *j* 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 *i* hit the lower bound while player *j* will best respond to that. The equilibrium choices of both, denoted by (x_i^e, x_j^e), is characterized by $x_i^e = -m$ and $x_j^e = -m + a_j$;[5] if $a_i a_j > 0$, since both players would like to move leftward, they will both hit the lower bound. The equilibrium is characterized by $x_i^e = x_j^e = -m$. To summarize, when $a_i + a_j < 0$, only the player whose target is greater than zero will not hit the lower bound. Therefore,

Proposition 1. *In a spatial beauty contest game with targets* (a_i, b_i) *and* (a_j, b_j) *where both players choose a location* (x, y) *satisfying* $x \in \{-m, ..., m\}, y \in \{-n, ..., n\}$, $a_i, a_j \leq 2m$ *and* $b_i, b_j \leq 2n$, *the equilibrium choices* (x_i^e, y_i^e) *are characterized by:*

[3] Note that this corresponds to the case where $\alpha\beta = 1$ in the two-person guessing (*p*-beauty contest) game in which one subject would like to choose α of her opponent's choice and the opponent would like to choose β of her choice.

[4] Due to symmetry, all other cases are isomorphic to this case.

[5] In all our games, since $a_j \leq 2m$, we do not need to worry about the possibility that x_j^e lies outside the upper bound *m* (i.e., $x_j^e = -m + a_j > m$). In general, if $a_j > 2m$, we have instead $x_j^e = m$.

$$\begin{cases} x_i^e = -m + a_i \cdot I\{a_i > 0\} & \text{if } a_i + a_j < 0 \\ x_i^e = m - a_i \cdot I\{a_i < 0\} & \text{if } a_i + a_j > 0 \end{cases} \text{ and } \begin{cases} y_i^e = -n + b_i \cdot I\{b_i > 0\} & \text{if } b_i + b_j < 0 \\ y_i^e = n - b_i \cdot I\{b_i < 0\} & \text{if } b_i + b_j > 0 \end{cases}$$ where

$I\{\cdot\}$ *is the indicator function.*

In addition to the equilibrium prediction, one may also specify various level-*k* predictions. A natural assumption is that an *L0* player will either choose the center (0,0) or choose any location on the map according to the uniform distribution. An *L1* player *i* with the target (a_i, b_i) would best respond to an *L0* opponent who either exactly chooses the center or chooses the center on average. If an *L0* player chooses the center, to best respond, an *L1* player would choose the location (a_i, b_i) when *m, n* is big enough so that we do not need to worry about falling outside the map.[6] 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 *m, n* is big enough). Repeating this procedure, one can determine the best responses of all higher level-*k* (*Lk*) types. Figure 1 shows the various level-*k* predictions of a 7x7 spatial beauty contest game for two players with targets (4, -2) and (-2, 4) labeled in **bold** and *italic and underlined* respectively.

To account for the possibility that choices may fall outside the map, we define the adjusted choice $R(m, n; (a, b))$. Formally, the adjusted choice is given by $R(m, n; (a, b)) \equiv (\min(m, \max(-m, a)), \min(n, \max(-n, b)))$. In words, if the ideal best response (which hits the target) is location (a, b), the adjusted choice $R(m, n; (a, b))$ gives us the closest feasible location on the map so the choice (x, y) is constrained to lie within the range $x \in \{-m, ..., m\}$, $y \in \{-n, ..., n\}$. 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 Appendix A2, since the grid map is of a finite size, eventually when the level of reasoning for a level-*k* type is large enough, the *Lk* prediction will coincide with the equilibrium. This is similar to the convergence to zero in the *p*-beauty contest. To summarize, we have

[6] Appendix A1 proves that if an *L0* player chooses any location on the grid map according to the uniform distribution, to best respond to such *L0*, an *L1* player would still choose the same location (a_i, b_i). This is true because our payoff structure is point symmetric by (0,0) over the grid map.

Proposition 2. *Consider a spatial beauty contest game with targets* (a_i, b_i) *and* (a_j, b_j) *where both players choose a location* (x, y) *satisfying* $x \in \{-m, ..., m\}, y \in \{-n, ..., n\}$, $a_i, a_j \leq 2m$ *and* $b_i, b_j \leq 2n$. *Denote the choice of a level-k player i by* (x_i^k, y_i^k), *then*

(1) $(x_i^k, y_i^k) = R(m, n; (a_i + x_j^{k-1}, b_i + y_j^{k-1}))$ *for* $k = 1, 2, ...,$ *and* $(x_i^0, y_i^0) = (x_j^0, y_j^0) \equiv (0, 0)$*;*

(2) there exists a smallest positive integer $\bar{k}$ *such that for all* $k \geq \bar{k}$, $(x_i^k, y_i^k) = (x_i^e, y_i^e)$.

PROOF: See Appendix A2.

In Table 1 we list all the 24 spatial beauty contest games used in the experiment, their various level-*k* predictions, equilibrium predictions and the minimum $\bar{k}$*'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 (2n-1) and (2n) (where n=1, 2, …, 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 $\bar{k}$*'s* for our 24 games are almost always 4, but some are 3 (Games 1, 10, 17) or 5 (Games 5, 6, 11, 12). 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 1, different levels make different predictions. In other words, various levels are strongly separated on the map.[7]

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. This indicates that a natural way to use lookups is to take the level-*k* reasoning processes literally in the following sense: For instance, for an *L2* player, 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 has to 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

[7] 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 in Table 1.

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 Proposition 3 and base our econometric analysis on this.[8]

Proposition 3. *For a spatial beauty contest game with targets* (a_i, b_i) *and* (a_j, b_j) *where both players choose a location* (x, y) *satisfying* $x \in \{-m, ..., m\}$, $y \in \{-n, ..., n\}$, $a_i, a_j \leq 2m$ *and* $b_i, b_j \leq 2n$. *Denote the choice of a level-k player i by* (x_i^k, y_i^k). *Assuming one carries out the reasoning process on the map, a level-k player i will look at the following locations in the level-k best response hierarchy* (x^0, y^0), …, (x_i^{k-2}, y_i^{k-2}), (x_j^{k-1}, y_j^{k-1}), (x_i^k, y_i^k).[9]

II.B Experimental Procedure

We conduct 24 spatial beauty contest games (with various targets and map sizes) 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 2A (which we denote as the GRAPH presentation), the other time as two one-dimensional choices chosen separately (See Figure 2B, denoted as the SEPARATE presentation).[10] Half of the subjects are shown the two-dimensional grid maps first, while the rest are shown the maps later. The results of the two presentations are quite similar, so we focus on the results of the two-dimensional presentation.[11]

[8] Note that this prediction is a bold one, and requires many assumptions. One should be surprised if it turns out to be a valid prediction.

[9] The player subscript of (x^0, y^0) is dropped since both *L0* players would choose the center.

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

[11] A comparison between final choices under these two representations is shown in the Appendix (Table S2).

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, Peters and Palmer, 2002). 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 record how subjects think before making their decision (and in fact test whether they think through the best response hierarchy implied by the level-*k* model).

In each round, when subjects use their eyes to fixate at a location, it will light up as red (as Figures 2A and 2B show). If they decide to choose that location, they could hit the space bar. Subjects are then asked to confirm their choices ("Do you confirm?"). They then have a chance to confirm their choice ("YES") or restart the process ("NO").

III. Results

This section analyzes subjects' lookups and final choices. We first report level-*k* classification using final choices. This gives us, for each subject, her choice-based type. Then, we summarize subjects' lookups and demonstrate the plausibility of Proposition 3, namely, subjects do look at and count on the map during their reasoning process. Thirdly, we analyze subjects' lookups with a Markov-switching model to classify them into various level-*k* types. As a part of the estimation, we employ Vuong's test to ensure separation between competing types. This gives us, for each subject, her lookup-based type. Finally, we compare choice-based types with lookup-based ones. We show that for more than a half of the subjects these two classifications coincide. When they differ, Vuong's test largely favors the lookup-based types if we look at the lookup data, while a resampling test casts doubts on the robustness of choice-based types if we look at the choice data. This demonstrates how lookup data can help us perform a sharper empirical classification of level-*k* types.

III.A Level-*k* Classification Based on Final Choices

We classify subjects into various (level-*k*) behavioral types based on their final choices, using the econometric analysis developed by Costa-Gomes and Crawford (2006). To evaluate the robustness of the classification, a resampling procedure is employed.

1. Econometric Analysis

We follow a procedure similar to Costa-Gomes and Crawford (2006) and perform a maximum likelihood estimation to classify each individual subject into a particular behavioral level-*k* type using the following logit error structure.[12] Let all possible level-*k* types be $k = 1,...,K$ and each subject goes through round $n = 1,...,N$. For a given round n, according to Proposition 2, a level-*k* subject's theoretic final choice is denoted as $c_n^k \in G_n$, where G_n is a finite countable choice set specified for round n. The choice set G_n depends on the map size of the game in that particular round, and $|G_n| = M_n$ is the number of elements in G_n.[13] Because of the logit error, a level-*k* subject may not choose c_n^k in round *n* with certainty. Instead, the logit error predicts a probabilistic choice $r_n(c_n^k) \in G_n$ which we will describe soon. Let $g_{n_1}, g_{n_2}, ..., g_{n_{M_n}}$ be typical elements of G_n. Define the distance $\left\| g_{n_m} - g_{n_{m'}} \right\|$ as the "steps" on the map (the sum of vertical and horizontal distance) between g_{n_m} and $g_{n_{m'}}$. Then, if a subject chooses a location g, her payoff (had c_n^k been her target) in this round is $S\left(\left\| g - c_n^k \right\|\right) = \overline{s} - \left\| g - c_n^k \right\|$ where $\overline{s}$ is a fixed initial payoff (endowment). We consider a logit error model and construct the choice density function d with precision λ_k as

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

[12] 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.

[13] For instance, suppose in round *n*, the grid map is as shown in Figure 1, then the choice set G_n consists of all 49 locations on the map.

In words, this means that the probability a level-*k* subject chooses a particular location $r_n(c_n^k)$ depends on how far this location is away from c_n^k, which is what a level-*k* player would choose according to the level-*k* model. Locations farther away from c_n^k are less likely. When $\lambda_k \to 0$, the subject randomly chooses from the choice set G_n. As $\lambda_k \to \infty$, the choice of the subject approaches to the level-*k* choice c_n^k. The log likelihood over all rounds can then be expressed as

$$\ln \prod_{n=1}^{N} d^k \left(r_n(c_n^k) \right). \tag{1}$$

For each *k*, we estimate the precision parameter λ_k by fitting the data with the logit error model to maximize empirical likelihood. Then we choose the *k* which maximizes empirical likelihood and classify the subject into this type.

We consider all the level-*k* types separable in our games: *L0, L1, L2, L3,* and *EQ*. Results shown in column (A) of Table 2 indicate, 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 results of the standard *p*-beauty contest games using Caltech subjects (but higher than normal subjects).[14] Moreover, to incorporate all empirically possible behavioral types, we follow Costa-Gomes and Crawford (2006) and include 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 their choices are better explained by each other's than by the pre-specified level-*k* types. Denoting pseudo-*i* the pseudotype constructed form subject-*i*, the results are reported in Appendix (Table S1). 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 Crawford (2006), they 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. Table 2 lists the classification with and without pseudotypes in columns (B) and (A)

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

respectively. The distribution of level-*k* types does not change much even if we include pseudotypes, having two *L0*, three *L1*, four *L2*, three *L3*, and three *EQ* (see column (B) of Table 2). The average of thinking steps is 2.13, nearly identical to that without pseudotypes.[15] This suggests that in our games, the level-*k* classification is quite robust to empirically omitted types.

2. Resampling Test for Robustness of Types

The above econometric model based on maximum likelihood estimation may not have enough power to distinguish between various types. For example, reading from Table S1, subject 14's 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 resampling procedure to attempt to deal with the issue of robustness.

Imagine that from the maximum likelihood estimation, a subject is classified as a particular level-*k* type. We evaluate the robustness of this classification with the following resampling test. Since our econometric estimation assumes each subject's 24 rounds of observations are independent, it is natural to maintain this assumption when resampling. Hence, we resample the data by randomly drawing one round out of the 24 rounds observed for each subject. By drawing (with replacement) 24 times, we obtain a new (resampled) dataset for this particular subject. Then, we estimate the subject's type with this resampled data. Since the resampled data is expected to resemble the empirical distribution, we should expect the maximum likelihood procedure gives us the same level-*k* type for sufficiently many resampled datasets. If the type estimated from a resampled dataset is not the same level-*k* type, we view this as a "misclassification," and count it against this particular classification *k*. By resampling 1000 times and calculating the total misclassification rate, we can measure the robustness of the original

[15] 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.

classification (against resampling error). This resampling test 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 resampling test are listed in Table 3. For each subject, we report the number of times that a subject is classified into *L0*, *L1*, *L2*, *L3* or *EQ* in the 1000 resampled datasets. The misclassification rate (percentage of times one classifies the subject as a type different from her original type having the largest likelihood using final choices) 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 resampling test. Subject 14 was 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 classification is not as good as one would hope, since only 8 subjects passed this resampling test with misclassification rate lower than 5% and could thus be unambiguously classified into a certain type. This suggests that choice data might not be enough to perform sharp classification. We turn to consider how the lookup data could help us further.

III. B Lookup Summary Statistics

Aggregate data regarding empirical lookups for all 24 Spatial Beauty Contest games are presented in Figures 3A and 3B: games with 1-dimensional target are presented in Figure 3A, and those with 2-dimensional targets in Figure 3B. 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 in each location, so bigger circles indicate longer time spent. The level-*k* choice predictions are labeled as O (*L0*), L1 (*L1*), L2 (*L2*), L3 (*L3*), E (*EQ*) for each game.

If Proposition 3 were true, the empirical lookups would concentrate on locations predicted by the level-*k* best response hierarchy. In fact, many big circles in Figures 3A

and 3B do coincide with locations of the level-*k* best response hierarchy. We attempt to quantify this concentration of attention. First, for every level-*k* type, we define the *Hit* area which is the minimal convex set enveloping the locations predicted by this level-*k* type's best response hierarchy. For instance, for an *L2* subject *i* (with opponent *j*), the best response hierarchy consists of (x^0, y^0), (x_j^1, y_j^1), (x_i^2, y_i^2). Thus we can construct a minimal convex set enveloping these three locations. We then take the union of the *Hit* areas of all level-*k* types and see if subjects' lookups are indeed within the union. Figure 4 shows an example of *Hit* areas for various level-*k* types in a 7x7 spatial beauty contest game with target (4, -2) and the opponent's target (-2, 4) (Game 16).

Figure 5 shows the empirical percentage of time spent on the union of *Hit* areas (aka "hit time"). Across the 24 games, average hit time is 0.62, ranging from 0.81 (in Game 9), to 0.36 (in Game 21). Since on average more than 60% of the lookup time is spent on the union of *Hit* areas, empirical lookups are indeed concentrated on locations predicted by the best response hierarchies of various level-*k* types.

However, hit time depends on the size of the area. If subjects only look at locations on the level-*k* best response hierarchy, the empirical hit time would be 1. However, 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 (aka "hit area size"). To correct for this hit area size bias, we calculate Selten's (1991) (linear) "difference measure of predicted success," i.e. the difference between empirical hit time and hit area size, and report it in Figure 6. These measures are all positive (except for Game 21), reflecting large hit area sizes alone cannot account for the high empirical hit time. This suggests that subjects indeed spent a disproportionally long time in the union of *Hit* areas.[16] With this aggregate result in mind, we now consider whether individual lookup data can be used to classify subjects into various level-*k* types and helps reduce the possible misclassifications based on final choices alone.

[16] In fact, sometimes subjects have hit time nearly 1. For example, Figure 7 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 *L2*, which is exactly her level-*k* type (based on lookups) according to the last column of Table 4.

III.C A Markov-Switching Model for Level-*k* Reasoning

According to Proposition 3, a level-*k* type subject *i* goes through the best response hierarchy during the reasoning process, and fixates at locations (x^0, y^0), …, (x_i^{k-2}, y_i^{k-2}), (x_j^{k-1}, y_j^{k-1}), (x_i^k, y_i^k). As subjects reason through the hierarchy from (x^0, y^0) to (x_i^k, y_i^k), we may consider lookups as a serially correlated time-series. We define some state variables to help us characterize the stage of the best response hierarchy the subject *i* is at. Define ***Lk*** as the state which indicates that she is (reasoning as) an *Lk* and hence her fixation is concentrated on the location (x_i^k, y_i^k). In general, we use the apostrophe to denote it is about the opponent. Hence, ***L(k-1)'*** is defined as the state which indicates that she is reasoning that her opponent *j* is an *L(k-1)* and hence her fixation is concentrated on the location (x_j^{k-1}, y_j^{k-1}). Lower states ***L(k-2)***, ***L(k-3)'***,..., etc. are defined similarly. Then, states corresponding to subject *i*'s level-*k* best response hierarchy of Proposition 3 can be expressed as "***L0***, …, ***L(k-2)***, ***L(k-1)'***, ***Lk***." For instance, the reasoning process of an *L2* subject *i* consists of three stages: First, she would be in state ***L0*** and fixate at (x_i^0, y_i^0) since she believes her opponent is *L1*, who believes she is *L0*. Then, she would be in state ***L1'*** and fixate at (x_j^1, y_j^1), thinking through her opponent's choice as an *L1*. Finally, she would be in state ***L2*** and best respond to the belief that her opponent is an *L1* by making her choice fixating at (x_i^2, y_i^2). These three states as ***L0*** (fixating at the location of (x_i^0, y_i^0)), ***L1'*** (fixating at the location of (x_j^1, y_j^1)), ***L2*** (fixating at the location of (x_i^2, y_i^2)) are expected to be passed through during the reasoning process of an *L2* subject. To account for the transitions of states, we employ a Markov switching model first used in macroeconomics by Hamilton (1989) to describe business cycles and characterize the transition of states by a Markov transition matrix.

We do not require a level-*k* subject to "strictly" obey a monotonic order of level-*k* thinking. In other words, they are not required to necessarily go from a lower state to a higher state. i.e., always moving upwards through the sequence ***L0***, …, ***L(k-2)***, ***L(k-1)'***, ***Lk***. Instead, we allow subjects to 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* type best responds to a level-(*k-1*) opponent, it

is difficult to imagine a subject jumping from the reasoning state of say ***L(k-2)*** to that of ***Lk*** without first going through the reasoning state of ***L(k-1)'***. Thus, the transition probability should be restricted to zero for all transitions that involve a jump in states.

To perform the classification using lookups on a subject-by–subject basis, we assume that each subject is the same level-*k* type in all rounds and the error structure under each state follows the same logit distribution. We perform a maximum likelihood estimation to obtain the transition probabilities and the logit error parameter λ, and classify subjects into various level-*k* types based on lookups.

1. Econometric Analysis

Let all possible level-*k* types be $k = 1,...,K$ and there are N rounds of games in which each round is indexed by $n = 1,...,N$. In round *n*, Let G_n be the (finite) choice set which depends on the size of the map in that round, and $|G_n| = M_n$ is the number of elements in G_n. Elements of G_n are denoted by $g_{n_1}, g_{n_2}, ..., g_{n_{M_n}}$. Since every element in G_n is a location on the map, each can be represented by a pair of coordinates. Recall that payoff is decreasing in the distance (or steps). Suppose $g_{n_m} = (x_{g_{n_m}}, y_{g_{n_m}})$ and $g_{n_{m'}} = (x_{g_{n_{m'}}}, y_{g_{n_{m'}}})$, as before, we define the distance as the sum of vertical and horizontal distance, $\| g_{n_m} - g_{n_{m'}} \| = \left| x_{g_{n_m}} - x_{g_{n_{m'}}} \right| + \left| y_{g_{n_m}} - y_{g_{n_{m'}}} \right|$.

For a level-k subject, we consider the *state* space Ω_k consisting of all stages in the best response hierarchy with (*k+1*) states {***L0***, …,***L(k-1)'***, ***Lk***}. We then define a state-to-lookup mapping $l_n : \Omega_k \to G_n$ for round *n* which assigns each state to a corresponding lookup location in G_n. For instance, if a level-2 player in round *n* is in state ***L0*** at a point of time, the l_n mapping would give us the location which a level-0 player would choose since at this particular point of time, when an *L2* is thinking about what an *L1* thinks an *L0* would choose, she would fixate at the location that an *L0* would choose. Similarly, if a level-2 player in round *n* is in state ***L1'*** (***L2***), then the l_n mapping would give us the location which a level-1 opponent (a level-2 subject) would choose.

In each game, we observe a sequence of lookups. We would like to infer, for each lookup in the sequence, which state a subject is in when her lookup is on that particular lookup location. However, the current state of a subject crucially depends on previous states since we assume a level-*k* subject goes through stages of the best response hierarchy (but allowing her to go back to double check). Therefore, we assume this transition is Markov, depending on the immediate previous state only, and estimate a constrained Markov transition matrix. We constrain the Markov model since the level-*k* model requires one to move up the hierarchy one step at a time (but has no restriction moving down). In other words, if we list the previous state in the row and the current state in the column and states are ordered from lower to higher, elements of the transition matrix have to be zero if they are more than one element above the diagonal (i.e. their column index is greater than the row index plus one).

In addition, we estimate a logit error model to describe the relationship between states and lookups. Suppose a level-2 player is inferred to be in state ***L1'***, then by the mapping l_n, her lookup should fall exactly on the location l_n(***L1'***). If her lookup is not on that location, we interpret this as an error. We assume a logit error structure for such errors so that looking at locations farther away from l_n (***L1'***) is less likely, and to estimate a precision parameter for this error structure.

To summarize, we estimate a state transition matrix and a precision parameter. Thus for any 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 from the state to the lookup gives us the lookup location corresponding to any state when there is no error. Coupled with the precision parameter, we can calculate the probability distribution of various errors and therefore the distribution of predicted lookup locations when errors are permitted. Using the state transition matrix and the precision parameter, we can calculate how well we are able to explain any observed sequence of lookups. The final step is to select the *k* in various level-*k* models that best explains the observed sequence of lookups for each subject.

Formally, the lookup sequence in round *n* is a time series over $t = 1,...,T_n$. Because of the logit error, a level-*k* subject may not look at a location with certainty. Therefore,

for time *t* (i.e. 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 $\mathcal{R}_n^t \equiv \{ r_n^1, \ldots, r_n^{t-1}, r_n^t \}$.

Suppose the subject is a particular level-*k*. Let $\mathbf{Z}^t$ be the random variable representing the state, drawn from the state space $\Omega_k = \{\textbf{\textit{L0}}, \ldots, \textbf{\textit{L(k-1)'}}, \textbf{\textit{Lk}}\}$, and its realization be z^t at time *t*. Denote the state history up to time *t* by $\mathcal{Z}^t \equiv \{z^1, ..., z^{t-1}, z^t\}$.[17]

Since lookups may be serially correlated, we model this by estimating a constrained Markov stationary transition matrix of states. Denote the transition probability from state $\mathbf{Z}^{t-1} = z^{t-1}$ to $\mathbf{Z}^t = z^t$ by

$$\Pr(\mathbf{Z}^t = z^t \mid \mathbf{Z}^{t-1} = z^{t-1}) = \pi_{z^{t-1} \to z^t} \tag{1}$$

Thus, the state transition matrix θ_k is

$$\theta_k = \begin{pmatrix} \pi_{0\to 0} & \cdots & \pi_{0\to k} \\ \vdots & \ddots & \vdots \\ \pi_{k\to 0} & \cdots & \pi_{k\to k} \end{pmatrix} = \begin{pmatrix} \pi_{0\to 0} & \pi_{0\to 1} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \vdots & & & \ddots & \pi_{k-1\to k} \\ \pi_{k\to 0} & \cdots & \cdots & \cdots & \pi_{k\to k} \end{pmatrix}.$$

where $\pi_{i\to j} = 0$ for all *j>i+1* since we do not allow for jumps.

Conditional on $\mathbf{Z}^t = z^t$, 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 $l_n(z^t)$), independent of lookup history $\mathcal{R}_n^t$. In other words,

$$\Pr(\mathbf{R}_n^t = r_n^t \mid \mathbf{Z}^t = z^t, \mathcal{R}_n^{t-1}) = \frac{\exp\left(-\lambda_k \left\| r_n^t - l_n(z^t) \right\|\right)}{\sum_{g \in G_n} \exp\left(-\lambda_k \left\| g - l_n(z^t) \right\|\right)}. \tag{2}$$

where $\lambda_k \in [0, \infty)$ is the precision. If $\lambda_k = 0$, subjects randomly look at locations in G_n. As $\lambda_k \to \infty$, subjects' lookups concentrate on the lookup location $l_n(z^t)$ predicted by the

[17] 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.

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

$$\Pr(\mathbf{R}_n^t = r_n^t \mid \mathcal{R}_n^{t-1}) = \sum_{z^t \in \Omega_k} \Pr(\mathbf{Z}^t = z^t \mid \mathcal{R}_n^{t-1}) \cdot \Pr(\mathbf{R}_n^t = r_n^t \mid \mathbf{Z}^t = z^t, \mathcal{R}_n^{t-1}) \tag{3}$$

where

$$\begin{aligned} &\Pr(\mathbf{Z}^t = z^t \mid \mathcal{R}_n^{t-1}) \\ &= \sum_{z^{t-1} \in \Omega_k} \Pr(\mathbf{Z}^{t-1} = z^{t-1} \mid \mathcal{R}_n^{t-1}) \cdot \Pr(\mathbf{Z}^t = z^t \mid \mathbf{Z}^{t-1} = z^{t-1}, \mathcal{R}_n^{t-1}) \\ &= \sum_{z^{t-1} \in \Omega_k} \Pr(\mathbf{Z}^{t-1} = z^{t-1} \mid \mathcal{R}_n^{t-1}) \cdot \pi_{z^{t-1} \to z^t}. \end{aligned} \tag{4}$$

The last equality follows since according to the Markov property, $\mathbf{Z}^{t-1} = z^{t-1}$ is sufficient to predict $\mathbf{Z}^t = z^t$. Note that (4) depends on the Markov transition matrix and the second term on the right hand side of (3) depends on the logit error. Hence, for a given round *n*, coupled with the initial distribution of states, the joint density of a level-*k* subject's empirical lookups, denoted by

$$\begin{aligned} f_n^k(r_n^1, ..., r_n^{T_n-1}, r_n^{T_n}) &\equiv \Pr(r_n^1, ..., r_n^{T_n-1}, r_n^{T_n}) \\ &= \Pr(r_n^1) \Pr(r_n^2 \mid r_n^1) \Pr(r_n^3 \mid r_n^1, r_n^2) ... \Pr(r_n^{T_n} \mid r_n^1, r_n^2, ..., r_n^{T_n-1}), \end{aligned} \tag{5}$$

can be derived. Assuming a level-*k* subject's lookups independently follow the same Markov process across all rounds, the log likelihood over *N* rounds is

$$L(\lambda_k, \theta_k) = \ln\left[\prod_{n=1}^{N} f_n^k(r_n^1, ..., r_n^{T_n-1}, r_n^{T_n})\right]. \tag{6}$$

To estimate λ_k and the Markov transition matrix θ_k, by (6), we need to start with an initial distribution of states. Since level-*k* reasoning starts from the lowest state (here ***L0***), we assume this initial distribution degenerates to a mass point at the lowest state

corresponding to ***L0***.[18] With this assumption, we estimate λ_k and the constrained Markov transition matrix θ_k using maximum likelihood estimation for each *k*, and classify subjects into the particular level-*k* type which has the largest likelihood.

2. Model Selection using the Vuong's Test

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

---

[18] Formally, we start with the assumption that $\Pr(\mathbf{Z}^0 = z^0) = 1$ when the initial state z^0 is ***L0*** and zero otherwise. Then we derive the following step by step. First, $\Pr(z^1) = \sum_{z^0 \in \Omega_k} \left[\Pr(z^0) \Pr(z^1 \mid z^0) \right]$, where $\Pr(z^0)$ is given by the initial distribution of states and $\Pr(z^1 \mid z^0)$ is given by the Markov transition matrix. Second, $\Pr(r_n^1) = \sum_{z^1 \in \Omega_k} \left[\Pr(z^1) \Pr(r_n^1 \mid z^1) \right]$, where $\Pr(z^1)$ is given by the first step and $\Pr(r_n^1 \mid z^1)$ is given by the logit error. Third, we update the state by the current lookup or $\Pr(z^1 \mid r_n^1) = \left[\Pr(z^1) \Pr(r_n^1 \mid z^1) \right] \Big/ \Pr(r_n^1)$, where terms in the numerator and the denominator are both derived in the second step. Fourth, we derive the next state from the current lookup or $\Pr(z^2 \mid r_n^1) = \sum_{z^1 \in \Omega_k} \left[\Pr(z^1 \mid r_n^1) \Pr(z^2 \mid r_n^1, z^1) \right] = \sum_{z^1 \in \Omega_k} \left[\Pr(z^1 \mid r_n^1) \Pr(z^2 \mid z^1) \right]$, where the second equality follows because by Markov, the transition to the next step only depends on the current state. Moreover, $\Pr(z^1 \mid r_n^1)$ is derived in the third step and $\Pr(z^2 \mid z^1)$ is given by the Markov transition matrix. Fifth, we derive the next lookup from the current lookup or $\Pr(r_n^2 \mid r_n^1) = \sum_{z^2 \in \Omega_k} \left[\Pr(z^2 \mid r_n^1) \Pr(r_n^2 \mid r_n^1, z^2) \right]$, where $\Pr(z^2 \mid r_n^1)$ is given by the fourth step and $\Pr(r_n^2 \mid r_n^1, z^2) = \Pr(r_n^2 \mid z^2)$ is given by the logit error. Sixth, as in the third step, we update the state by the lookups up to now or $\Pr(z^2 \mid r_n^1, r_n^2) = \left[\Pr(z^2 \mid r_n^1) \Pr(r_n^2 \mid r_n^1, z^2) \right] \Big/ \Pr(r_n^2 \mid r_n^1)$, where terms in the numerator and the denominator are both derived in the fifth step. Seventh, as in the fourth step, we derive the next state from the lookups up to now or $\Pr(z^3 \mid r_n^1, r_n^2) = \sum_{z^2 \in \Omega_k} \left[\Pr(z^2 \mid r_n^1, r_n^2) \Pr(z^3 \mid r_n^1, r_n^2, z^2) \right] = \sum_{z^2 \in \Omega_k} \left[\Pr(z^2 \mid r_n^1, r_n^2) \Pr(z^3 \mid z^2) \right]$, where $\Pr(z^2 \mid r_n^1, r_n^2)$ is derived in the sixth step and $\Pr(z^3 \mid z^2)$ is given by the Markov transition matrix. Eighth, as in the fifth step, we derive the next lookup from the lookups up to now or $\Pr(r_n^3 \mid r_n^1, r_n^2) = \sum_{z^3 \in \Omega_k} \left[\Pr(z^3 \mid r_n^1, r_n^2) \Pr(r_n^3 \mid r_n^1, r_n^2, z^3) \right]$, where $\Pr(z^3 \mid r_n^1, r_n^2)$ is given by the seventh step and $\Pr(r_n^3 \mid r_n^1, r_n^2, z^3) = \Pr(r_n^3 \mid z^3)$ is given by the logit error. Continuing in this fashion and multiplying altogether the second step, the fifth step, the eighth step, and so on, we will be able to 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 $\Pr(\mathbf{Z}^0 = z^0) = 1/(k+1)$ for all $z^0 \in \Omega_k$. 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 $\Pr(\mathbf{Z}^0 = z^0) = 1$ when the initial state z^0 is ***L0*** and zero otherwise.

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.[19] For example, a level-2 subject has 3 states (***L0***, ***L1'***, and ***L2***) and five (Markov) parameters, but a level-1 subject has only 2 states (***L0'*** and ***L1***) 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 (and predicts more lookup locations).[20]

However, usual tests for model restrictions may not apply, since the parameters involved in different level-*k* types could be non-nested. For instance, the state space of a level-1 type, (***L0'***, ***L1***), is nested in the state space of a level-3 type, (***L0'***, ***L1***, ***L2'***, ***L3***), but is not nested in the state space of a level-2 type, (***L0***, ***L1'***, ***L2***).

In order to evaluate the classification, we follow Vuong's test for overlapping models (1989). 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 $(\lambda_{k^a}, \theta_{k^a})$. We want 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. In order to do so, 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.[21] 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

[19] 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 $(k+1)\times(k+1)-(k+1)-[k(k-1)]/2=[k(k+3)]/2$ parameters.

[20] Overfitting is an issue pertaining to lookup data since higher level-*k* types have more parameters. This is not an issue in choice data since every type has only one logit error parameter and makes only one final choice anyway.

[21] Since overfitting may be a problem, when both models are equally good, we consider the number of parameters in each model, and conservatively select the model with the lower number of parameters to avoid the possibility that we may select a higher type simply because it has more parameters.

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

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

To perform the Vuong's test, we construct the log-likelihood ratio round by round and define

$$m_n = lr_n(\lambda_{k^*}, \theta_{k^*}) - lr_n(\lambda_{k^a}, \theta_{k^a}) \text{ for round } n=1,\ldots, N.$$

Let $\bar{m} = \frac{1}{N}\sum_{n=1}^{N} m_n$. Vuong (1989) proposes a sequential procedure (p.321) for overlapping models. Its general results 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, the Vuong's test is performed by 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.)

Recall that in our case Lk^* is the type with the largest likelihood based on lookups. Let the alternative type Lk^a be the type having the next largest likelihood among all

lower types.[22] If according to Vuong's test, (V2) applies so that Lk^* is a better model than Lk^a, we can be assured that the maximum likelihood criterion does not pick up the reported type (instead of the second largest type) by mere chance. Thus, we conclude that the lookup-based type is Lk^*. If instead we find that according to Vuong's test, (V1) applies so that Lk^* and Lk^a are equally good, then we conservatively classify the subject as the second largest lower type Lk^a. (V3) does not apply since $V > 0$ by construct.

Table 4 shows the results of the maximum likelihood estimation and Vuong's test for each subject. For each subject, we list her Lk^* type, her Lk^a type and her lookup-based type according to the Vuong' test in the last column. Six of the seventeen subjects (subjects 1, 5, 6, 8, 11, 13) pass the 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 (C) of Table 2. After employing the Vuong's test, the type distribution for (*L0*, *L1*, *L2*, *L3*, *EQ*) is (1, 6, 4, 4, 2).[23] 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. The average number of thinking steps is 2.00.[24]

[22] Recall that the reason why we look at the 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.

[23] If we ignore the two pseudo-17 subjects (subjects 3 and 17, both classified as *L1* based on lookups) since their choices suggest that they may not behave according to the level-*k* theory, then the type distribution for (*L0*, *L1*, *L2*, *L3*, *EQ*) is (1, 4, 4, 4, 2).

[24] Two points are worth noting here. 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, we also perform the Hansen's test (Hansen, 1992). Results are reported in columns of Table S3, and are nearly identical to those based on the Vuong's test. The only potential difference is subject 6 having a marginally significant *p*-value of 0.053 (while in Vuong's test the test statistic is V=2.40>1.96, significant at the 5% level). In other words, even when we switch to use Hansen test when Lk^* and Lk^a are strictly nested, the result is almost the same. Secondly, note that we only perform the 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 type that has the next largest likelihood. It is possible that the lower type with the next largest likelihood is still not different from the even lower type with the even next largest likelihood (and so on). Hence, one might wonder whether we should stop here. Nonetheless, even if we employ an iterative Vuong's test and classify subjects as the type that is, for the first time, significantly different from a lower type of which the likelihood is immediate lower, we can 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 Table S3.

There are several possible explanations to why we observe higher than typical type distributions reported in previous studies. First of all, as stated before, Caltech subjects are reported to have more steps of thinking than usual subjects. Moreover, the spatial beauty contest game is intuitive and does not require mathematical multiplication (as compared with say, the standard *p*-beauty contest game). Hence, this may explain why subjects could perform more steps of reasoning in this easier task.[25]

However, one might wonder whether the results reported in Table 4 is due to a misspecification of possible types. After all, many assumptions are required for Proposition 3 to hold. Unfortunately, we cannot perform a pseudotype test as in Costa-Gomes and Crawford (2006) since the length of the lookup sequence differs across subjects. However, we can compare the classification based on lookups with those using final choices alone, and see if types are aligned between the two classifications. We turn to this now.

III.D Final Choices vs. Lookups

In Table 2, the choice-based and lookup-based classification results look similar, though the choice results indicate slightly more steps of reasoning (2.12 for choice-based types without pseudotypes instead of 2.00 for look-up based types with Vuong's test). This suggests that the lookup-based estimation (and the underlying Proposition 3) is not completely out of the 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. For each subject, using the lookup data, we consider her lookup-based type (denoted by Lk^l, as reported in the last column of Table 4) and her choice-based type (denoted by Lk^c, as reported in the second column of Table 3). We perform the Vuong's test between these two types (using the lookup data) and report the *V* statistics for Vuong's test in the second to last column of Table 5. For the ease of comparison, in Table 5 we also reproduce the misclassification rate of choice-based types in the last column for each subject (reported originally in the

[25] For example, Chou et al. (2009) show that a graphical presentation of the standard *p*-beauty contest game yields results closer to equilibrium.

last column of Table 3). Overall, using the lookup data, for ten of the seventeen subjects, their lookup-based types and the choice-based types are the same. Such alignment in classification results would be surprising if one thought Proposition 3 was too strong a claim. Nevertheless, given that for more than half of the subjects, both classifications are the same, it is hard not to accept the conclusion that Proposition 3 (and its underlying assumptions) does have some prediction power. Moreover, for these ten subjects, all but three of them have (choice-based) misclassification rates lower than 0.05, suggesting that their classifications are truly sharp.[26]

On the other hand, among the seven subjects whose lookup-based classification and choice-based classification differ, using lookup data, results of Vuong's tests suggest that the lookup-based classifications are significantly better models than the choice-based models for four subjects (see the column labeled as Vuong's statistic *V* in the bottom panel of Table 5). This suggests that based on lookup data, the lookup-based types indeed can separate well from the choice-based types for these four subjects. For the remaining three subjects (whose two classifications fit equally well), the lookup-based classifications all have fewer parameters than choice-based models. Hence if we worry about overfitting using the lookup data, since according to Vuong's test, the lookup-based models and the choice-based models are equally good but the lookup-based ones have fewer parameters, this, in a conservative sense, makes the lookup-based models better models to explain the lookup data. Moreover, among these seven subjects (except subject 8), for six subjects, if we use the choice data, their choice-based type all have 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. A closer look at Table 3 would in fact reveal that for these six subjects, they are actually classified into lookup-based type not so infrequently when we resample their choices. Their lookup-based types are almost always the second most frequent type they

[26] One of the three subjects (subject 17) is a pseudotype. 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. In contrast, except for subject 8, all other six subjects whose lookup-based types are different from their choice-based types have misclassification rates at least 0.184. This suggests that when the lookup-based types and the choice-based types are the same, the classification is quite sharp. On the other hand, when they differ, the classification based on choice is not that sharp.

are classified into in the resampling test.[27] This provides another piece of evidence that the lookup-based types can explain part of their choices, suggesting that their lookup-based types might not be a bad candidate if we are to classify them properly.[28]

Altogether, when lookup-based types differ from choice-based types, using lookup data, lookup-based types are either better models than choice-based types or they have fewer parameters than choice-based types. On the other hand, using choice data, choice-based type typically have misclassification rate not so low. Moreover, the lookup-based types are typically the second most frequent types they are classified into if we resample their choices. From these we probably can conclude lookup data do have some truth in classifying subjects properly and may help us separate types.

To summarize, these results show that lookup data can help us confirm classification results based on choices alone and even provide better classification results. In particular, without the lookup data, we could have classified subjects into certain types based on insignificantly larger likelihoods.

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 Proposition 3. Results in Table 5 show that the level-*k* model does indeed hold up under this test. One ought to keep in mind that explaining the reasoning process is a hard one, if not harder than explaining choices. The result that in our dataset, for more than a half of subjects, their lookup-based types are aligned with their choice-based types could 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.

---

[27] For instance, for subject 6, her lookup-based type is *EQ* while her choice-based type is *L2*. In 1000 times of resampling of choices, she is classified to *EQ* 228 times. The only exception is subject 14, whose choice-based type is *EQ*, but is reclassified as *L3* 228 times and as *L1* (her lookup-based type) 185 times.

[28] If we ignore the two pseudo-17 subjects (subjects 3 and 17) since their choices suggest that they may not behave according to the level-*k* theory, the results are even stronger. Among the remaining fifteen subjects, for nine subjects, their lookup-based types and the choice-based types are the same. Among these nine subjects, except for subjects 2 and 4, the misclassification rates of their choice-based types are all lower than 0.05 For subjects 2 and 4, the misclassification rates are 0.076 and 0.110, both at the margin. For the six subjects whose lookup-based types differ from their choice-based types, for four of them, using the lookup data, their lookup-based types are better models than their choice-based types according to Vuong's test. For two of them, using the lookup data, their lookup-based types and their choice-based types are equally good according to Vuong's test. But their lookup-based types have fewer parameters.

Finally, one might wonder if it is the case that subjects in fact do perform lookups that resemble higher levels of strategic thinking, but somehow "downgrade" their choices to a lower level, possibly realizing that even if they could perform higher levels of thinking, their opponents may not. However, among the seven subjects whose lookup-based classification disagrees with the choice-based one, three of them have higher lookup-based types (subjects 6, 9, 11), while the remaining four (subjects 3, 8, 14, 15) have higher choice-based types. So, this explanation could at most account for only half of the disagreements in our data.

IV. Conclusion

We introduce a new spatial beauty contest game, and provide theoretical predictions based on the equilibrium and the level-*k* theory. The theoretical predictions of the level-*k* theory yield a plausible conjecture on the decision-making process when people actually play the game. 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 could be classified into various level-*k* types, 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, a resampling test shows that we might misclassify subjects into their choice-based types due to insignificantly larger likelihoods. On the other hand, Vuong's test on lookups shows that lookup-based types are either better models than choice-based types or have fewer parameters than choice-based types. This suggests that lookups may give us a stronger separation of types.

Comparing the distribution of level-*k* types based on final choices with that based on lookups, we find that some subjects have higher level-*k* types using lookup data. This could be due to imprecise choice-based classification. Another possibility is that subjects may perform lookups that resemble high levels of strategic thinking, but decide to "downgrade" their choices to a lower level, possibly realizing that their opponents may not perform equally high levels of thinking. Our eyetracking data show that the latter explanation can at most account for a half of the difference between the two distributions.

However, this explanation is of special interest for future research because it is inconsistent with the cognitive hierarchy model, as in Camerer and Ho (2004), which assumes subjects' beliefs about others are tied with their own levels of cognition. However, this could be explained by other level-*k* models, such as Stahl and Wilson (1995), Costa-Gomes, Crawford and Broseta (2001) and Costa-Gomes and Crawford (2006), which assume that subjects are fully rational (capable of any high level thinking), but their beliefs about others may not be consistent with the choices of others. Therefore, this explanation points to a subtle difference between the two classes of level-*k* models in the literature, and should be explored with more experimental evidence in the future.

Reference

Brainard, David H. "The Psychophysics Toolbox." *Spatial Vision*, 1997, *10*, pp. 433-36.

Camerer, Colin. "Progress in Behavioral Game Theory." *Journal of Economic Perspectives*, 1997, *11*(4), pp. 167-188.

Camerer, Colin. "Behavioral Game Theory: Experiments in Strategic Interaction." 2003, Princeton University Press.

Camerer, Colin, Eric Johnson, Talia Rymon, and Sankar Sen. "Cognition and Framing in Sequential Bargaining for Gains and Losses." 1993, *Frontier of Game Theory*, p. 27-47, ed. by Ken Binmore, Alan Kirman, and Piero Tani, MIT Press.

Camerer, Colin, Teck-Hua Ho, and Juin-Kuan Chong. "A Cognitive Hierarchy Model of Games" *Quarterly Journal of Economics*, 2004, *119*(3), pp. 861-98.

Chou, Eileen, Margaret McConnell, Rosemarie Nagel, and Charles R. Plott. "The Control of Game Form Recognition in Experiments: Understanding Dominant Strategy Failures in a Simple Two Person "Guessing" Game," *Experimental Economics*, 2009, *12*(2), pp. 159-79.

Cornelissen, Frans W., Enno M. Peters and John Palmer. " The Eyelink Toolbox: Eye Tracking with Matlab and the Psychophysics Toolbox." *Behavior Research Methods, Instruments & Computers*, 2002, *34*, pp. 613-17.

Costa-Gomes, Miguel, Vincent Crawford, and Bruno Broseta. "Cognition and Behavior in Normal-Form Games: An Experimental Study." *Econometrica*, 2001, *69*(5), pp. 1193-1235.

Costa-Gomes, Miguel, and Vincent Crawford. "Cognition and Behavior in Two-Person Guessing Games: An Experimental Study." *American Economic Review*, 2006, *96*(5), pp. 1737-68.

Gabaix, Xavier; David Laibson, Guillermo Moloche and Stephen Weinberg. " Information Acquisition: Experimental Analysis of a Boundedly Rational Model." *American Economic Review*, 2006, *96*(4), pp. 1043-68.

Grosskopf, Brit and Rosemarie Nagel. "The Two-person Beauty Contest." *Games and Economic Behavior*, 2008, 62, pp. 93-99.

Hamilton, James D. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica*, 1989, *57*(2), pp. 357-384.

Hamilton, James D. "Time Series Analysis." 1994, Princeton University Press.

Hansen, B. E. "The Likelihood Ratio Test under Nonstandard Conditions: Testing the Markov Switching Model of GNP." *Journal of Applied Econometrics*, 1992, *7*, pp. S61-S82.

Hansen, B. E. "Erratum: The Likelihood Ratio Test under Nonstandard Conditions: Testing the Markov Switching Model of Gnp." *Journal of Applied Econometrics*, 1996a, *11*(2), pp. 195-198.

Hansen, B. E. "Inference When a Nuisance Parameter Is Not Identified under the Null Hypothesis." *Econometrica*, 1996b, *64*(2), pp. 413-430.

Ho, Teck-Hua, Colin Camerer and Keith Weigelt. "Iterated Dominance and Iterated Best Response in Experimental "P-Beauty Contests." *American Economic Review*, 1998, *88*(4), pp. 947-69.

Johnson, Eric, Colin Camerer, Sankar Sen, and Talia Rymon. "Detecting Failures of Backward Induction: Monitoring Information Search in Sequential Bargaining." *Journal of Economic Theory*, 2002, *104*(1), pp. 16-47.

Nagel, Rosemarie. "Unraveling in Guessing Games: An Experimental Study." *American Economic Review*, 1995, *85*(5), pp. 1313-1326.

Pelli, Denis G. "The Videotoolbox Software for Visual Psychophysics: Transforming Numbers into Movies." *Spatial Vision*, 1997, *10*, pp. 437-42.

Salmon, Timothy C. "An Evaluation of Econometric Models of Adaptive Learning." *Econometrica*, 2001, *69*(6), pp. 1597-628.

Selten, Reinhard. "Properties of a Measure of Predictive Success." *Mathematical Social Sciences*, 1991, *21*(2), pp. 153-167.

Stahl, Dale, O. and Paul W. Wilson. "On Players' Models of Other Players: Theory and Experimental Evidence." *Games and Economic Behavior*, 1995, *10*(1), pp. 218-54.

Vuong, Quang. "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica*, 1989, *57*(2), pp. 307-33.

Wang, Joseph Tao-yi, Michael Spezio and Colin F. Camerer. "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth-Telling and Deception in Sender-Receiver Games." *American Economic Review*, 2009, *forthcoming*.

Figures and Tables

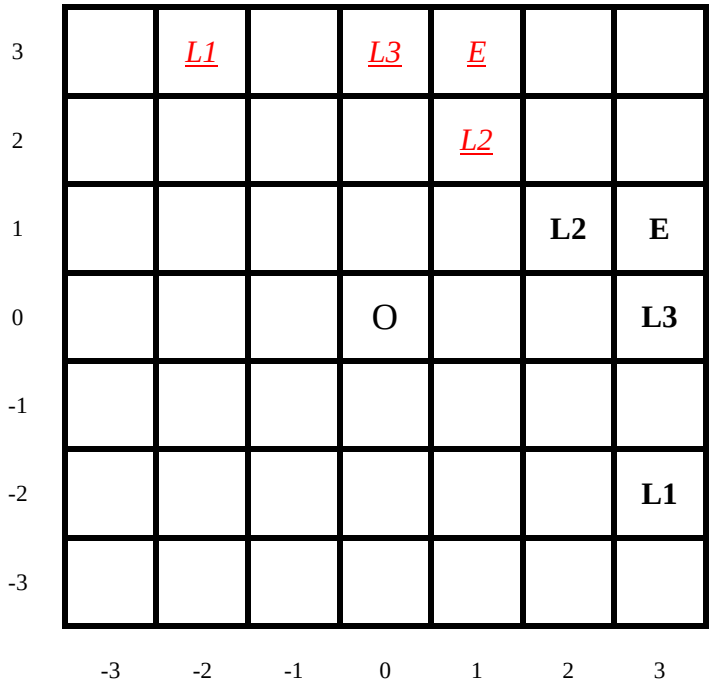


Figure 1: Equilibrium and Level-*k* Predictions of a 7x7 Spatial Beauty Contest Game with Targets (4, -2) and (-2, 4) (Game 16). Predictions for the player with Target (4,-2) are in **bold**, and predictions for the player with Target (-2,4) are in *italic and underlined*. O stands for the prediction of *L0* for both players. Note that **Lk** and *Lk* are the best responses to *L(k-1)* and **L(k-1)**, respectively. For example, L2's choice (2,1) is the best response to L1 since (-2,3) + (4, -2) = (2, 1).

Figure 2A: Screen Shot of the GRAPH Presentation

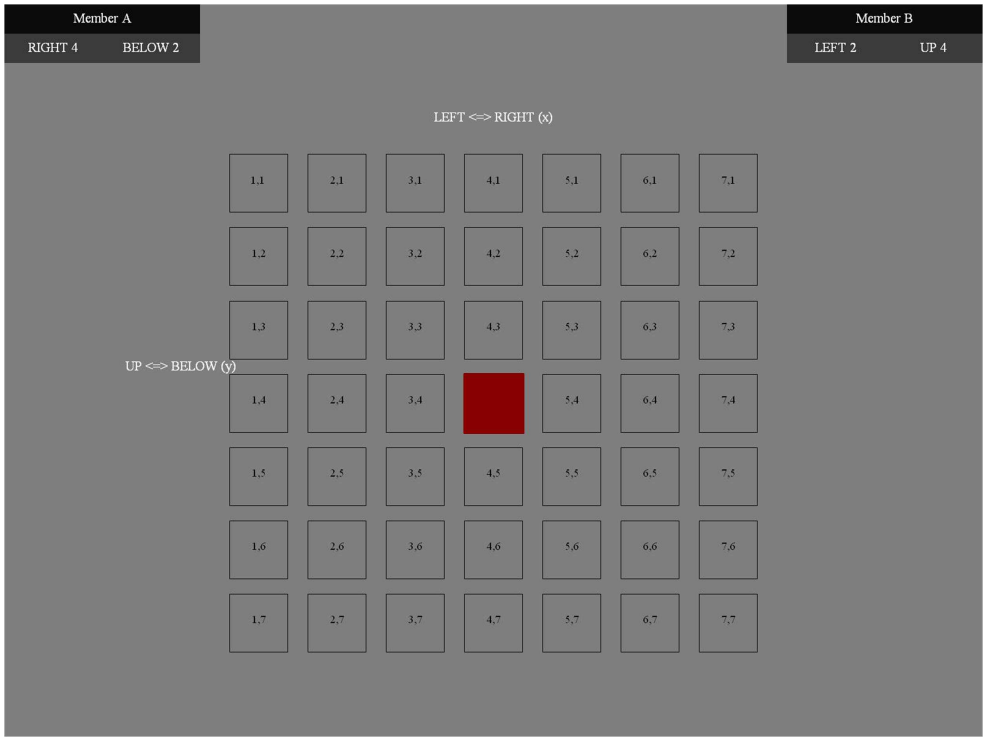


Figure 2B: Screen Shot of the SEPARATE Presentation

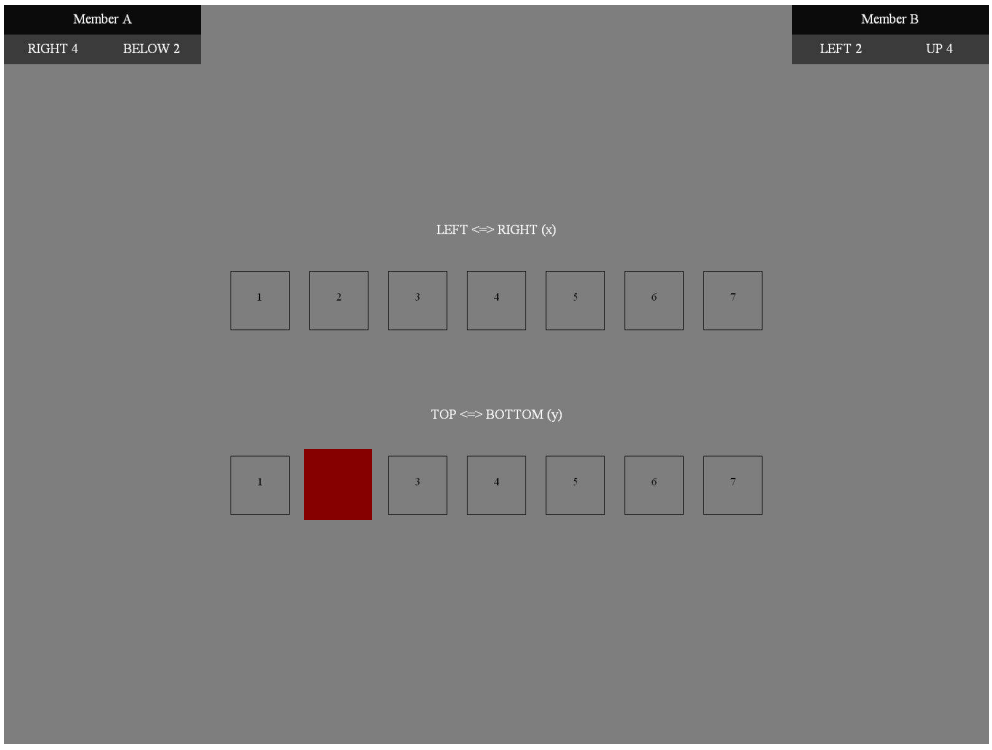
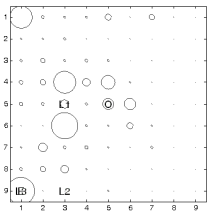


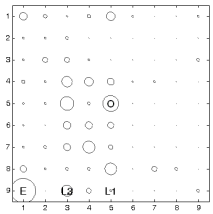
Figure 3A: Aggregate Empirical Percentage of Time Spent on Each Location for Games with 1-dimensional Targets (**GAME 1- GAME 12**). The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, …, E are player *i*'s predicted choices of various level-*k* types.

GAME1



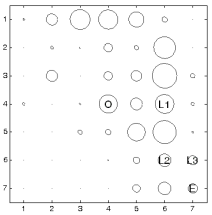
size	*Player i*	*Player j*
9×9	(-2,0)	(0,-4)

GAME 2



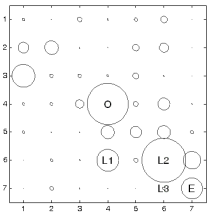
size	*Player i*	*Player j*
9×9	(0,-4)	(-2,0)

GAME 3



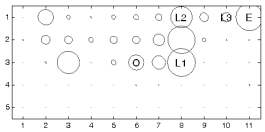
size	*Player i*	*Player j*
7×7	(2,0)	(0,-2)

GAME 4



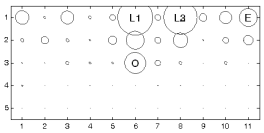
size	*Player i*	*Player j*
7×7	(0,-2)	(2,0)

GAME 5



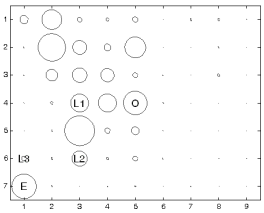
size	*Player i*	*Player j*
11×5	(2,0)	(0,2)

GAME 6



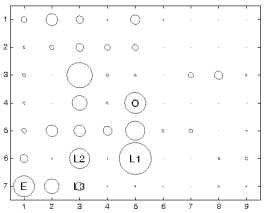
size	*Player i*	*Player j*
11×5	(0,2)	(2,0)

GAME 7



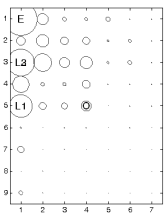
size	*Player i*	*Player j*
9×7	(-2,0)	(0,-2)

GAME 8



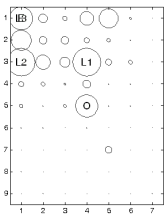
size	*Player i*	*Player j*
9×7	(0,-2)	(-2,0)

GAME 9



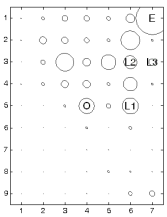
size	*Player i*	*Player j*
7×9	(-4,0)	(0,2)

GAME 10



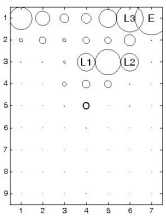
size	*Player i*	*Player j*
7×9	(0,2)	(-4,0)

GAME 11



size	*Player i*	*Player j*
7×9	(2,0)	(0,2)

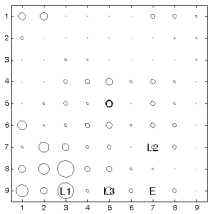
GAME 12



size	*Player i*	*Player j*
7×9	(0,2)	(2,0)

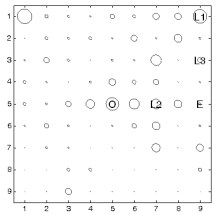
Figure 3B –Aggregate Empirical Percentage of Time Spent on Each Location for Games with 2-dimensional Targets (**GAME 13- GAME 24**). The radius of the circle is proportional to the average percentage of time spent on each location, so bigger circles indicate longer time spent. O, L1, …, E are player *i*'s predicted choices of various level-*k* types.

GAME 13



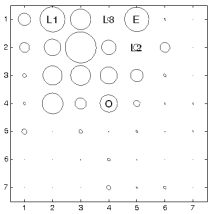
size	*Player i*	*Player j*
9×9	(-2,-6)	(4,4)

GAME1 4



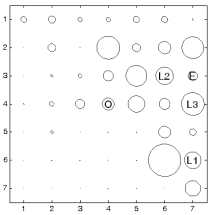
size	*Player i*	*Player j*
9×9	(4,4)	(-2,-6)

GAME 15



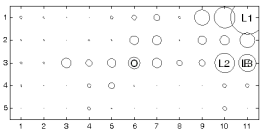
size	*Player i*	*Player j*
7×7	(-2,4)	(4,-2)

GAME 16



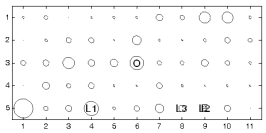
size	*Player i*	*Player j*
7×7	(4,-2)	(-2,4)

GAME 17



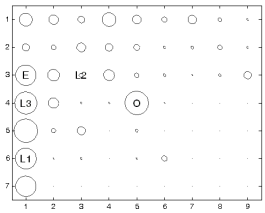
size	*Player i*	*Player j*
11×5	(6,2)	(-2,-4)

GAME 18



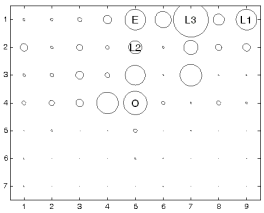
size	*Player i*	*Player j*
11×5	(-2,-4)	(6,2)

GAME 19



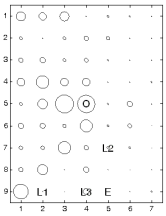
size	*Player i*	*Player j*
9×7	(-6,-2)	(4,4)

GAME 20



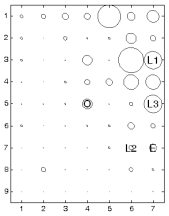
size	*Player i*	*Player j*
9×7	(4,4)	(-6,-2)

GAME 21



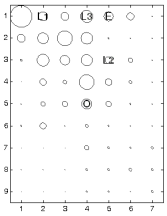
size	*Player i*	*Player j*
7×9	(-2,-4)	(4,2)

GAME 22



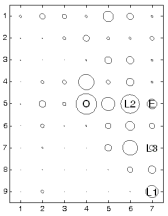
size	*Player i*	*Player j*
7×9	(4,2)	(-2,-4)

GAME 23



size	*Player i*	*Player j*
7×9	(-2,6)	(4,-4)

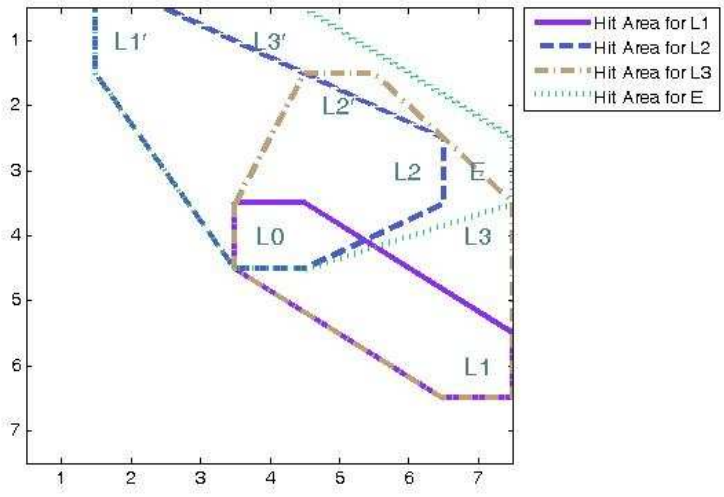
GAME 24



size	*Player i*	*Player j*
7×9	(4,-4)	(-2,6)

Figure 4: *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.

Hit area



Note: In general, if we use the apostrophe to denote it is about the opponent and follow the notations defined in III.C, the minimal convex set enveloping the locations (***L0***, …, ***L(k-2)***, ***L(k-1)'***, ***Lk***) predicted by various level-*k* types are illustrated.

Figure 5: Aggregate Empirical Percentage of Time Spent on the Union of *Hit* Areas ("Hit Time") in Each Game

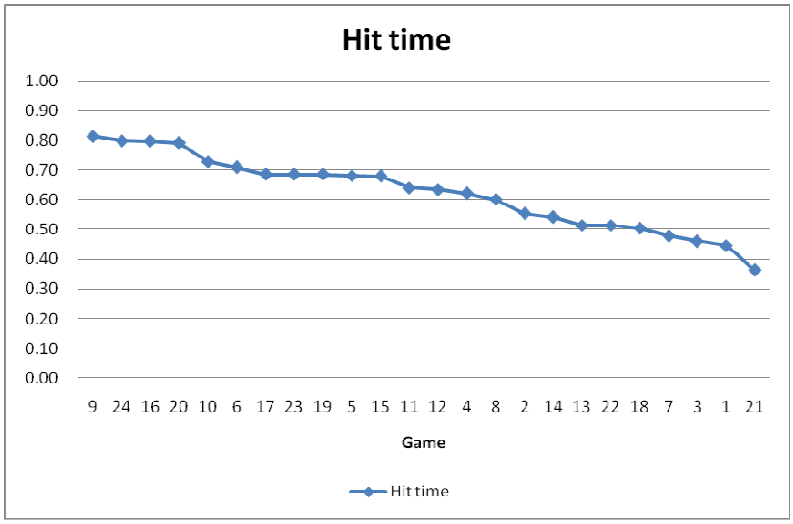


Figure 6: Aggregate Linear Difference Measure of Predicted Success in Each Game. It measures the difference between hit time and the size percentage of the union of the *Hit* area.

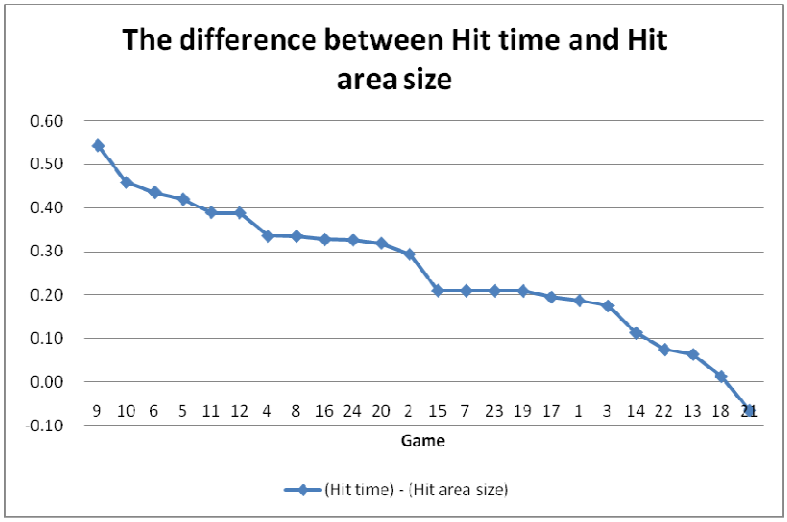


Figure 7: Subject 2's Eye Lookups in Round 17 (as Member B). The radius of the circle is proportional to the length of that lookup, so bigger circles indicate longer time spent.

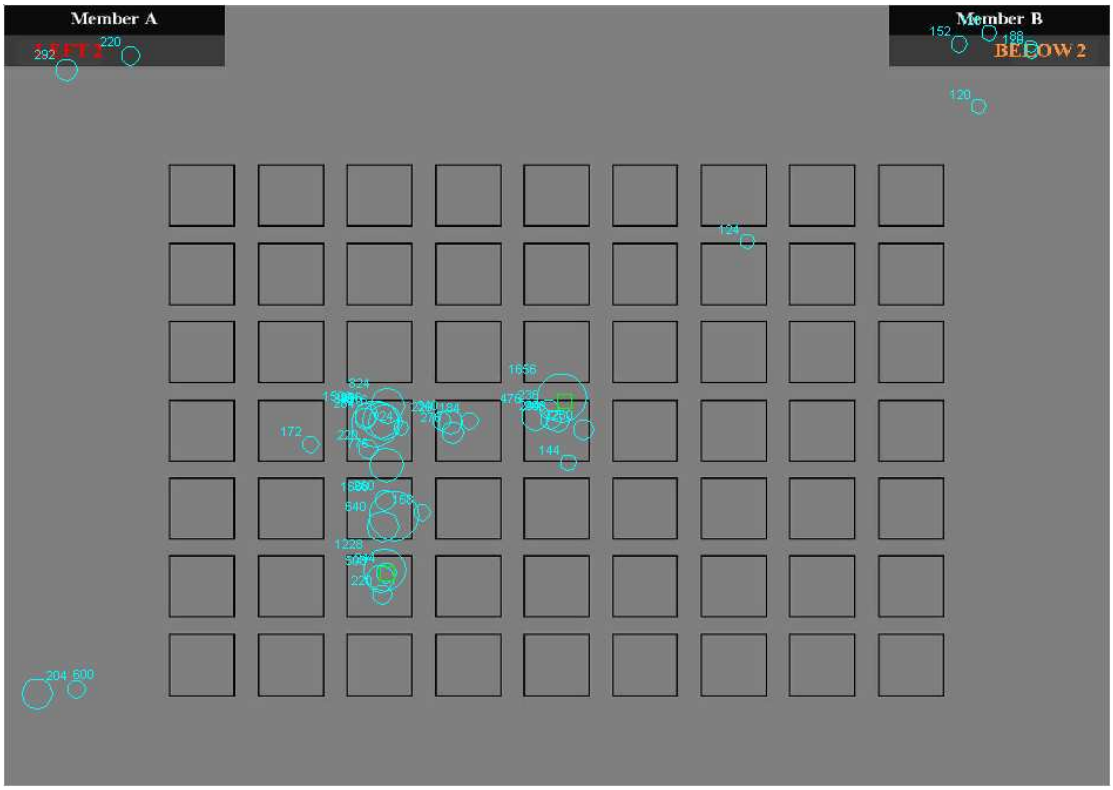


Table 1: Level-*k*, Equilibrium Predictions and Minimum $\bar{k}$'s in All Games

Game	*Map size*	*Player 1 target*	*Player 2 target*	*L0*	*L1*	*L2*	*L3*	*EQ*	$\bar{k}$
1	9×9	-2, 0	0, -4	0, 0	-2, 0	-2, -4	-4, -4	-4, -4	3
2	9×9	0, -4	-2, 0	0, 0	0, -4	-2, -4	-2, -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	2, 2	2, 2	5, 2	5
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	-3, 2	-3, 2	-3, 4	4
10	7×9	0, 2	-4, 0	0, 0	0, 2	-3, 2	-3, 4	-3, 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	5, 0	5, 0	3
18	11×5	-2, -4	6, 2	0, 0	-2, -2	3, -2	2, -2	3, -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*, (6) the theoretic prediction of *L1*, (7) the theoretic prediction of *L2*, (8) the theoretic prediction of *L3*, (9) the theoretic prediction of *EQ*, and (10) the minimum $\bar{k}$ 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 underlined.

Table 2: Distribution of Types under Various Specifications

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

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), uses the choice data in which pseudotypes are not included. The second classification, reported in column (B), also uses the choice data but in addition, pseudotypes are included. The third classification, reported in column (C), 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 types (and to the type with the next largest likelihood among all lower types otherwise).

Table 3. Distribution of Types in 1000 Times of Resampling of Choice Data

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

Note: * indicates misclassification rate less than 0.05.

+ indicates each subject's lookup-based classification of type in Table 4. Notice that the lookup-based types are typically the second most frequent types subjects are classified into (if not the most frequent types) if we resample their choices. The only exceptions are subject 7 and 14.

Each row corresponds to a subject and contains the following information in order: (1) the subject number, (2) her choice-based level-*k* type denoted by Lk^c, (3) the number of times that she is classified as an *L0* in 1000 times of resampling of her choice data, (4) the number of times that she is classified as an *L1* in 1000 times of resampling of her choice data, (5) the number of times that she is classified as an *L2* in 1000 times of resampling of her choice data, (6) the number of times that she is classified as an *L3* in 1000 times of resampling of her choice data, (7) the number of times that she is classified as an *EQ* in 1000 times of resampling of her choice data, and (8) the misclassification rate, i.e., the number of times that she is not classified as her choice-based level-*k* type or Lk^c in 1000 times of resampling of her choice data divided by 1000.

Table 4: Distribution of Types Based on Lookup Data

subject	Lk^*	Lk^a	Vuong's statistic *V*	Lk^l
1	L3	L2	4.425 +	L3
2	L3	L2	0.689	L2
3	L3	L1	1.577	L1
4	L3	L1	1.597	L1
5	EQ	L2	2.977 +	EQ
6	EQ	L2	2.400 +	EQ
7	L2	L0	1.582	L0
8	L3	L1	2.812 +	L3
9	EQ	L2	1.001	L2
10	L3	L1	1.226	L1
11	L3	L2	2.087 +	L3
12	L3	L1	0.853	L1
13	L3	L1	3.939 +	L3
14	L3	L1	1.692	L1
15	L3	L2	1.470	L2
16	L3	L2	1.342	L2
17	L3	L1	1.778	L1

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

Lk^* denotes the type with the largest likelihood.

Lk^a denotes the alternative lower type which has the second-largest likelihood.

Lk^l denotes the classified type based on Vuong's test result.

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 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. 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. The result in (5) is summarized in column (C) of Table 2.

Table 5: Comparison between Choice-based and Lookup-based Classifications

subject	Lk^l	Lk^c	Vuong's statistic V	*Misclassification Rate of* Lk^c
1	L3	L3	.	0.000*
2	L2	L2	.	0.076
4	L1	L1	.	0.110
5	EQ	EQ	.	0.012*
7	L0	L0	.	0.034*
10	L1	L1	.	0.000*
12	L1	L1	.	0.010*
13	L3	L3	.	0.004*
16	L2	L2	.	0.000*
17	L1	L1	.	0.232
3	L1	L3	-1.577	0.244
6	EQ	L2	2.400+	0.236
8	L3	EQ	2.636+	0.000*
9	L2	L0	2.981+	0.472
11	L3	L2	2.087+	0.365
14	L1	EQ	1.395	0.413
15	L2	L3	-1.470	0.184

Note: + indicates the Vuong statistic V is significant or $|V|>1.96$.
* indicates p-value less than 0.05.

Lk^l denotes a subject's lookup-based type.

Lk^c denotes a subject's choice-based type.

Each row corresponds to a subject and contains the following information in order: (1) the subject number, (2) her lookup-based type (as reported in the last column of Table 4), (3) her choice-based type (as reported in the second column of Table 3), (4) Vuong's statistic on whether her lookup-based type and her choice-based type are equally good models (based on lookup data), (5) the misclassification rate of her choice-based type in 1000 times of resampling (as reported in the last column of Table 3). Subjects whose lookup-based and choice-based classifications coincide are listed in the top panel; those who differ are listed in the bottom.