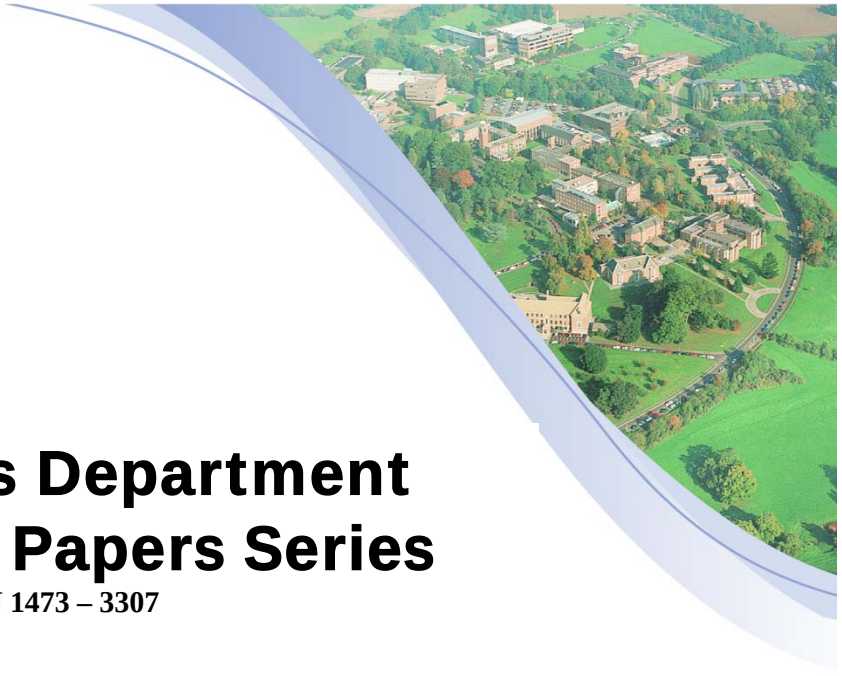

# Economics Department Discussion Papers Series

ISSN 1473 – 3307

## Explaining Behavior in the "11-20" Game

Lawrence C.Y Choo and Todd R. Kaplan

**Paper number 14/01**

URL: http://business-school.exeter.ac.uk/economics/papers/
URL Repec page: http://ideas.repec.org/s/exe/wpaper.html

# Explaining Behavior in the "11-20" Game

Lawrence C.Y Choo*and Todd R. Kaplan†

January 2, 2014

**Abstract**

We investigate whether subjects' behavior in the Arad and Rubinstein (2012) "11-20" game could be well explained by the k-level process described by the authors. We replicated their game in our baseline experiment and provided two other variations that retained the same mixed-strategy equilibrium but resulted in different predicted behavior by the k-level process. Our experiments results suggest that k-level process leads to inconsistent predictions. In contrast to the standard k-level process as in Arad and Rubinstein, we allow players to best respond stochastically in our "SK" model and compared the model's statistical fit against the Quantal Response Equilibrium and Cognitive Hierarchy Model. The SK model and Cognitive Model were able to outperform the QRE in a statistical sense and performed as well as each other. In addition, the Cognitive Hierarchy and to lesser extend the SK model, demonstrate consistent estimates. Our findings suggest that the behavioral assumptions of Arad and Rubinstein k-level process does not fully explain behavior in the "11-20" and better explanations could be obtained when one allows for stochastic best responds as in the SK and Cognitive Hierarchy Models.

**Keywords:** *k-level, Cognitive Hierarchy, Quantal Response Equilibrium, "11-20" money request game*
**JEL Classification:** C73, C91

## 1 Introduction

Deviations from the equilibrium predictions are a well-documented phenomenon in the literature of economic and game theory experiments. The challenge in this field is the provision of better explanatory models. One explanation posit that people often avoid the circular concepts embedded in equilibrium outcomes, in preferences for rule-of-thumb behaviors (Crawford, Costa-Gomes, and Iriberri, 2013). Furthermore, this explanation suggest that such rule-of-thumb behaviors are closely associated with the steps of iterative reasoning employed by players. The k-level (Nagel, 1995; Stahl and Wilson, 1994, 1995; Costa-Gomes, Crawford, and Broseta, 2001) model is a leading candidate in this field.

The standard k-level model partitions the population of players into specific $L_k$ types ($k \in \mathbb{N}$). The model thereafter posits a set hierarchal behavioral rules depending on the types of players. This hierarchy begins with a non-strategic $L_0$ type, who is assumed to follow some common knowledge behavioral rule. Such behavior is analogous to the players' *instinctive reaction* in the game and is often taken to be the uniform randomization over all strategies.[1] Higher $L_k$ types ($k > 0$), assume that all other players are type $L_{k-1}$ and best respond to these

*University of Exeter: cylc201@exeter.ac.uk
†University of Exeter and University of Haifa

[1]Whether the uniform randomization is the appropriate specification of the $L_0$ type behavioral is by itself a debate. We find it difficult to see how even a non-strategic player will assign equal weights to strategies that are payoff dominant and dominated.

beliefs via iterative thought-experiments.[2] A $L_1$ type selects a strategy that is the best respond to the $L_0$ behavior, a $L_2$ type to a $L_1$, a $L_3$ type to a $L_2$ and so forth. Higher types are hence strategic in the same tradition of "rationalizable" behaviors (Bernheim, 1984; Pearce, 1984). In games, the aggregate strategies are attributed to a specific proportion of $L_k$ types. The model is simple and intuitive, however applications to wider economic settings first requires some prior on the plausible proportion of types. The popularity of the model has led to a growing collections of experiment and games that sought to elicit or estimate the proportion of $L_k$ types in a generalized population.[3] The mode in most experiments was found to be the $L_2$ type.

A common concern, with any k-level investigation, is the exogenous specification of the $L_0$ type behavior which the entire model anchors upon. If such specification is not salient amongst the populations of subjects, then any estimates proportion of $L_k$ types will most likely be misspecified. To overcome this problem Arad and Rubinstein (2012) - henceforth known as AR - proposed the "11-20" game, which the authors argue to possess a salient $L_0$ specification. The game involves two players simultaneously choosing an amount of payoff between 20.00 to 11.00 (integers), which they are certain to receive. In addition, a player receives a bonus payoff of 20.00, if his choice is exactly 1.00 less than the other player. The game has no pure strategy equilibrium, only a mixed-strategy equilibrium that assigns positive probabilities to the strategies 20.00 to 15.00. In their experiments, AR found that subjects' behaviors were significantly different from the equilibrium distribution and proposed the *k-level process* to explain such deviations. AR make three assumptions in their analysis (1) $L_0$ type player will always choose 20.00, (2) Higher $L_k$ types will always perfectly best respond to their beliefs and (3) there exist a highest type $L_{\bar{K}} = 9$. As such, a $L_1$ type player best responds with 19.00, a $L_2$ with 18.00, a $L_3$ with 17.00 and so forth.[4] AR argue that the $L_0$ type behavior is salient in their game since it represented the highest payoff a player could receive without consideration for the behavior of other players. Given their assumptions, the relative proportion of $L_k$ types could be inferred from the observed strategies in the experiment. Types $L_1$, $L_2$ and $L_3$ are now frequently found in the proportion 0.12, 0.30 and 0.32 respectively. Subsequent adaptions of the "11-20" game have been utilized by Lindner and Sutter (2013) to study decision making under time pressure and Alaoui and Penta (2013) to study iterative reasoning formation.

This paper is motivated by concerns as to whether subjects' behavior in the "11-20" were well explained by the k-level process. While the game is simple and intuitive, might it be too simple to be described by the k-level process? The premise of the k-level process is that subjects who do $k$ steps of iterative thought-experiments do not expect any other players to do $k+1$ steps, otherwise they would respond with $k+2$ steps. The justification is usually found in the psychological evidence of overconfidence in one's abilities (see Camerer and Lovallo, 1999; DellaVigna, 2009). Therefore, we should expect each additional step of iterative thought-experiments to be less obvious or more *cognitively* demanding. However, the nature of the "11-20" game meant that subjects who do one step of iterative thought-experiment could easily extend it to two steps or more without incurring significantly cognitive cost.[5] If

[2]In the k-level literature, players are not necessarily restricted to the $L_k$ types. For example Costa-Gomes, Crawford, and Broseta (2001) examine a k-level model that includes the four different $L_0$ types, *Altruistic, Pessimistic, Naive* and *Optimistic*. The authors also considered five strategic types starting from our $L_2$ type, the $D_1$ & $D_2$ types that eliminates dominated strategies of one and two steps, the *Sophisticated* type that best responds to the actually probabilities distribution of his opponent and the *Equilibrium* type that makes equilibrium choices.

[3]Examples of k-level investigation in the Guessing Game: Bosch-Domènech, Montalvo, Nagel, and Satorra (2002); Grosskopf and Nagel (2008); Burchardi and Penczynski (2010); Chou, McConnell, Nagel, and Plott (2009); Costa-Gomes and Crawford (2006); Duffy and Nagel (1997); Kocher and Sutter (2005); Weber (2003); in normal form games Costa-Gomes, Crawford, and Broseta (2001); Stahl and Wilson (1994, 1995), in Rosenthal (1981) centipede game Kawagoe and Takizawa (2012); in auctions Crawford and Iriberri (2007); in coordination games Costa-Gomes, Crawford, and Iriberri (2009); Weber (2001).

[4]Without the highest $L_{\bar{K}}$ type, the k-level process in the "11-20" game induces *cycles,* such that the choice of 20.00 can be attributed to $L_0$, $L_{10}$, $L_{20}$ types and so forth.

[5]One of the most frequently discussed game in the k-level literature is Nagel (1995) guessing game. Here a group ($n \geq 2$) of players

subjects were indeed adhering to the k-level process as described by AR, shouldn't even higher types, e.g., $L_4$, $L_5$, $L_6$ be more frequently observed. Alternatively, could subjects' behavior also be explained by some statistical distortion of the mixed-strategy equilibrium such as the Quantal Response Equilibrium (McKelvey and Palfrey, 1995, 1996, 1998) or some other coherent distortion owning to the game's structure. Eventually with experimental data, there could be multiple competing explanations. The central question in this paper is whether the k-level process proposed by AR is the dominant explanation and the challenge is to put forth a suitable experiment design and hypothesis to test this question.[6]

Denoting the "11-20" game as the baseline game, we take up this challenge with two simple extensions, the Medium and Extreme game. In the Medium game, players choose from following payoffs 20.00, 19.50, 19.00,..., 11.00 which they are certain to receive. The bonus of 20.00 is only awarded if the player's choice is 0.50 or 1.00 less than the other player. In the case of the Extreme game, players choose from the payoffs 20.00, 19.75, 19.50, 19.25, 19.00,..., 11.00 which they are certain to receive. However, the bonus is now only awarded if the player's choice is 0.25, 0.50, 0.75 or 1.00 less than the other player. All games - Baseline, Medium and Extreme - share the same structure and decisional problem. In addition, the games have equivalent mixed-strategy equilibrium distributions (see Table 1). For example, the strategy of {20.00} in the Baseline game, {20.00, 19.50} in the Medium game and {20.00, 19.75, 19.50, 19.25} in the Extreme game are all predicted to be chosen with 5% probability. By the k-level reasoning process, the $L_0$ type in all games will always select the strategy 20.[7] However, the behavioral prediction for higher types are remarkably different. The $L_1$ type now best respond with 19.00, 19.50 and 19.75 in the Baseline, Medium and Extreme games respectively. The corresponding best respond for the $L_2$ types are 18.00, 19.00 and 19.50, for the $L_3$ type 17.00, 18.50 and 19.25, and so forth.

We hence refer to a logical consequence of the k-level process, such that if players were randomly recruited from the same population into the various games, the k-level process should predict the same proportion of types in all games. This leads to a simple hypothesis test. The additional benefit of such design controls for any behavior that might be attributed to the structure of the game and focuses the discussions on the k-level process predictions. Our experiment procedure involves four classroom experimental sessions over two cohort of students. Students in the first cohort were recruited into the Baseline and Medium games, whilst those in the second cohort were recruited in the Medium and Extreme games. The predicted proportion of $L_k$ types as inferred by AR's k-level process were compared between sessions of the same cohort and were found to be significantly different in all comparisons.[8] Hence, the k-level reasoning process is unable to demonstrate consistency at any reasonable level. Though this finding does not preclude the possibility that subjects in the Baseline, Medium or Extreme games were employing some other form of iterative thought-experiments, it suggest that the k-level process was not the dominant explanation.

This conclusion is also shared by Goeree, Louis, and Zhang (2013) who also replicated the "11-20" game and provided two other extensions games which shared the same predicted behaviors for each $L_k$ type but have different mixed-strategy equilibriums. They found that the k-level process explains the data no better than the mixed-

choose a number from 0 to 100. A fixed prized is awarded to that player whose number is closest to 2/3 of the average. If a $L_0$ type player is assume to uniformly randomize across all numbers, a $L_1$ best responds to the uniform randomization. A $L_2$ best respond to the best respond of a uniform randomization. Owning to the game design, the best respond task becomes more challenging as the step of iterative thought-experiments increases.

[6] In their paper, AR consider two other extension of the "11-20" game, the *costless iteration* and *cycle* versions. Both extensions sought to investigate the saliency of the $L_0$ type behavior assuming that the data was generated by the k-level process.

[7] Despite the games' differences, there should be no reason why a $L_0$ type player will behave any differently in either games. Choosing 20.00 in either games still accords $L_0$ type player the highest payoff without regard for the behavior of the other player.

[8] Comparisons between sessions of the same cohort controls for the demographic differences between sessions.

strategy equilibrium when the predicted proportion of $L_k$ types in their replication "11-20" experiment was fitted onto the data of the other two extensions. The authors argue that the out-of-sample fit could be enhanced if one assumes the "common knowledge noise" as in the Quantal Response Equilibrium (QRE) or the Noisy Introspection (NI) Model (Goeree and Holt, 2004). Allowing for such noise makes the QRE and NI more flexible than the k-level process but induces the discussion as to whether the performance of the NI and QRE represent a better description of subjects' behavior or mere statistical observations.

We hence revisit the k-level process allowing for such noise - henceforth known as the SK model to distinguish from the k-level process. In addition, we also considered the closely related Cognitive Hierarchy (CH) model by Camerer, Ho, and Chong (2004), a close "cousin" of the SK model, that differs on the beliefs of higher type players. The SK, CH and QRE were fitted onto the respective sessions data via econometric methods. Employing Vuong (1989) likelihood ratio test, we show that the CH and SK models were able to fit the data significantly better than the QRE (and mixed-strategy equilibrium). However, the CH and SK models fitted the data as well as each other. Returning to the central hypothesis test, the estimated proportion of $L_k$ types in the CH model was found to be consistent between sessions of the same cohort. The SK model estimated proportion of types were found to be consistent for sessions in the second cohort and to a much lesser extend, the first cohort.

This paper is organized as followed: Section 2 details our experimental procedure, Section 3 provides an overview of the data and investigate AR's k-level process, Section 4 introduces the SK and CH models, Sections 5 reports the estimated results of the SK and CH models and finally, Section 6 concludes.

| Cohort | | **2012** | | **2013** | |
|---|---|---|---|---|---|
| Strategies | Eq, | B(2012) | M(2012) | M(2013) | E(2013) |
| 2000-1925 | .050 | .034 | .120 | .110 | .132 |
| 1900-1825 | .100 | .231 | .359 | .374 | .374 |
| 1800-1725 | .150 | .265 | .188 | .154 | .088 |
| 1700-1625 | .200 | .231 | .077 | .088 | .066 |
| 1600-1525 | .250 | .085 | .077 | .066 | .055 |
| 1500-1425 | .250 | .026 | .085 | .044 | .121 |
| 1400-1325 | .000 | .077 | .034 | .066 | .088 |
| 1300-1225 | .000 | .026 | .017 | .033 | .033 |
| 1200-1125 | .000 | .009 | .009 | .011 | .011 |
| 1100 | .000 | .017 | .034 | .055 | .033 |
| $N$ | | 117 | 117 | 91 | 91 |

Table 1: Summary of Observed Choice Frequencies and Mixed Strategy Distributions

## 2 Experiment Procedure

Four classroom experimental sessions were conducted at the University of Exeter, over two cohorts of Intermediate Microeconomics students. The subjects were mostly economics majors and with no formal training in game theory. We denote each session by the game which the subjects were enrolled into, followed by the cohort which they were recruited from. For example, session B(2012) refers to the Baseline game conducted with subjects from cohort 2012. All sessions were conducted during the first lecture class of the course (approximately 250-300 students in each class), subjects were informed that their participation was voluntary and to refrain from conversing with each other.

In each cohort, the layout of the lecture class had consisted of three separated seated columns. With cohort

2012 and 2013, subjects in the center seated column received the instructions for sessions B(2012) and M(2013) respectively. With cohort 2012 and 2013, subjects in the two other side columns received instructions for sessions M(2012) and E(2013) respectively. For the respective games, the instructions were as followed:

> **Baseline (B) Game:** *You and another player will simultaneously request an amount of payoff from the set {2000, 1900, 1800, 1700, ..., 1100} denoted in ECU. Each player will receive his chosen amount. In addition, a player will receive a bonus of 2000 if his request amount is 100 ECU less than the other player.*
>
> **Medium (M) Game:** *You and another player will simultaneously request an amount of payoff from the set {2000, 1950, 1900, 1850, ..., 1100} denoted in ECU. Each player will receive his chosen amount. In addition, a player will receive a bonus of 2000 if his request amount is (a) 50 ECU or (b) 100 ECU less than the other player.*
>
> **Extreme (E) Game:** *You and another player will simultaneously request an amount of payoff from the set {2000, 1975, 1950, 1925, 1900, ..., 1100} denoted in ECU. Each player will receive his chosen amount. In addition, a player will receive a bonus of 2000 if his request amount is (a) 25 ECU, (b) 50 ECU, (c) 75 ECU or (d) 100 ECU less than the other player.*

Subjects had to circle their choice on a table consisting of all the relevant request amounts. In addition, subjects were to include their contact details and a brief feedback of their behavior. The sessions were completed within 15 minutes and the instruction sheets were thereafter collect by the experimenters. In each cohort, ten pairs of subjects were randomly selected for cash payment (they were privately contacted via email) at the exchange rate of 100 ECU to £1. A total of 130, 140, 114 and 94 subjects participated in sessions B(2012), M(2012), M(2013) and E(2013), respectively.

We choose to split the sessions by the seated columns for ease of instructions distribution and to avoid any confusion created by subjects seeing the other instructions. Furthermore, this procedure is consistent to that of AR experiments which were also conducted in classroom settings. However, the same experimental procedure induces concerns that there might be some natural difference in behavior due to the seated positions of subjects.[9]

To address such concerns, the respective sessions were immediately followed up by the Guessing Game (Nagel, 1995).[10] Here each player chooses a number between 0 to 100 and a fixed prized is awarded to the player whose chosen number is closest to a target number, that is derived to be 2/3 multiplied by the average of all chosen numbers. Subjects in each cohort competed against each other for a fixed prize of £50, were informed that the Guessing Game was a different experiment from the previous sessions and that their participations was voluntary. The Guessing Game instructions sheets were distributed and collected within 20 minutes. A total of 274 and 206 subjects participated in the Guessing Game for cohorts 2012 and 2013 respectively.

To control for our concerns in the sessions' data, we had firstly excluded all observations where subjects had not participated in the Guessing Game. Thereafter, in each cohort, we employed the k-mean clustering algorithm to identify equal session sample sizes, such that the cumulative distribution of Guessing Game numbers in each session sample was not significantly different from each other.[11] This resulted in 117 and 91 observations in each session of cohort 2012 and 2013 respectively.

[9] A common observation in our lecture class was that the attentive students had tended to occupy the frontal rows of the center columns.

[10] We employed the Guessing Game since it was one the most frequently studied games in the literature of $k$-level model.

[11] We verified these results with the Kolmogorov–Smirnov test which reports a pvalue of 0.242 (0.453) in cohort 2012 (2013).

# 3 Investigation AR's k-level Process

The summary of the sessions' results are reported in Table 1. The first column refers to the strategies, the second column the mixed-strategy equilibrium and the third column onwards, the aggregated strategies chosen in the respective sessions. As an empirical warm-up, we first investigate if subjects' behaviors in the respective sessions were consistent with the mixed-strategy equilibrium. Here, Fisher's exact test finds all sessions' data to be significantly different (two-sided Fisher $\rho < 0.001$ for all comparisons).[12]

> **Result 1:** *Subjects' behaviors in sessions B(2012), M(2012), M(2013) and E(2013) were found to be significantly different from the mixed strategy equilibrium.*

A prominent difference pertains to the strategies 1600-1425. These were predicted to be chosen by 50% of the subjects in each session, but were observed to be chosen by no more than 18% of subjects in any session. Comparing between sessions of the same game, the B(2012) session data was not found to be significantly different from AR's results (two-sided Fisher $\rho = 0.323$), confirming their results. This finding was also shared by replications of the baseline session by Lindner and Sutter (2013) and Goeree, Louis, and Zhang (2013).

> **Result 2:** *Subjects' behaviors in our B(2012) session was not found to be significantly different to those of Arad and Rubinstein (2012) experiment.*

Similarly, the M(2012) and M(2013) sessions' data were not found to be significantly different (two-sided Fisher $\rho = 0.483$). The findings insofar suggest that there might be some coherent structure in the behavior of subjects in the respective games. The central question here is whether such behavior is well explained by the AR's k-level process.

As mentioned, AR's k-level process is characterized by three assumptions (1) $L_0$ type player will always choose 2000 in all games, (2) Higher $L_k$ types will always perfectly best respond to their beliefs and (3) there exist a highest type $L_{\bar{K}} = 9, 16, 36$ in the Baseline, Medium and Extreme game respectively. The predicted proportion of $L_k$ types were hence directly inferred from the relative choices in the sessions' data. We report on table 2, the predicted proportion of $L_k$ types (truncated at the $L_8$ type), as inferred from the sessions' data, given the assumptions of AR's k-level process.

We hence test hypothesis that the predicted proportion of $L_k$ types by the k-level model is consistent across sessions of the same cohort. In sessions cohort 2012, the predicted proportions are significantly different (two-sided Fisher $\rho < 0.001$). In session B(2012), 73% of subjects were classified as types $L_1 - L_3$ whilst the same classification only pertains to 41% of subjects in M(2012).

Similarly, we the predicted proportion of $L_k$ types in cohort 2013 were found to be significantly different (two-sided Fisher $\rho < 0.001$). Here, whilst 40% of subjects in session M(2013) were classified as types $L_1 - L_3$, only 7% of subjects in session E(2013) fall under the same classification. Furthermore, a quarter of all subjects in session E(2013) had chosen the amount 1900, which corresponds to the type $L_4$.

> **Result 3:** *The k-level process proposed by Arad and Rubinstein (2012) leads to significantly different predicted proportion of $L_k$ types between sessions of the same cohort.*

[12] For the purposes of our analysis, we chooses the Fisher Exact test over the conventional $r \times c$ contingency table chi-square test, since the test statistics in the latter test requires each cell to have an expected value of at least 1 and that 20% of the cells to have an expected value of at least 5 (Sheskin, 2003).

This result suggest that the k-level process remains an incomplete description of subjects' behavior. One may of course disagree with our hypothesis test. More specifically, why should we expect the proportion of $L_k$ types to be similar across sessions of the same cohort? In our view, this alternative is merited if the respective sessions involved games that were intrinsically different. However, in the setting of our experiment, this alternative propounds that small modifications to the game results in its own unique prediction of $L_k$ types. Whilst such outcome cannot be exclude, we find it unhelpful, especially if the ambitions of such research is its applicability to wider economic settings.

It is important to emphasis that result 3 merely suggest, that if subjects in the Baseline, Medium and Extreme games were performing some form of iterative thought-experiments, this form is unlikely to be the k-level process as proposed by AR. To investigate if subjects' behavior are actually characterized by some generalized k-level model, we introduce a common knowledge noise to the best responding behavior of higher types.

| Cohort | **2012** | | **2013** | |
|---|---|---|---|---|
| | B(2012) | M(2012) | M(2013) | E(2013) |
| $L_0$ | .034 | .068 | .088 | .066 |
| $L_1$ | .231 | .051 | .022 | .022 |
| $L_2$ | .265 | .162 | .209 | .022 |
| $L_3$ | .231 | .197 | .165 | .022 |
| $L_4$ | .085 | .128 | .099 | .253 |
| $L_5$ | .026 | .060 | .055 | .088 |
| $L_6$ | .077 | .051 | .044 | .011 |
| $L_7$ | .026 | .026 | .044 | .022 |
| $\geq L_8$ | .026 | .256 | .275 | .495 |
| $N$ | 117 | 117 | 91 | 91 |

Table 2: Inferred proportion of $L_k$ types by the Arad and Rubinstein (2012) k-level Reasoning Process

# 4 Modeling Stochastic Best Response

In this section, we relax AR's assumption that all higher types best respond to their beliefs, allowing for a common knowledge noise $\lambda \geq 0$ in their best responding behavior. We shall hence refer to this as the Stochastic k-level (SK) model. This naturally leads to comparisons with the QRE, the rational expectation "statistical refinement" of the mixed-strategy equilibrium. In addition, we also consider the CH model. To provision for a common platform of comparison, we will assume that the individual probability choice functions takes the *logistic* functional form (McFadden, 1976). In the remaining sub-sections, we will detail the SK and CH models. Discussions of the QRE model are omitted since it is well known in the literature.

## 4.1 The SK and CH model

The SK and CH model belong to a class of bounded rationality models. They consider a hierarchical of $L_k$ types but differ on the assumed beliefs of each higher types. In application to our Baseline, Medium and Extreme games, both models consider the $N = \{1, 2\}$ set players where each player $i \in N$ simultaneously chooses the strategy $a_i \in A$. Denote $\pi_i(a_i, a_{-i}) > 0$ as the payoff to player $i$ for choosing strategy $a_i$ if the other player chooses $a_{-i}$.

The models anchor upon a $L_0$ type player who is assumed to always choose the action 2000. For any higher $L_k$ type player $i$, let $b_i^k(g) \in [0, 1]$ denote the proportion of $L_g$ type players he believes to exist in the game. We assume

that $b_i^k(g) = 0$ for all $g \geq k$, implying that each $L_k$ type player $i$ ignores the possibility that other players might do the same or more steps of iterative thought-experiments than himself.[13] The SK model assumes that each higher $L_k$ type player believes everyone else to be exactly one type below, resulting in beliefs

$$b_i^k(g) = \begin{cases} 1 & \text{if } g = k-1, \\ 0 & \text{if } g \neq k-1, \end{cases} \quad \forall k > 0,$$

The CH model on the other hand assumes that each higher $L_k$ type player believes everyone else to be a mixture of lower types, distributed according to a normalized Poisson distribution. More specifically, for any population of players, let $f(k) \in [0,1]$ denote the true proportion of $L_k$ type players. The CH model thereafter assumes that $f(0)$, $f(1)$, ..., $f(k)$,... follows a Poisson distribution with the mean and variance $\tau$, where $f(k|\tau) = \tau^k \exp(-\tau)/k!$. Since the CH model also assumes that each higher type players know the true relative proportion of lower types, they hence hold beliefs

$$b_i^k(g) = \frac{f(g|\tau)}{\sum_{h=0}^{k-1} f(h|\tau)} \quad \forall k > 0,\ g < k,$$

If the true proportion of types are clustered around the lower types, then an interesting consequence of the CH model relative to the SK model, is that the beliefs of higher types players in the former model become more precise as $k$ increases, whereas the beliefs in latter becomes less precise.

Let $p^k(a_i) \geq 0$ denote the probability of a higher type player $i$ choosing strategy $a_i \in A$. The choice probability function of the higher types in either SK and CH model is assumed to be

$$p^k(a_i) = \frac{\exp(\lambda \pi_i(a_i, \cdot))}{\sum_{a_i' \in A} \exp(\lambda \pi_i(a_i', \cdot))} \quad \forall k > 0$$

where $\pi_i(a_i, \cdot) = \sum_{a_{-i} \in A} \pi_i(a_i, a_{-i}) \{\sum_{g=0}^{k-1} b_i^k(g) \cdot p^g(a_{-i})\}$ denotes the expected payoff for a higher $L_k$ type player $i$ with choosing strategy $a_i$.[14,15] With the CH model, Rogers, Palfrey, and Camerer (2009) describe such choice probability function as the the behavior equivalent to a limiting version of the truncated Heterogenous QRE model. The parameter $\lambda \geq 0$ here denotes the common knowledge error rate in the games. As $\lambda \to \infty$, each higher $L_k$ places more weights to the action that accords to him the greatest payoff. Likewise as $\lambda \to 0$, each higher type $L_k$ uniformly randomizes across all strategies.

With data, the SK and CH model are fitted through econometric methods. The econometric results make two predictions, the common knowledge noise $\lambda$ and the proportion of $L_k$ types. We are primarily interest in the latter predictions. The estimation of the SK model first requires some prior arbitrary specification of $L_{\bar{K}} = 2, 3, 4, ...$ the highest $L_k$ type one believes to exist in the data. Thereafter, the proportion of types, $L_0$ through to $L_{\bar{K}}$ and the

[13] Solving a model where $b_i^k(g) \neq 0$ for $g \geq k$ might also be more complex and involve finding a fixed point at each step of iterative thought-experiment (Camerer, Ho, and Chong, 2004).

[14] One could also model the choice probability function with a normalized power function form

$$p^k(a_i) = \frac{(\pi_i(a_i, \cdot))^\lambda}{\sum_{a_i' \in A} (\pi_i(a_i', \cdot))^\lambda} \quad \forall k > 0$$

as in Östling, Wang, Chou, and Camerer (2011), and the results will most probability be identical. We decided upon the *Logistic* functional form for natural comparison against the QRE model.

[15] An alternative specification is to assume that the higher $L_k$ types will uniformly randomize with probability $\varepsilon_k \in [0,1]$ or choose the action which accords the highest expected payoff with probability $(1 - \varepsilon)$ as in Costa-Gomes, Crawford, and Broseta (2001). This alternative may not be immediately applicable to the CH model. Since our objective is to restrict any behavior difference in the Sk-level and CH model to assumptions on $L_k$ type beliefs, we choose not to adapt the alternative specification.

noise $\lambda$ are estimated from the data (this results in $\bar{K}+1$ free parameters). Since the SK model does not impose any parametric requirements on the proportion of $L_k$ types, it presents one with certain amount of *flexibility* in increasing the statistic fit by considering different $L_{\bar{K}}$. Estimation of the CH model usually involve setting an arbitrary high $L_{\bar{K}}$. Thereafter, the parameters $\tau$ and $\lambda$ are estimated from the data given the conditions that $f(\bar{K}|\tau) > 1-\varepsilon$. One should note that given the parametric assumptions on the proportion of types, the CH model is slightly more *restrictive* than the SK model. However, is such restriction tantamount to a significantly worst fit?[16]

# 5 Estimating Models of Stochastic Best Respond

The estimates from the SK and CH models and the QRE were derived through maximum likelihood estimation (see Appendix for discussion of MLE procedure). Deriving for the QRE was straightforward and involved econometrically fitting the data for the noise parameter $\lambda$. The SK model was less straightforward since it involved some prior specification for $L_{\bar{K}}$. To avoid overfitting the SK model ($\bar{K}+1$ free parameters), we first generated the estimates in session B(2012) for $L_{\bar{K}} = 3,4,5,6,7,8,9$. The likelihood ratio test prefers the estimates where $L_{\bar{K}} = 6$ (1% level significance). We hence estimated the other three sessions with the same highest, $L_{\bar{K}} = 6$. The CH model was estimated with an arbitrary high type $L_{\bar{K}} = 20$.

We report on Tables 3 and 4 the estimation results for sessions in cohort 2012 and 2013 respectively. Each table comprises of three panels. The top panel depicts the observed choices and the predicted choices by the SK, CH and QRE estimates. The middle panel reports the test statistics of Vuong (1989) likelihood ratio test. The bottom panel reports the estimated proportion of $L_k$ types by the SK and CH models. We also present on Figure 1, the predicted frequencies of choices by the QRE (dotted lines), SK (solid lines) and CH (dash lines) estimates.

Before we discuss our results, it is important to emphasis the central difference between the various models. The QRE describes a perfectly rational equilibrium where each player best responds to the noisy best respond of each other player. This corresponds to an eloquent fixed-point argument based on the common knowledge noise $\lambda$. Recent research by Goeree, Holt, and Palfrey (2005) sought to provide some structural forms to plausible quantal response, however there still remains some ambiguity to the constitution and determinants of $\lambda$. Namely, what is noise and what drives it? Is it specific to population and/or to the game? These are important questions beyond the scope of this paper. We view such noise as some form of *indecisiveness* in the game.[17] However, we remain nonchalant about the plausible determines of such noise. Furthermore, we should highlight that such noise in the QRE is not necessarily equivalent to those in the CH and SK models. The latter two models do not rely on fixed-point arguments and belong to a class of bounded rationality models. In our interpretation, they could very well refer to different measures of *indecisiveness.* It will hence not be prudent to make comparisons amongst the $\lambda$ estimates.

[16] In applications to a series of Guessing Game results, Camerer, Ho, and Chong (2004) adaption of the CH estimated $\tau \approx 1.61$ (types $L_1$ and $L_2$ most frequent). They found that the CH model fitted the data as well as the conventional k-level model. Given that the different in the behavior of players between the two models only differ from type $L_2$ onwards, we do not find their results surprising since most k-level study on the guessing game also found types $L_1$ and $L_2$ to be most frequent. Results in this experiment may potentially be different since the relative "ease" of employing iterative thought-experiments should imply that higher types e.g. $L_3$, $L_4$ and $L_5$, are more frequently found.

[17] One could take a more process driven view and see $\lambda$ as some estimation smoothing parameter.

| | B(2012) | | | | M(2012) | | | |
|---|---|---|---|---|---|---|---|---|
| Strategies | Obs. | QRE | SK | CH | Obs. | QRE | SK | CH |
| 2000-1950 | .034 | .128 | .055 | .113 | .120 | .220 | .146 | .173 |
| 1900-1850 | .231 | .197 | .230 | .190 | .359 | .268 | .341 | .303 |
| 1800-1750 | .265 | .220 | .264 | .268 | .188 | .187 | .179 | .209 |
| 1700-1650 | .231 | .188 | .230 | .216 | .077 | .111 | .086 | .088 |
| 1600-1550 | .085 | .120 | .085 | .082 | .077 | .072 | .065 | .061 |
| 1500-1450 | .026 | .062 | .025 | .041 | .085 | .051 | .054 | .050 |
| 1400-1350 | .077 | .034 | .075 | .031 | .034 | .037 | ,045 | .041 |
| 1300-1250 | .026 | .022 | .015 | .025 | .017 | .027 | .038 | .034 |
| 1200-1150 | .009 | .016 | .012 | .020 | .009 | .020 | .033 | .029 |
| 1100 | .017 | .012 | .010 | .016 | .034 | .008 | .014 | .012 |
| $\lambda$ | | .0028 | .0020 | .0021 | | .0027 | .0015 | .0017 |
| $\tau$ | | | | 4.09 | | | | 3.90 |
| $-\mathcal{L}$ | | 228.42 | 217.70 | 225.10 | | 308.61 | 302.68 | 304.00 |

| LR test | | CH | QRE | | CH | QRE |
|---|---|---|---|---|---|---|
| | SK | $2.12^a$ | $2.81^a$ | SK | 0.60 | $1.67^b$ |
| | CH | | $1.66^b$ | CH | | $2.27^b$ |

$^a$ : $\rho < 0.1$; $^b$ : $\rho < 0.05$ and $^c$ : $\rho < 0.01$ (one-sided test)

| Types | | $L_0$ | $L_1$ | $L_2$ | $L_3$ | $L_4$ | $L_5$ | $L_6^\dagger$ |
|---|---|---|---|---|---|---|---|---|
| SK Model | B(2012) | .00 | .19 | .26 | .25 | .08 | .01 | .21 |
| | M(2012) | .02 | .06 | .51 | .21 | .10 | .02 | .08 |
| CH Model | B(2012) | .02 | .07 | .14 | .19 | .20 | .16 | .22 |
| | M(2012) | .02 | .08 | .15 | .20 | .20 | .15 | .20 |

$^\dagger$ : Refers to types $L_6$ and greater in the CH model

Table 3: Cohort (2012): SK, CH and QRE Estimates

| | M(2013) | | | | E(2013) | | | |
|---|---|---|---|---|---|---|---|---|
| Strategies | Obs. | QRE | SK | CH | Obs. | QRE | SK | CH |
| 2000-1925 | .110 | .210 | .151 | .188 | .132 | .154 | .218 | .282 |
| 1900-1825 | .374 | .239 | .331 | .289 | .374 | .189 | .330 | .268 |
| 1800-1725 | .154 | .173 | .139 | .174 | .088 | .169 | .099 | .112 |
| 1700-1625 | .088 | .116 | .081 | .088 | .066 | .133 | .078 | .078 |
| 1600-1525 | .066 | .082 | .071 | .067 | .055 | .102 | .068 | .066 |
| 1500-1425 | .044 | .061 | .062 | .056 | .121 | .080 | .060 | .057 |
| 1400-1325 | .066 | .046 | ,055 | .048 | .088 | .065 | .052 | .049 |
| 1300-1225 | .033 | .035 | .048 | .040 | .033 | .053 | .046 | .043 |
| 1200-1125 | .011 | .027 | .042 | .034 | .011 | .044 | .040 | .037 |
| 1100 | .055 | .011 | .019 | .015 | .033 | .010 | .009 | .008 |
| $\lambda$ | | .0024 | .0012 | .0015 | | .0015 | .0012 | .0013 |
| $\tau$ | | | | 3.64 | | | | 3.11 |
| $-\mathcal{L}$ | | 247.58 | 241.38 | 243.02 | | 314.20 | 299.97 | 301.50 |

| LR test | | CH | QRE | | CH | QRE |
|---|---|---|---|---|---|---|
| | SK | 0.69 | $1.89^b$ | SK | 0.81 | $3.01^a$ |
| | CH | | $2.68^a$ | CH | | $2.73^a$ |

$^a$ : $\rho < 0.1$; $^b$ : $\rho < 0.05$ and $^c$ : $\rho < 0.01$ (one-sided test)

| Types | | $L_0$ | $L_1$ | $L_2$ | $L_3$ | $L_4$ | $L_5$ | $L_6$ |
|---|---|---|---|---|---|---|---|---|
| SK Model | M(2013) | .04 | .03 | .93 | .00 | .00 | .00 | .00 |
| | E(2013) | .04 | .06 | .90 | .00 | .00 | .00 | .00 |
| CH Model | M(2013) | .03 | .10 | .17 | .21 | .19 | .14 | .16 |
| | E(2013) | .04 | .14 | .22 | .22 | .17 | .11 | .10 |

$^\dagger$ : Refers to types $L_6$ and greater in the CH model

Table 4: Cohort (2013): SK, CH and QRE Estimates

## 5.1 Comparing Statistical Fit

A natural extension to any econometric results is to examine which model provides the best explanatory power or statistical fit. Readers will usually expect some goodness of fit measures, e.g., $R^2$, pseudo-$R^2$, or information criterions such as AIC or BIC. We believe there to be *interpretative* issues with such approaches. For example, comparisons of $R^2$ favors the model with the highest $R^2$. This approach does not discuss whether the $R^2$ of the favored model is significantly different from the next favored model. Information criterions penalizes additional parameters in preference for more parsimonious models. Whilst such approach might ideal when comparing nested-models, interpretation of such criterion is difficult when one considers models which are non-nested or based on fundamentally different assumptions.

In light of these issues, we adopt a statistical comparison which focuses on the differences in the log-likelihood values of each model. For our purposes, we employ the Vuong (1989) likelihood ratio test, which allows for pairwise comparisons between non-nested models.[18] Assuming that there exist a *true* model, the test investigates which

[18]Vuong (1989) test is not unconventional in the literature of experiments. For example, in their study of the Rosenthal (1981) centipede game Kawagoe and Takizawa (2012) employ the test to compare the statistical fit of their iterative reasoning models against the equilibrium driven models. Similarly, Harrison and Rutström (2009) employ the test to compare between decision theory models.

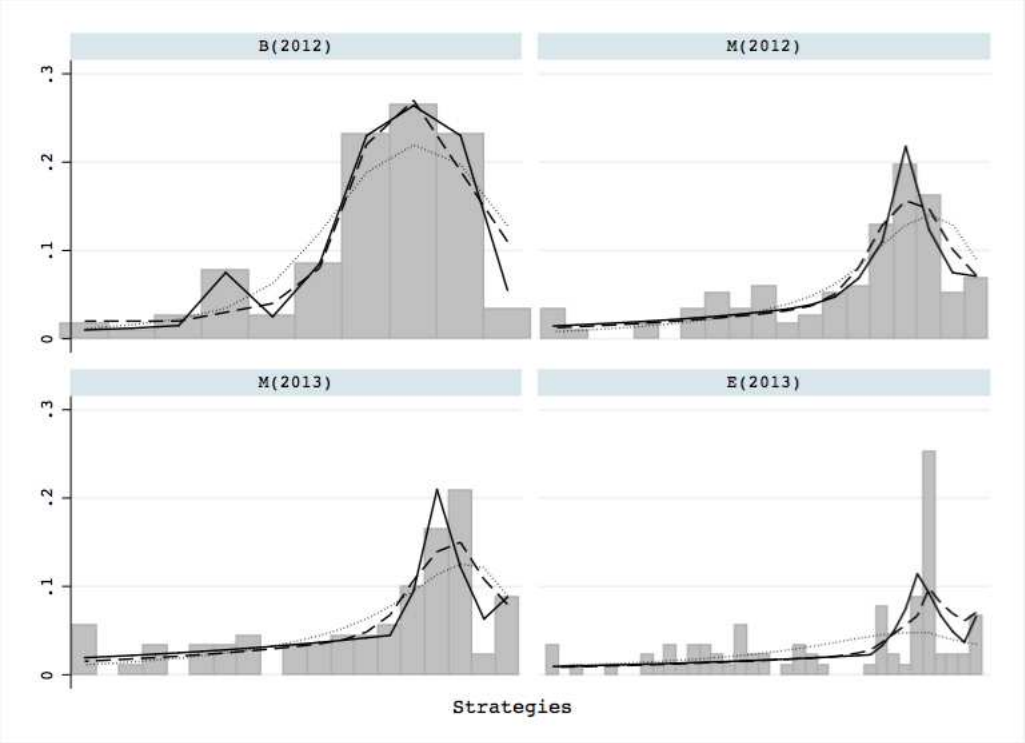


Figure 1: Actual and Predicted Strategy Frequencies - *SK (Solid Lines); CH (Dash Lines); QRE (Dotted Lines)*

of two models is "closer" to the true model. The *null* hypothesis is for both models to be equally close to the true model and two one-sided *alternative* hypotheses, that one of the two models is significantly closer to the true model.[19] The test statistics (reported on the middle panel of Tables 3 and 4) is assumed to follow a standard normal distribution.[20,21] For the ease of interpretation, the test statistics are presented in the following manner: With pairwise comparison, the model with more (less) favorable log-likelihood is position in the row (column) - this ensures that the test statistics must be positive. This allows us to conduct a simple one-sided test to evaluate if the row model is significantly "closer" to the true model or fits the data better, relative to the column model.

**B(2012):** The test prefers the SK and CH models to the QRE ($\rho = 0.002$ & $\rho = 0.048$ respectively). However, the SK is also preferred to the CH model ($\rho = 0.017$). From top left box of Figure 1, the test results become more apparent. The SK model tracks the strategies 2000-1400 much better than the other two models. However, this could also be driven by the fact that such strategies largely correspond to the behavior profiles of $L_0 - L_6$ types, which are by construct free-parameters in the SK model. Although the predicted strategies of the QRE and CH are observed to correctly peak at 1800, the QRE is noticeably under-predicting (over-predicting) the strategies 1800 and 1700 (1600 and 1500) relative to the CH model.

[19] The Vuong (1989) test suffers from some *logical* issues if two fundamentally different models i.e. Rational expectation and Bounded Rationality Models, are found to equally close to the true model. Without loss of generality, the *null* hypothesis can be interpreted as the condition where we are unable to distinguish between the statistical fit of both models.

[20] Corrections for degree of freedom are often employ in applications of (Vuong, 1989) likelihood ratio test, to penalize estimates with more parameters. Whilst such approach might be sensible with nested models, we do not agree that such premises for the purposes of our study since the models are based on drastically different assumptions.

[21] Clarke (2003) proposes a non-parametric alternative test to the Vuong (1989) test. In our results, the conclusion remain consistent if one considers either likelihood ratio test.

**M(2012):** The test prefers the SK and CH models to the QRE ($\rho = 0.048$ & $\rho = 0.012$ respectively). The test is unable to distinguish between the SK and CH ($\rho = 0.274$). From top right box of Figure 1, the SK and CH predicted strategies are observed to correctly peak at 1850 whilst the QRE, at 1800. Furthermore, the QRE under-predicts the three most frequent strategies (1750, 1800 and 1850) relative to the SK and CH. The SK seems to track the strategies better than the CH but such differences were not significant.

**M(2013):** The test again prefers the SK and CH models to the QRE ($\rho = 0.029$ & $\rho = 0.004$ respectively). The test is unable to distinguish between the SK and CH ($\rho = 0.245$). From bottom left box of Figure 1, the SK model is the only model that could account from the sharp drop in strategy frequency at 1950. Again the strategy of 1950 largely corresponds to the behavior profile of the $L_1$ type which is a free parameter in the SK model. However, whilst the QRE and CH are observed to correctly peak at 1900, the SK is observed to peak at 1850. The QRE is noticeably under-predicting the three most frequent strategies (1800, 1850 and 1900) relative to the CH. The performance of the SK relative to the QRE is probably due to its ability to track the lower strategy (1750-1100) better.

**E(2013):** The test prefers the SK and CH models to the QRE ($\rho = 0.001$ & $\rho = 0.003$ respectively). The test is unable to distinguish between the SK and CH ($\rho = 0.209$). From bottom right box of Figure 1, the performance of the SK and CH over the QRE is fairly obvious. Here, the QRE fit is observed to be a small hump with predicted frequencies of around 4% at each strategy 2000-1750 and 3-1% and each strategy 1725-1100. The data exhibits a sharp peak at 1900 (25%) and surprising only CH was able to correct track this peak. However, the predicted strategy frequency at this peak is still nearly 2 times lower. The SK model is again found to peak one choice away from the true peak at 1875.

> **Result 4.** *The SK and CH estimates were found to have fitted the data significantly better than the QRE estimates in all four session. The SK estimates had fitted the data significantly better than the CH estimates in one session and fitted as well in the other three sessions.*

In general, the results from the Vuong test suggest that the SK and CH are able to explain the respective session's data better than the QRE and the mixed-strategy equilibrium in a similar test.[22] The SK model was only found to explain the data significantly better than the CH model in the B(2012) session. When the same test in session B(2012) was conducted with the SK model where $L_{\bar{K}} = 5$, the SK was again found to fit the data significantly better than the QRE but as well as the CH model. This suggest that the superior performance of the SK model over the CH model was primarily driven by the provision of the $L_6$ type. Given these findings, we are confident that despite the Poisson assumptions, the CH model was still on average able to fit the respective session's data as well as the SK model.

## 5.2 Estimated proportion of $L_k$ types

In this sub-section, we hence focus on the estimated proportion of $L_k$ types by the SK and CH models. Given our experimental design and procedure, the theoretical predictions for both models is for the similar estimated proportion of types between sessions of the same cohort.

**Cohort 2012 (SK Model)**: The estimated proportion of types are reported on the bottom panel of Table 3. The $L_2$ types was most frequently estimated in both sessions. However, the proportion of types $L_0$ to $L_6$ in the B(2012) and M(2012) sessions were found to be significantly different (two-sided Fisher $\rho < 0.001$). Concerned

[22] The same conclusions were also made for the QRE.

that such findings were primarily driven by the prior specification of $L_{\bar{K}}$, We conducted the same test for $L_{\bar{K}} = 3, 4, 5, 6, 7, 8, 9$. However, the proportion of types in both sessions were still found to be significantly different (1% level) for each $L_{\bar{K}}$ considered.[23] Returning back to estimates on Table 3, the differences are prominent for type $L_1$ types (0.19 and 0.06), $L_2$ type (0.26 and 0.51) and $L_6$ type (0.21 and 0.08). The estimation procedure of the SK model is of course sensitive to the distribution of data. Hence we consider a lesser hypothesis test focusing on the aggregated estimated proportion of $L_1 - L_3$ types. Here the corresponding frequency in the B(2012) and M(2012) were 0.70 and 0.78 respectively, not significantly different (two-sided Fisher $\rho = 0.295$).

**Cohort 2012 (CH Model)**: The estimates of $\tau$ was found to be 4.09 and 3.90 in sessions B(2012) and M(2012) respectively, suggesting that types $L_3$ and $L_4$ should be most frequently found in both sessions. Since the estimated proportion of types in the CH model are assumed to follow a Poisson distribution, characterized by the mean and variance $\tau$, the reader should naturally expect some formal test on the equality of $\tau$. There is an extended literature on such test, building on the pioneering works of Przyborowski and Wilenski (1940). However, such test assumes that the data generating process follows a Poisson distribution. This is not the case with the CH model, since the assumptions is for the unobservable proportions of types to follow a Poisson distribution and not the data. We therefore take an alternative approach, where we derived the estimated proportions of types from $L_0$, $L_1$,..., $L_5$ and the final proportion that includes type $L_6$ and above (see Table 3). Here, the estimated proportion of types in both sessions were not found to be significantly different (two-sided Fisher $\rho = 0.995$).

**Cohort 2013 (SK Model)**: The estimated proportions of types are reported on the bottom panel of table 4. The $L_2$ types was most frequently estimated in both sessions (at least 0.90). Furthermore, the proportion of types $L_0$ to $L_6$ were not found to be significantly different (two-sided Fisher $\rho = 0.797$).

**Cohort 2013 (CH Model)**: The estimated $\tau$ in sessions M(2013) and E(2013) were found to be 3.64 and 3.11 respectively, suggesting that types $L_3$ were most frequently found in both sessions. The estimated proportion of types are reported on table 4 . Here, the proportion of $L_k$ types were not found to be significantly different (two-sided Fisher $\rho = 0.772$).

> **Result 5a.** *The SK estimated proportions of $L_k$ types were not found to be significantly different between sessions of cohort 2013 and to the lesser extend sessions in cohort 2012.*
>
> **Result 5b.** *The CH estimated proportions of $L_k$ types were not found to be significantly different between sessions of cohort 2012 and cohort 2013.*

One immediate observation is the obvious difference in estimated types between the CH and SK model. Consistent with prior literature on k-level, the SK model found $L_2$ types to be most frequent in all sessions. The average $\tau$ estimated in the CH model was found to be 3.68, which suggest that types $L_3$ and $L_4$ are most frequent. Furthermore, such $\tau$ estimates suggest that types $L_5$ and $L_6$ are still quite substantial. How does one explain such discrepancy? Are the CH model's estimates too high? Firstly (and unfortunately), we believe that any estimated proportion of types is contextualized by the model a researcher employs and the design of the game. One could also investigate if the low types frequency observed in SK model could due to the stipulated $L_{\bar{K}}$. Here one could consider $L_{\bar{K}} = 8, 9, 10, ..., 100$ but such process will obviously be computationally daunting. However, it should be noted that high $\tau$ are not unusual in the literature. For example in their seven week CH model investigation of the Swedish Lottery LUPI game, Östling, Wang, Chou, and Camerer (2011) estimated $\tau$ to be above 4 from week 3 onwards. In some games (not this game), behavior of subjects converges to the Nash equilibrium as $\tau \to \infty$ (Camerer, Ho, and

[23] Even in the most parsimonious case where $L_{\bar{K}} = 3$ the estimated proportions of $L_0$, $L_1$, $L_2$ and $L_3$ types were found to be 0.00, 0.19, 0.34 and 0.47 in session B(2012) and 0.02, 0.05. 0.63 and 0.30 in session M(2012).

Chong, 2004). If one observes adherence to equilibrium behavior in such games, could one inversely argue that such behavior is not possible because it can only be attributed to a extremely high $\tau$. Another plausible explanations is that the $L_k$ types as described in the CH and SK model might not be equivalent. In a recent paper, Kawagoe and Takizawa (2012) estimated a variety of bounded rationality models to investigate behavior in Rosenthal (1981) centipede game. Amongst the models considered, the authors also estimated a close variation of the SK and CH models described in this paper. Their SK estimates found types $L_2$ to be most frequent, with types $L_3$ onwards nearly nonexistent. However, their CH estimated $\tau$ was found to be 5.17 which suggested that types $L_5$ were most frequent.

Another anomaly, pertains to the sharp differences in SK model's estimated proportion of types in the M(2012) and M(2013) sessions. The results suggest that the SK model might be extremely sensitive to the demographic pool of subjects being investigated. On the other hand, the estimated proportion of types between both Medium game sessions by the CH model was again found to be consistent (two-sided Fisher $\rho = 0.989$).

### Remark

We were also interested to investigate the influence of the $L_0$ type behavioral specification on the consistency of the CH model's estimates. Here that a $L_0$ type player uniformly randomizes across all strategies with probability $z \in [0, 1]$ or chooses 2000 with probability $(1 - z)$ - the above estimates were derived with $z = 0$. With the CH model, we estimated the respective sessions for $z = 0, 0.25, 0.50, 0.75, 1$. The CH model estimated proportion of $L_k$ types were found to be consistent between sessions of the same cohort for $z = 0, 0.25, 0.50$. When $z = 0.75, 1$, they were found to be significantly different.

## 6 Discussion

We began this paper by investigating if subjects' behavior in Arad and Rubinstein (2012) "11-20" game could be explained by the k-level process proposed by the authors. To do so, we replicated their "11-20" game in our Baseline design and proposed two other variations: the Medium and Extreme games. All games (Baseline, Medium and Extreme) have equivalent mixed-strategy equilibriums but are differentiated by their predicted behaviors for each $L_k$ types. Whilst subjects' behavior in our experimental sessions were significantly different from the mixed-strategy equilibrium, the k-level reasoning process predicts significant different proportion of $L_k$ types between sessions of the same cohort of subjects. This is contrary to the theoretical predictions of the k-level reasoning process given our experimental design and procedure, suggesting that the k-level reasoning process does not well explain the behavior of subjects in the respective sessions.[24]

However, when we allow for some flexible in the best responding behaviors of higher types as in the SK and CH model, we now observed consistent estimated proportion of $L_k$ types in all sessions of the same cohort for the CH model and to the lesser extend the SK model. Further support for the SK and CH models were found from the subjects' experimental feedback. Here 8.5%, 32%, 38% and 30% of the feedbacks from sessions B(2012), M(2012), M(2013) and E(2013) respectively were either empty or clearly corresponded to random behavior.[25] With the remaining feedbacks, the following two observations were made.

[24] The alternative view is that subjects' behavior were well explained by the k-level process. This view must therefore accept that small changes in any games will result in significantly different prediction of $L_k$ type proportions. We find this alternative unhelpful in extending the k-level reasoning process into wider economic settings.

[25] The feedbacks were independently evaluated by a graduate student.

**(i)** *Iterative thought-experiments anchoring on 2000.* Most subjects in session B(2012) described their behaviors as a consequence of an iterative process from 2000 ("I think that a lot of people will choose 1900 because it is 100 lower than the maximum amount. So I have gone for 1800, which is one step lower than that"). Similar descriptions are also observed in session M(2012) and M(2013) ("I hope that the other person will think that I have ignore the bonus and thus pick 1950. I therefore picked 1900"). In session E(2013), the descriptions are less straight forward, but nevertheless involve the discussion of the choice 2000.

**(ii)** *Subjects expect other subjects to best respond stochastically.* This is a prominent observation in sessions M(2012), M(2013) and E(2013) - to some extend in session B(2012). For example a atypical feedback in E(2013) session is as followed "Many people will expect others to choose 2000 and hence themselves choose 1975, 1950, 1925 or 1900. I therefore choose 1875 to get the bonus".

If subjects' feedback were truthful, their behavioral are consistent with decision process commonly attributed to those models of iterative reasoning.[26] It is however unclear if such behaviors were more closely associated with the SK or CH model. How is it therefore possible that two models (SK and CH) of iterative reasoning with substantially different assumptions on the $L_k$ types could explain the data equally well? We conjecture that the *true* model could include some mixture of the SK, CH model and quite possibly some "multiple types k-level process". For example, the Extreme game could be populated by three $L_1$ types. The first who always chooses 1975, the second who randomizes between 1975 to 1875, and the third who best respond stochastically.

To conclude, like the findings of Goeree, Louis, and Zhang (2013), we disagree with AR's k-level process for studying behavior in the "11-20" game. In addition, we find that some modification, such as the introduction of noise in the players behavior can allow experimenters to explain a wider variety of data. Our findings, suggest that the "11-20" game has the potential to discriminate among models of iterative thought-experiments and this will be a target for future research.

# References


Alaoui, L., and A. Penta (2013): "Endogenous Depth of Reasoning," Working Paper.

Arad, A., and A. Rubinstein (2012): "The 11-20 Money Request Game: A Level-k Reasoning Study," *American Economic Review*, 107(7), 3561–3573.

Bernheim, D. B. (1984): "Rationalizable Strategic Behavior," *Econometrica*, 52(4), 1007–1028.

Bosch-Domènech, A., J. G. Montalvo, R. Nagel, and A. Satorra (2002): "One, Two, (Three), Infinity, ...: Newspaper and Lab Beauty-Contest Experiments," *American Economic Review*, 92(5), 1687–1701.

Burchardi, K. B., and S. P. Penczynski (2010): "Out Of Your Mind: Estimating The Level-k Model," Working Paper.

Camerer, C. F., T.-H. Ho, and J.-K. Chong (2004): "A Cognitive Hierarchy Model of Games," *Quarterly Journal of Economics*, 119(3), 861–898.

Camerer, C. F., and D. Lovallo (1999): "Overconfidence and Excess Entry: An Experimental Approach," *American Economic Review*, 89(1), 306–318.


[26] Although were not incentivize to provide accurate feedbacks, we find little reasons for subjects to lie.

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

Clarke, K. A. (2003): "Nonparametric Model Discrimination in International Relations," *Journal of Conflict and Resolution*, 47(1), 72–93.

Costa-Gomes, M. A., and V. P. Crawford (2006): "Cognition and Behavior in Two-Person Guessing Games: An Experimental Study," *American Economic Review*, 96(5), 1737–1768.

Costa-Gomes, M. A., V. P. Crawford, and B. Broseta (2001): "Cognition and Behavior in Normal-Form Games: An Experimental Study," *Econometrica*, 69(5), 1193–1235.

Costa-Gomes, M. A., V. P. Crawford, and N. Iriberri (2009): "Comparing Models of Strategic Thinking in Van Huyck, Battalio, and Beil's Coordination Games," *Journal of the European Economic Association*, 7(2), 365–376.

Crawford, V. P., M. A. Costa-Gomes, and N. Iriberri (2013): "Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications," *Journal of Economic Literature*, 51(1), 5–62.

Crawford, V. P., and N. Iriberri (2007): "Level-k Auctions: Can a Nonequilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions?," *Econometrica*, 75(6), 1721–1770.

DellaVigna, S. (2009): "Psychology and Economics: Evidence from the Field," *Journal of Economic Literature*, 47(2), 315–372.

Duffy, J., and R. Nagel (1997): "On the Robustness of Behaviour in Experimental 'Beauty Contest' Games," *Economic Journal*, 107(445), 1684–1700.

Goeree, J. K., and C. A. Holt (2004): "A Model of Noisy Introspection," *Games and Economic Behavior*, 46(2), 365–82.

Goeree, J. K., C. A. Holt, and T. R. Palfrey (2005): "Regular Quantal Response Equilibrium," *Experimental Economics*, 8(4), 347–67.

Goeree, J. K., P. Louis, and J. Zhang (2013): "Noisy Introspection in the "11-20" Game," Working Paper.

Grosskopf, B., and R. Nagel (2008): "The Two-person Beauty Contest," *Games and Economic Behavior*, 62(1), 93–99.

Harrison, G. W., and E. E. Rutström (2009): "Expected Utility Theory and Prospect Theory: One Wedding and a Decent Funeral," *Experimental Economics*, 12(2), 133–158.

Kawagoe, T., and H. Takizawa (2012): "Level-k Analysis of Experimental Centipede Games," *Journal of Economic Behavior and Organization*, 82(2), 548–566.

Kocher, M. G., and M. Sutter (2005): "The Decision Maker Matters: Individual Versus Group Behaviour in Experimental Beauty-Contest Games," *Economic Journal*, 115(500), 200–223.

Lindner, F., and M. Sutter (2013): "Level-k Reasoning and Time pressure in the 11-20 Money Request Game," *Economic Letters*, 120(3), 542–545.

McFadden, D. L. (1976): "Quantal Choice Analysis: A Survey," *Annals of Economic and Social Measurement*, 5(4), 363–390.

McKelvey, R. D., and T. R. Palfrey (1995): "Quantal Response Equilibria for Normal Form Games," *Games and Economic Behavior*, 10(1), 6–38.

——— (1996): "A Statistical Theory of Equilibrium in Games," *Japanese Economic Review*, 47(2), 186–209.

——— (1998): "Quantal Response Equilibria for Extensive Form Games," *Experimental Economics*, 1(1), 9–41.

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

Nelder, J. A., and R. Mead (1965): "A Simplex Method for Function Minimization," *The Computer Journal*, 7(4), 308–313.

Östling, R., J. T. Wang, E. Y. Chou, and C. F. Camerer (2011): "Testing Game Theory in the Field: Swedish LUPI Lottery Games," *American Economic Journal: Microeconomics*, 3(3), 1–33.

Pearce, D. G. (1984): "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica*, 52(4), 1029–1050.

Przyborowski, J., and H. Wilenski (1940): "Homogeneity of Results in Testing Samples from Poisson Series: With an Application to Testing Clover Seed for Dodder," *Biometrika*, 31(3), 313–323.

Rogers, B. W., T. R. Palfrey, and C. F. Camerer (2009): "Heterogeneous Quantal Response Equilibrium and Cognitive Hierarchies," *Journal of Economic Theory*, 144(4), 1440–1467.

Rosenthal, R. W. (1981): "Games of Perfect Information, Predatory Pricing, and the Chain Store Paradox," *Journal of Economic Theory*, 25, 92–100.

Sheskin, D. J. (2003): *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman and Hall.

Stahl, D. O., and P. W. Wilson (1994): "Experimental Evidence on Players' Models of Other Experimental Evidence on Players' Model of Other Players," *Journal of Economic Behavior and Organization*, 25(3), 309–327.

——— (1995): "Models of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior*, 10(1), 218–254.

Vuong, Q. H. (1989): "Likelihood Ratio Tests for Model Selection and non-nested Hypotheses," *Econometrica*, 57(2), 307–333.

Weber, R. A. (2001): "Behavior and Learning in the "Dirty Faces" Game," *Experimental Economics*, 4(3), 229–242.

——— (2003): "'Learning' with No Feedback in a Competitive Guessing Game," *Games and Economic Behavior*, 44(1), 134–144.

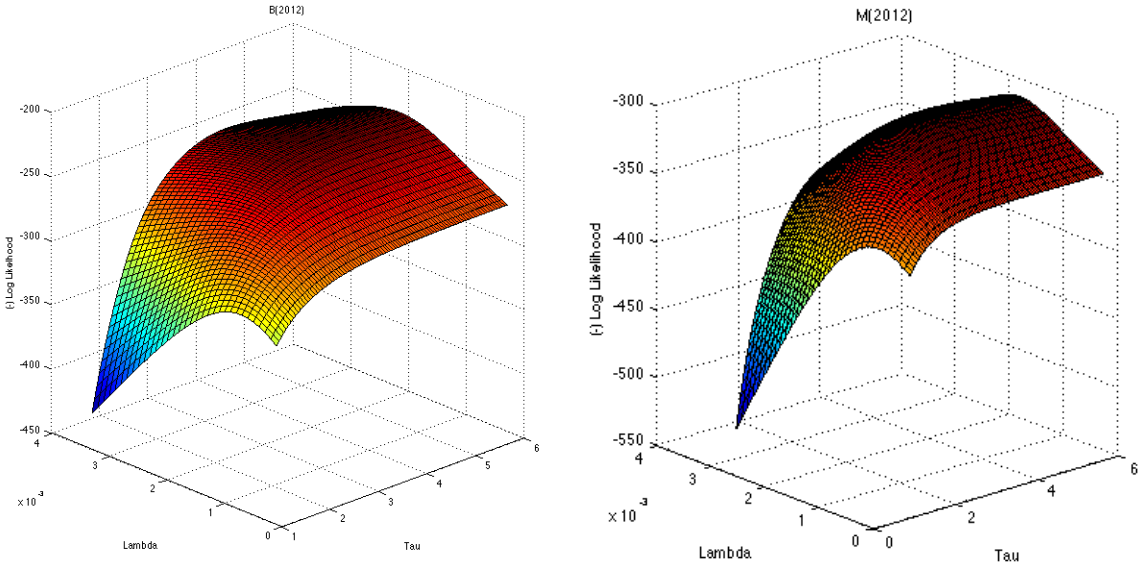


Figure 2: CH Model Log-Likelihood Functions for Sessions B(2012) and M(2012)

# Appendix

### Estimating the Cognitive Hierarchy Model

The model was estimated using the maximum likelihood techniques. Let $p(a)$ denote the probability of observing action $a \in A$ in the game and $y_i$ the $i = 1, 2, ..., N$ observation. Given the model's construct, one is able to rewrite

$$p(a|\tau, \lambda) = p^0 f(0|\tau) \prod_{k=1}^{\bar{K}} p^k(a|\lambda, \tau) f(k|\tau)$$

where $p^k(a)$ is the probability of a $L_k$ type choosing $a \in A$, $\bar{K}$ refers to the highest $L_{\bar{K}}$ type in the data and $f(k|\tau) \in [0, 1]$ is the relative proportion of $L_k$ type given the Poisson assumptions that $f(k|\tau) = \tau^k \exp(-\tau)/k!$. The log-likelihood function is therefore

$$log(L) = \sum_{i=1}^{N} log(p(y_i|\lambda, \tau))$$

which was optimized given the constraints $\sum_{k=0}^{\bar{K}} f(k|\tau) > 1 - \varepsilon$, where $\varepsilon = 0.001$, and the boundary conditions $\tau \in [0, \bar{K}]$ and $\lambda \in [0, 100]$. We were uncertain if the log-likelihood function was concave or kinked and thus employed the direct search, Nelder and Mead (1965) optimization technique. Cautious of such approach, we explored a fine search termination criteria of 0.0000001 and checked if our estimates ($\tau$ and $\lambda$) were robust for $\bar{K} = 6, 9, 20$. The estimates were to be robust and the log-likelihood function was observed to be concave (see Figure 2), which suggest that our estimates were indeed the global maximum.

**Estimating the Sk-level Model**

The maximum likelihood technique involves $\bar{K}+1$ free parameters. We hence expressed $p(a)$ as

$$p(a|\alpha_0, \alpha_1, ..., \alpha_{\bar{K}}, \lambda) = p^0\alpha_0 \prod_{k=1}^{\bar{K}} p^k(a|\lambda, \tau)\alpha_k$$

where $\alpha_k \in [0,1]$ denotes the proportion of $L_k$ types in the data, given the constraints that $\alpha_{\bar{K}} = 1 - \alpha_0 - \alpha_1 - ... - \alpha_{\bar{K}-1}$. We again employed the same estimation techniques as in the CH model. To ensure that our estimates are the global maximum, we considered multiple random starting values for the parameters $\alpha_0, \alpha_1, ..., \alpha_{\bar{K}-1}$. Given this criteria, we repeated the estimation process 10 times for each session and the estimates were found to be identical each time. This suggest that our estimates are also the global maximum.