

Consistent Depth of Reasoning in Level-k Models

David J. Cooper

Florida State University and University of East Anglia

Enrique Fatas

University of East Anglia

Antonio J. Morales

Universidad de Málaga

Shi Qi

College of William and Mary

Abstract: The level-k model is often implemented with an assumption that individuals employ a fixed depth of reasoning across different games. To study the validity of this assumption, we have subjects make choices in a series of games designed to identify inconsistent depth of reasoning without relying on the results of an econometric model. Most subjects' choices are *not* consistent with them having a fixed depth of reasoning even in extremely closely related games. Econometric analysis verifies that this result is quite robust and illustrates the nature of the inconsistency. The likelihood of inconsistency increases with cognitive ability, suggesting that it is not solely due to confusion. Higher optimization premiums are correlated with greater depth of reasoning, but do not reduce the likelihood of inconsistency per se. We argue that depth of reasoning, like many other varieties of individual choice, is subject to stochastic choice.

“A foolish consistency is the hobgoblin of little minds ...”

Ralph Waldo Emerson, 1841

1. Introduction: Since the path-breaking work of Nagel (1995), it is well established that limited depth of reasoning accounts for important features of experimental data missed by models based on full rationality. The level-k model, with its intuitive and tractable structure, has emerged as the most commonly used model of limited depth of reasoning.¹ The model is based on a hierarchy of levels. Level-0 individuals make decisions in a fashion that is not based on strategic considerations. Level-1 individuals best respond (possibly with noise) to the distribution of choices by level-0 individuals, level-2 individuals best respond to the distribution of choices by level-1 individuals, etc. Many papers have attempted to identify the distribution of levels (e.g. Stahl and Wilson, 1995; Costa-Gomes, Crawford, and Broseta, 2001; Costa-Gomes and Crawford, 2006) or used level-k models to explain behavior in a variety of settings (e.g. Crawford and Iriberri, 2007; Arad and Rubinstein, 2012; Östling, Wang, Chou, and Camerer, 2011).²

Several recent papers have pointed out flaws in the level-k approach.³ The results of Georganas, Healy, and Weber (2015) are particularly germane for the work reported below. They econometrically estimate each subject’s level (level 1, level 2, etc.) for two classes of games, undercutting games and guessing games. Only 27% of subjects have the same estimated level for both classes. They find positive evidence for consistent levels within the class of undercutting games but not within the class of the guessing games. The results within classes of games are less convincing than the results between classes due to data limitations.⁴ One potential explanation for the observed lack of consistency is that depth of reasoning can reflect beliefs about *others’* limited depth of reasoning rather than limits to one’s own depth of reasoning (Agranov, Potamites, Schotter, and Tergiman, 2012). To the extent that different classes of games trigger different beliefs about others’ depth of reasoning, changes in levels across classes of games will occur.

¹ Other models that incorporate limited depth of reasoning include Cognitive Hierarchies (Camerer, Ho, and Chong, 2004), Noisy Introspection (Goeree and Holt, 2004), and sophisticated EWA (Camerer, Ho, and Chong, 2002).

² The literature on level-k models is far too large for us to list all of the papers that have used this approach. For a recent summary of the literature, see Crawford, Costa-Gomes, and Iriberri (2013).

³ Costa-Gomes and Wiezsacker (2008) and Ivanov, Levin, and Niederle (2010) demonstrate that subjects’ choices are not consistent with best-responding to beliefs. Hargreaves Heap, Rojo Arjona, and Sugden (2014) show that level-0 behavior responds to the strategic features of games when it should not.

⁴ Estimated types for within-class comparisons are based on a single observation and generated using an assignment rule. See their fn. 24 for discussion of this issue.

Another possibility is that stochastic choice creates the appearance of inconsistency. If subjects are only capable of noisy optimization, their estimated levels based on observed choices can be inconsistent even though the same depth of reasoning is used throughout.⁵

Our goal is to show that inconsistent depth of reasoning is a pervasive phenomenon that cannot easily be explained away. We use an experimental design that does not solely rely on econometric estimation to identify inconsistency and provide strong evidence that inconsistency is common both between *and within* classes of games. The latter point is particularly important, since there is no obvious reason why beliefs about others' levels should vary much between such closely related games. Fitting structural econometric models, we show that the inconsistencies cannot be explained by stochastic choice and are robust to a wide variety of different model specifications. Although the level-k model with consistent depth of reasoning does poorly in predicting individual behaviors, we show that it does well at predicting aggregate behaviors out of sample. The level-k model remains a valuable tool for understanding aggregate behaviors, but must be used carefully given the pervasive inconsistency in depth of reasoning. Depth of reasoning, like many other varieties of individual choice, appears to be inherently stochastic.

Going into detail, subjects in our experiments make a series of choices in 2-player games drawn from five oft-studied classes of games. Each class consists of four games that systematically vary the two players' payoffs. The first three classes (imperfect price competition, minimum effort, and traveler's dilemma) yield strong predictions about how an individual who uses a consistent depth of reasoning (i.e. consistently being level-0, level-1, level-2, etc.) responds to changes in his own and his rival's payoffs. Namely, a level-0 individual does not respond to changes in either player's payoffs; a level-1 individual responds to changes in his own payoff but not to changes in his rival's payoffs; and a level-2 individuals responds to changes in his rival's payoffs but not in his own payoffs. Higher levels display the same alternating pattern.⁶ If subject's levels are *consistent within a class of games*, we should observe a predictable pattern of changing decisions between different games of the same class.

⁵ See also Kline (2017). This paper is primarily concerned with the econometrics of estimating models of strategic reasoning with heterogeneous types, but includes an estimation exercise using guessing game data from Costa-Gomes and Crawford (2006). Level-k types are a subset of the decision rules considered by the model. The estimation exercise finds that the most common type within the population uses multiple decision rules, primarily different levels of "unanchored reasoning" which is closely related to rationalizability. The distinction between anchored and unanchored reasoning is not crucial for our non-econometric analysis of consistency.

⁶ Specifically, level-3 types only respond to changes in their own payoffs, level-4 types only respond to changes in their rival's payoffs, etc.

We find little evidence of consistent levels *within classes of games*, and even less evidence *between classes of games*. A subject is defined as “strongly consistent” with level-1 within a class of games if his responses to changes in his *own* payoffs are consistent with being level-1 and his responses to changes in his *rival’s* payoffs are *not* consistent with being level-2. A subject is defined as “strongly consistent” with level-2 in an analogous manner. Only about 20% of subjects are strongly consistent with either level-1 or level-2 for a given class of games. Virtually no subjects are strongly consistent with the *same* level for the first three classes of games.

The fourth class of games, Arad and Rubinstein (2012) “11 – 20” game, is included to make a simple point. Individual behavior is not consistent with subjects possessing a fixed depth of reasoning, but *aggregate* behavior is in line with predictions by the level-k model. The 11-20 game was designed to give a specific pattern of choices that are consistent with the level-k model rather than Nash equilibrium. Our 11-20 data closely resembles Arad and Rubinstein’s data, and is consistent with predicted pattern from the level-k model.

The fifth and final class of games, all-pay auctions, features a large number of dominated strategies. This class of games was always played at the end of the experiment (when subjects were tired and presumably most likely to make random errors) to test whether subject’s choices were consistent with a minimal level of rationality. Subjects rarely played dominated strategies in the all-pay auctions. It is unlikely that the pervasive inconsistency observed in the first three classes of games can be attributed to confusion or arbitrary mistakes.

A strength of our approach is that we do not rely on an econometric fitting exercise to identify inconsistency. However, the observed inconsistencies could reflect stochastic choice (i.e. noisy best responses) rather than a lack of consistent depth of reasoning. We address this issue by fitting several structural econometric models. Our baseline model allows for three “consistent” types (level-0, level-1, and level-2) and two “inconsistent” types that mix across the three levels. A “pure-mixing” type randomly draws a level of reasoning (0, 1, or 2) for each game with the mixing probabilities fit from the data. A “semi-mixing” type is identical to a “pure-mixing” type, except that, rather than drawing a new level of reasoning for every game, a semi-mixing type draws a new level for every class of game but uses the same level within a class. Subjects are assumed to use a noisy best response to their beliefs, allowing for stochastic choice. The baseline model assigns 89% of the population to the two mixing types, with 43% classified as the pure-mixing type and 46% as the semi-mixing type. The structural model identifies more consistency than our

non-econometric approach, but it remains true that the vast majority of subjects display an inconsistent depth of reasoning. This finding is robust across a wide variety of alternative model specifications.

Three additional results from the econometric exercise are worth noting. First, we show that behavior is sensitive to cognitive ability as measured by scores on a Raven's Progressive Matrix (RPM) test. This is not due to a change in the likelihood of being an inconsistent type. Rather, the probability of choosing a higher level, *subject to mixing*, is an increasing function of the RPM score. Our result mirrors that of Gill and Prowse (2016). Making the reasonable assumption that subject confusion is a decreasing function of cognitive ability, this result provides additional evidence that the pervasive inconsistency we observe cannot be attributed to subject confusion.

Second, we find that behavior is sensitive to the optimization premium (the increased payoff from using a greater depth of reasoning), consistent with the work of Alaoui and Penta (2016). *Subject to mixing*, the probability of choosing higher levels is increasing in the optimization premium. This does not change our conclusion that most subjects do not use a consistent depth of reasoning, but indicates that a coherent pattern underlies their inconsistency.

Finally, we find that not accounting for inconsistent depth of reasoning cause problems when attempting to estimate the distribution of levels (i.e. level-0, level-1, level-2, etc.). A typical approach has subjects play a large number of games without feedback and then fits an econometric model to estimate the distribution of levels *assuming that individuals use a consistent depth of reasoning across all games*. If subjects do *not* use a consistent depth of reasoning, estimation methods that assume consistency will confound mixing between levels with noisy optimization. A comparison of our baseline model with a model that only includes the three consistent types confirms this intuition – forcing consistency causes a modest shift in the realized distribution of levels (i.e. the distribution of levels *after* mixing has occurred) toward higher levels and a large increase in the estimated amount of noise in subjects' choices.

The assumption that individuals use a consistent depth of reasoning is *not* an essential component of the level-k model. For many applications, it is sufficient that the model can predict the *aggregate* distribution of choices. We examine the predictive ability of the model by fitting our baseline model to four classes of games and then simulating data for the fifth class. The model does well at predicting aggregate behavior in the fifth class. This reflects a basic feature of the

level-k model: the ability to predict out of sample relies on the *distribution* of levels being stable across classes of games, not on individuals having a consistent depth of reasoning. The differing classes of games we study are sufficiently similar that the distribution of levels changes little.

Our primary contribution is demonstrating the pervasive presence of inconsistent depth of reasoning. The inconsistency occurs even between very closely related games and cannot be attributed to details of the model's specification, subject confusion, or arbitrary choices.

The method we use for identifying inconsistencies is also an important contribution of our paper. We make heavy use of econometric modeling to verify and extend the main finding, but our experimental design makes it possible to identify inconsistencies without relying solely on econometrics. We view the two approaches as complements – we are more confident about the pervasiveness of inconsistent depth of reasoning because this finding is corroborated by both approaches.

The level-k model remains a valuable tool whether or not individuals employ a consistent depth of reasoning – our prediction exercise should make this point clear. Features of how mixing occurs, such as the correlation between cognitive ability and the likelihood of mixing or the correlation between optimization premiums and the weight on higher levels, suggest a coherent reasoning process underlies the use of inconsistent depth of reasoning. Economists have become comfortable with stochastic choice in individual choice (e.g. Agranov and Ortoleva, 2017), and choosing one's depth of reasoning for a game is simply another example of individual choice. Rather than viewing inconsistent depth of reasoning as a flaw in the level-k model, we hope theorists and experimenters become comfortable with stochastic choice between different heuristics for thinking about games and devote their efforts to further exploration of what heuristics are being used and how individuals choose between them from game to game.

2. Experimental Design and Procedures: Subjects make choices in five classes of 2-player games, with four games in each class, yielding a total of 20 games. Each experimental subject made 20 decisions, one for each game, without feedback. This section introduces the five classes of games, discusses predictions for these games, and describes the experimental procedures.

2.1. The Classes of Games: In all 20 games, two players simultaneously choose actions from the discrete set $X = \{110, 120, 130, \dots, 200\}$. Let $x_1 \in X$ and $x_2 \in X$ denote the actions chosen by

Players 1 and 2 respectively. Within each class, $C \in \{1,2,3,4,5\}$, the players' payoffs π_1^C and π_2^C are functions of the actions x_1 and x_2 , conditioned on two payoff parameters, α_1 and α_2 . In each class of games $C \in \{1,2,3,4,5\}$, a game $G^C(\alpha_1, \alpha_2)$ is defined by the payoff parameters α_1 and α_2 . In our experimental design the payoff parameters can take on either a high or a low values, specifically 20 or 80. Within each class $C \in \{1,2,3,4,5\}$ we consider four games, generated by systematically varying the values of the payoff parameters: $G^C(20,20)$, $G^C(20,80)$, $G^C(80,20)$ and $G^C(80,80)$. Each class includes two symmetric games and two asymmetric games.

Subject to relabeling, the payoff functions are identical for the two players: If $x_1 = a$, $x_2 = b$, $\alpha_1 = c$, and $\alpha_2 = d$, then $\pi_1^C(a, b|c, d) = \pi_2^C(b, a|d, c)$. Given that the payoff functions are basically identical for the two roles, Player 1 and Player 2, there is usually no need to distinguish between roles. We therefore use the following notation which refers to a “generic” player in either role: **π^C , x_i and α_i refers to a player's own payoff function, own action, and own payoff parameter respectively, and notation x_j and α_j refers to his rival's action and payoff parameter respectively.** The payoff functions for all five classes are constructed such that a player's payoff is a function of x_i , x_j , and α_i , but not α_j . In other words, changing α_i changes a player's own payoffs, but not his rival's payoffs, holding both players' actions fixed. The payoff parameters are common knowledge. Given that payoff functions are identical for both roles, subject to relabeling, all subjects face the same four decisions in each class regardless of role.

All five classes are based on games previously studied in the experimental literature. A brief introduction for each class follows.

Class 1: Imperfect Price Competition (Capra et al, 2002): The two players simultaneously choose prices. A player's payoff equals his price if he submits the lower of the two prices. His payoff is a proportion of his rival's price if he submits the higher price, with the proportion equal to $\alpha_i/100$. In case of a tie, the player is paid his expected payoff based on a 50/50 chance of being considered the low price. The resulting payoff function is given by (1).

$$(1) \pi^1(x_i, x_j | \alpha_i) = \begin{cases} x_i & \text{if } x_i < x_j \\ \frac{100 + \alpha_i}{200} x_i & \text{if } x_i = x_j \\ \frac{\alpha_i}{100} x_j & \text{if } x_i > x_j \end{cases}$$

If $\alpha_i < 83\frac{1}{3}$, a player's best response to their rival's choice x_j is choosing $x_j - 10$ if $x_j > 110$ and choosing 110 if $x_j = 110$. For all values of α_1 and α_2 used in our experiment, the unique Nash equilibrium is for both players to choose 110.

Class 2: Minimum Coordination Game (Goeree and Holt, 2005): The two players simultaneously choose effort levels. Each player earns the minimum of the two effort levels minus a proportion of her chosen effort level, with the proportion equal to $\alpha_i/100$. The payoff function is given by (2).

$$(2) \pi^2(x_i, x_j | \alpha_i) = \min\{x_i, x_j\} - \frac{\alpha_i}{100} x_i$$

If $\alpha_i < 100$, a player's best response to x_j is choosing x_j . Hence, all symmetric pairs ($x_1 = x_2$) are Nash equilibria of the game for all values of α_1 and α_2 used in our experiment.

Class 3: Travelers' Dilemma (Capra et al, 1999): The two players simultaneously choose claims. Each player earns the minimum of the two claims and an additional quantity, equal to α_i , is added (subtracted) if hers is (not) the minimum claim. In case of a tie, there is no additional quantity to be paid/received. The payoff function is given by (3).

$$(3) \pi^3(x_i, x_j | \alpha_i) = \begin{cases} x_i + \alpha_i & \text{if } x_i < x_j \\ x_i & \text{if } x_i = x_j \\ x_j - \alpha_i & \text{if } x_i > x_j \end{cases}$$

When $\alpha_i > 10$, a player's best response to x_j is choosing $x_j - 10$ if $x_j > 110$ and choosing 110 if $x_j = 110$. Hence, mutual choice of 110 is the unique Nash equilibrium of the game for all values of α_1 and α_2 used in our experiment.

Class 4: The "11-20" Game (Arad and Rubinstein, 2012): The two players simultaneously choose numbers. Each player receives her chosen number plus an additional quantity, equal to α_i , if her chosen number is exactly 10 below her rival's chosen number. The payoff function is given by (4).

$$(4) \pi^4(x_i, x_j | \alpha_i) = \begin{cases} x_i + \alpha_i & \text{if } x_i + 10 = x_j \\ x_i & \text{otherwise} \end{cases}$$

When $\alpha_i > 10$, a player's best response to x_j is choosing $x_j - 10$ if $\alpha_i > 210 - x_j$ and choosing 200 otherwise. This game has no pure strategy Nash equilibrium.

Class 5: All-pay Auction (Gneezy and Smorodinski, 2006): The two players simultaneously choose bids. Each player gets 110 minus her bid and the high bidder receives an amount, equal to α_i . In case of a tie, each player wins with probability one half and is paid the expected payoff. The payoff function is shown in (5).

$$(5) \pi^5(x_i, x_j | \alpha_i) = \begin{cases} 110 - x_i + \alpha_i & \text{if } x_i > x_j \\ 110 - x_i + \frac{\alpha_i}{2} & \text{if } x_i = x_j \\ 110 - x_i & \text{if } x_i < x_j \end{cases}$$

For $\alpha_i = 80$, the best response to $x_j < 180$ is $x_j + 10$ and 110 otherwise.⁷ For $\alpha_i = 20$, a player is indifferent between choosing 110 and 120 if $x_j = 110$ and is indifferent between choosing 110, 120, and 130 if $x_j = 120$. Otherwise, choosing 110 is a strict best response. The game $G^5(20,20)$ has four weak Nash equilibria, two symmetric equilibria with mutual choice of 110 or 120 and two asymmetric equilibria where one player chooses 110 and the other chooses 120. Given that 110 weakly dominates 120, the equilibrium where both players choose 110 is the most plausible. For either asymmetric game, the game has a unique weak Nash equilibrium where the player with the low value of α chooses 110 and the player with the high value of α chooses 120. The game $G^5(80,80)$ has no pure strategy Nash equilibrium.

2.2 Theoretical Predictions: A central feature of the level-k model is that individuals who use level-0 reasoning are non-strategic. They may not follow a uniform distribution as is often assumed, but their distribution of actions cannot be rationalized as a best response to some beliefs

⁷ For $x_j = 180$, Player i is indifferent between choosing 190 and 110.

about their rivals' behavior. This implies that the distribution of play for level-0 individuals is invariant to changes in the payoff parameters. For all five classes of games, a player's payoff is a function of their own payoff parameter α_i but *not* of their rival's payoff function parameter α_j . Together, the preceding observations imply that the expected payoff function for a level-1 individual takes on the following form where $p(x_j)$ is the pdf of actions for his rival.

$$(6) \quad E\pi^c(x_i|\alpha_i) = \sum_{x_j=110}^{200} \pi^c(x_i, x_j|\alpha_i)p(x_j)$$

This expected payoff function is a function of α_i , but not α_j . Observation 1 follows.

Observation 1: A level-1 individual will react to changes in their own payoff parameter (α_i) but not to changes in their rival's payoff parameter (α_j).

A level-2 individual best responds to the choice of a level-1 individual. For simplicity, assume that level-1 individuals best respond *without noise* to level-0 individuals. For Classes 1 – 3, a player's best response function does not depend on α_i for the range of α_i used in our experiment. Since the choices of a level-1 individual only responds to his own payoff parameter, it follows that a level-2 individual's choices respond to changes in her rival's payoff parameter (α_j), but not to changes in her own payoff parameter (α_i).

Observation 2: For Classes 1 – 3, a level-2 player will react to changes in her rival's payoff (α_j) parameter but not to changes in her own payoff parameter (α_i).

For Classes 1 – 3, level-3 individuals will respond to changes in α_i , but not to changes in α_j . This follows from Observation 2. A level-3 individual best responds to a level-2 individual. A level-2 individual only responds to changes in her rival's payoffs, which are own payoffs from the point of view of the level 3 individual. Similar logic dictates that level-4 individuals will respond to changes in α_j , but not to changes in α_i , level-5 individuals will respond to changes in α_i , but not to changes in α_j , and so forth.

Observations 1 and 2, along with their extension to level-3 and higher, imply that data from Classes 1 – 3 can be used to detect consistency without relying solely on econometric analysis. If subjects have consistent depth of reasoning throughout the experiment, the following prediction applies to all subjects who are level-1 or higher. Given that we expect few level-0 individuals, Prediction 1 should apply to the vast majority of subjects.

Prediction 1: Subjects should only respond to changes in their own payoff parameter or should respond only to changes in their rival's payoff parameter. They should not respond to changes in both payoff parameters.

The values of 20 and 80 for the payoff parameters are chosen to generate large responses to changing the payoff parameters. To give a sense of the likely magnitude of responses, level- k predictions for the different classes of games are displayed in Table 1. The values of the payoff parameter of Player 1 are given by the rows and those of Player 2 by the columns. These predictions are from the point of view of a Player 1, assuming that choices of level-0 individuals are distributed uniformly while choices for level-1 and level-2 individuals are best responses without noise. Predictions with more than one number (e.g. all cells for Class 2) reflect indifference between two actions. For Classes 1 – 3, a level-1 (level-2) subject is predicted to respond strongly to a change in their own (rival's) payoff parameter and not respond at all to a change in their rival's (own) payoff parameter.

[Insert Table 1]

Classes 4 and 5 are less useful than Classes 1 – 3 for detecting consistency, but are included in the experimental design for other reasons. Observation 2 does not hold for Classes 4 and 5,⁸ and in Table 1 we see that the predicted shifts are only weakly consistent with Observations 1 and 2. In practice we predict no shifts in response to changing payoff parameters for Classes 4 and 5.

Class 4 (11 – 20 games) is useful for two reasons. First, the game with high symmetric payoff parameters $G(80,80)$ closely resembles the original Arad and Rubinstein version in that higher levels choose smaller numbers in an ordered way (up to level 8). We use data from this game to confirm that our subjects' behavior looks similar *on aggregate* to what has been observed for a canonical game in the level- k literature. The failure of our subjects to exhibit consistent depth of reasoning is not due to behavior that is wholly inconsistent with the basic patterns of play predicted by level- k models and observed in earlier research.

⁸ It is trivial to construct examples for Classes 4 and 5 where a level-2 individual responds to changes in her own payoff parameter. For Class 4, suppose all level-0 individuals choose 160. Fix $\alpha_j = 80$, implying that a level-2 individual best responds to a choice of 150. If $\alpha_i = 20$, the best response is 200. If $\alpha_i = 80$, the best response is 140. For Class 5, suppose all level-0 individuals choose 140. Fix $\alpha_j = 80$, implying that a level-2 individual best responds to a choice of 150. If $\alpha_i = 20$, the best response is 110. If $\alpha_i = 80$, the best response is 160.

Second, when subjects have a low payoff parameter ($\alpha_i = 20$), choices smaller than 180 are strictly dominated by the choice of 200. Class 5, the all-pay auction, serves a similar purpose since choices greater than 130 were strictly dominated for $\alpha_i = 20$. Experimental subjects had already taken 12 choices in 12 different environments when they reached Classes 4 and 5. If they behaved randomly due to boredom, fatigue, or lack of salience, this should be reflected in frequent play of the dominated strategies in Classes 4 and 5. Rare play of dominated strategies suggests that inconsistency cannot be attributed to these causes.⁹

Prediction 2: Subjects will not use dominated strategies in Classes 4 and 5.

2.3 Experimental Procedures: All sessions were run at LINEEX at the University of Valencia in 2014 and 2015. The subjects were undergraduate students with no previous exposure to experiments with any of the five classes of games.

At the beginning of the experiment, experimental subjects were randomly allocated to one of two possible roles, Player 1 or Player 2. Roles were kept constant along the whole duration of the experiment. Table 2 summarizes the sessions that were conducted.

[Insert Table 2 here]

The experiments were run using paper and pencil. After experimental subjects were seated and types were allocated, subjects were given an initial set of general instructions (see Appendix A). We read all instructions aloud as well providing subjects with printed copies. The general instructions emphasize how to read the payoff matrices, but also explained how role assignment would be done, how pairings would work, and how payment would be made.

Following the general instructions, experimental subjects faced the five classes of games sequentially. The order of Classes 1 – 3 was rotated across sessions, but Classes 4 and 5 were always the last two classes. This was done to increase any possible effects of fatigue or boredom in Classes 4 or 5. A separate packet was handed out for each class. Each packet had a set of instructions along with copies of the payoff matrices for the four games. The payoff tables show the payoffs for both roles, maintaining common knowledge of payoffs.

The packet instructions included a brief recapitulation of the general instructions for the experiment and a detailed explanation of the game being played with an emphasis on

⁹ Some use of dominated strategies is expected in a level-k framework due to level-0 individuals.

understanding the payoffs. The packet instructions stressed that the four games being played within a class were *not* the same. For instructions after the first class of games, it was also stressed that the games being played changed from packet to packet. Beyond going through the mechanics of reading the payoff table, the packet instructions gave a brief intuitive explanation of the structure of games in the class. For example, the instructions for the minimum game stated, "... the two participants receive the smaller number [of the two chosen], minus a percentage (20% or 80%) of the number they have chosen."

Experimental subjects were asked to make decisions in all four games of a given class at the same time, and they could fill the decision sheet out for the four games in any order they wished. Once decisions for a given class were done, papers were collected and the packets for the next class were handed out, so subjects could not go back (nor forward) to a different class of games. At no point did subjects receive feedback about others' choices or outcomes of the games. Each payoff matrix was printed on a single sheet of paper. Using paper and pencil rather than computerizing the experiment was intended to make it as easy as possible for subjects to compare payoff tables within a class or go back to the instructions.

[Insert Figure 1 here]

After all twenty games had been played, subjects took a 15 item version of Raven's progressive matrices (RPM) test as a measure of cognitive ability. This was computerized, using z-tree (Fischbacher, 2007), rather than run by hand. Each item showed subjects a 3x3 matrix of geometric figures (see Figure 1 for an example). They were asked to deduce what figure was needed to complete the sequence from a menu of eight possibilities. Subjects were given thirty seconds to complete each question and were paid 0.25 euros for each item completed correctly. The median score was 12 out of 15 items. The RPM test is a well-known instrument for testing reasoning ability. Gill and Prowse (2016) show a positive relationship between RPM test scores and depth of reasoning in a level-k model.¹⁰ We administered the abbreviated RPM test to study whether there is a relationship between cognitive ability and consistency.

¹⁰ Gill and Prowse use a sixty-question version of the RPM test taken before the games. We use a shortened version administered after the games. This reflects the differing goals of the two papers – we are primarily interested in consistency and wanted to eliminate any possibility that the RPM test could affect behavior in the games.

At the end of the experiment, subjects in the Player 1 role were randomly matched with subjects in the Player 2 role. They were paid based on their choices for one randomly chosen game out of the twenty. We paid on one randomly chosen game to avoid any possibility of hedging. The payoff tables were denominated in ECU, with a conversion rate of 10 ECU = 1 euro. The average duration of a session was around 90 minutes and the average payoff was about 18 – 20 euros, including a 5 euro show-up fee.

3. Experimental Results: This section begins by confirming that our data is consistent with previous experiments studying these five classes of games and with the level-k model predictions at the *aggregate* level. We then show that the individual data is largely consistent with Prediction 2, but not Prediction 1. The latter implies that subjects do not employ a consistent depth of reasoning, a finding that we confirm with formal econometric analysis in Section 4.

3.1 Aggregate results: Table 3 displays the average choices for all twenty games, sorted by class. The layout parallels Table 1, with the values of a subjects' own payoff parameter (α_i) given by the rows and those of his rival (α_j) by the columns. See Appendix B for a more detailed breakdown of subjects' choices by class of game and payoff parameters within class.

[Insert Table 3 here]

Our aggregate data has the same basic patterns as previous studies using the same classes of games. We drew the Imperfect Price Competition game from Capra *et al.* (2002). They study symmetric versions of the game, comparing behavior with high and low payoff parameters. Even though changing the payoff parameter does not affect the Nash equilibrium, Capra *et al.* find that higher values of the payoff parameter lead to higher choices (prices). Comparing the top left and bottom right corners for Class 1 in Table 3, the same pattern is observed as the distribution of choices shifts to the right with the higher value of α .

Goeree and Holt (2005) use the minimum effort game to make a similar point. They study symmetric versions of the game, comparing behavior with high and low payoff parameters. Changing the payoff parameter does not affect the set of Nash equilibrium, but Goeree and Holt find that higher values of the payoff parameter (costs) lead to lower choices (efforts). Comparing the top left and bottom right corners for Class 2 in Table 3, the same pattern is observed as the distribution of choices shifts to the left with the higher value of α .

The story is similar for Class 3, the Traveler’s Dilemma. Capra *et al.* (1999) study how choices (claims) depend on the reward parameter in symmetric versions of the game. They find that higher values of the payoff parameter lead to lower choices. The same pattern is seen in our data if the top left and bottom right corners are compared for Class 2 in Table 3. When α is increased for the symmetric game, the distribution of choices shifts to the left.

[Insert Figure 2 here]

In the “11-20” game, as introduced by Arad and Rubinstein’s (2012), subjects chose integers in the interval [11,20] with a reward of 20. Arad and Rubinstein find that more than 80% of chosen numbers were 17 or larger, meaning that experimental subjects are at most level-3. Figure 2 displays the distribution of choices (scaled by a factor of 10) from Arad and Rubinstein (2012), together with our symmetric 11-20 games with high and low rewards. Our 11-20 game with high rewards is the most similar to Arad and Rubinstein’s, albeit with a lower reward (80 vs 200, scaled). The data from our game with high rewards is shifted to the right relative to Arad and Rubinstein’s data, reflecting the lower reward, but like them we see fewer choices of 200 (equivalent to their 20) than 190 or 180 and rare choice of numbers consistent with more than level-3 reasoning – only 3% of our observations are smaller than 170.

None of the preceding speaks to the issue of consistency. Rather, the point is simply that there is nothing inherently unusual about our data. Subjects respond in *aggregate* to changes in the payoff parameters in exactly the way we would expect from earlier experiments. In the 11 – 20 game, a game “that naturally triggers level-k reasoning,” (Arad and Rubinstein, p. 3562), our data has the same basic features as Arad and Rubinstein’s data.

A different concern with our data is that a large fraction of subjects might be confused or inattentive and making choices randomly. If this was the case, the problem should be especially severe for Classes 4 and 5 which were always played at the end of the experiment. In the 11 – 20 game with $\alpha_i = 20$, it is strictly dominated to choose a number below 180. We find that 92% of the choices are undominated strategies. For Class 5, the All-pay Auction, choices greater than 130 were strictly dominated for $\alpha_i = 20$. 98% of the choices are undominated.¹¹ Even at the end of the experiment, most subjects’ choices are consistent with a basic level of rationality. This implies

¹¹ Choices other than 110 were strictly dominated for the low payoff parameter ($\alpha_i = 20$). 87% of choices were 110 with the low payoff parameter, giving even stronger support to our conclusion that subjects displayed a basic level of rational choice even at the end of the experiment.

that subjects were paying attention and responding to the payoffs in the games rather than making random decisions.

Result 1: Our data is similar to what has been observed for these classes of games in previous experiments. We see little evidence of purely random choice, consistent with Prediction 2.

Comparing Tables 1 and 3, the average changes in response to shifts in own (rival's) payoff parameters are in the directions predicted for level-1 (level-2) individuals. Table 3 displays a strong pattern that subjects respond more strongly on average to changes to their own payoffs than changes to their rivals' payoffs (the change is larger in vertical comparisons than in horizontal comparisons). On average, subjects play more like level-1 than level-2 individuals, but this does not address the issue of consistency. In line with the predictions shown in Table 1, the responses to changes in the payoff parameters are far stronger for Classes 1 – 3 than Classes 4 – 5.

Result 2: On aggregate, subjects respond strongly to changes in their own payoff parameter (α_i) in Classes 1 – 3 and weakly to changes in their rival's payoff parameter (α_j). These patterns of play are consistent with a level- k model with more level-1 than level-2 (or higher) individuals.

3.2 Individual Level Data and Consistency: This subsection examines reactions to changes in own and rival's payoff parameters at the *individual* level, checking whether individuals' choices are in line with Prediction 1 which would imply that subjects employ a consistent depth of reasoning. Recall that α_i denotes a subject's *own* payoff parameter and α_j denotes their *rival's* payoff parameter. Within each class of games, there are two possible shifts in α_i holding α_j fixed: from $(\alpha_i = 20; \alpha_j = 20)$ to $(\alpha_i = 80; \alpha_j = 20)$ and from $(\alpha_i = 80; \alpha_j = 80)$ to $(\alpha_i = 20; \alpha_j = 80)$. There are also two shifts in α_j holding α_i fixed: from $(\alpha_i = 20; \alpha_j = 20)$ to $(\alpha_i = 20; \alpha_j = 80)$ and from $(\alpha_i = 80; \alpha_j = 80)$ to $(\alpha_i = 80; \alpha_j = 20)$.

Definition 1: A subject's reaction to a change of their own payoff parameter (α_i) is "consistent" with level-1 if their choice moves strictly in the predicted direction for a level-1 individual. Likewise, a subject's reaction to a change of their rival's payoff parameter (α_j) is "consistent" with level-2 if their choice moves strictly in the predicted direction for a level-2 individual.¹²

¹² Table 1 assumes a uniform distribution over actions for level-0 individuals, but for Classes 1 – 3 the specific distribution assumed doesn't matter for directional predictions as long as L0 is not deterministic. For Classes 4 and 5, we can make a directional prediction *if a shift occurs* but typically expect no response to changing the payoff parameters as per the predictions reported in Table 1. For these two classes, having no change in response to a change in your own (rival's) payoff parameter was counted as being consistent with level-1 (level-2).

We now get to the central issue of the paper. For Classes 1 – 3, having a consistent depth of reasoning (level-1 or level-2) implies a specific pattern of reactions to changing payoff parameters. A level-1 individual should only respond to changes in their own payoff parameter and level-2 individuals should only respond to changes in their rival’s payoff parameter. Within each class of games, a subject has two chances to be consistent with level-1 and two chances to be consistent with level-2. For each subject, we compute the number of reactions consistent with level-1 and with level-2 within each class of games.

Definition 2: A subject is “weakly consistent” with level-1 within a class of games if his two reactions to changes in his own payoff parameter are consistent with level-1. He is defined as “weakly consistent” with level-2 if his two reactions to changes in his rival’s payoff parameter are consistent with level-2.

Weak consistency with a specific level only requires movement in the predicted direction without any restrictions on the magnitude of the change and also allows for changes which are consistent with a different level.

Definition 3: A subject is “strongly consistent” with level-1 within a class of games if he is weakly consistent with level-1 and neither of his two reactions to changes in his rival’s payoff parameter are consistent with level-2. A subject is “strongly consistent” with level-2 within a class of games if he is weakly consistent with level-2 and neither of his two reactions to changes in his own payoff parameter are consistent with level-1.

In other words, a subject is strongly consistent with level-1, for example, if he responds in the predicted direction for a level-1 individual to both changes in his own payoff parameter (α_i) and does not respond to either change in his rival’s payoff parameter (α_j) in the predicted direction for a level-2 individual. Compared with weak consistency, strong consistency restricts how an individual classified as a level-1 (level-2) can respond to changes in his rival’s (own) payoff parameter. This restriction is weaker than what the theory calls for, namely no response to his rival’s (own) payoff parameter.¹³

¹³ Directionally, the predicted shifts for a level-3 are the same as for a level-1, the predicted shifts for a level-4 are the same as for a level-2, etc. This implies that a subject who is a consistent level-3 will be classified as strongly consistent with level-1, a subject who is consistent level-4 will be classified as strongly consistent with level-2, etc. This is not a major issue since our focus is identifying whether subjects have a consistent depth of reasoning, not what specific depth of reasoning they are using. Also, a subject who switches from being a level-1 and a level-3, for example, is classified as being strongly consistent with being a level-1. This biases our approach in favor of finding consistency.

Table 4 displays the percentage of subjects classified as weakly/strongly consistent with level-1 and level-2 within a class of games, broken down by Classes 1 – 3. The final row gives the percentage of subjects classified as weakly/strongly consistent *with the same level* for all three classes. This is the highest possible level of consistency, requiring consistency *across* classes of games as well as *within* classes.

[Insert Table 4 here]

Within any given class of games, a bit more than half the subjects are weakly consistent with level-1. This drops to only about a fifth of the subjects if we look at those who are strongly consistent with level-1. It is striking how little these percentages vary across the three classes of games. Less than a quarter of subjects are weakly consistent with level-1 for all three classes and only one individual out of 224 subjects is strongly consistent with level-1 for all three classes. Consistency with level-2 is even rarer. For any one class of games, we see less than a fifth of the subjects are weakly consistent with level-2 and almost none are strongly consistent with level-2. Once again these percentages are similar for all three classes of games. Only a single subject is weakly consistent with level-2 across all three classes of games and none is strongly consistent with level-2 across all three classes.

To check whether subjects are consistent *within* classes but inconsistent *between* classes, we calculate how many subjects are weakly/strongly consistent with some level for all three classes without requiring that they be consistent *with the same level* for all three classes. For example, a subject could be consistent with level-1 for Classes 1 and 2 and consistent with level-2 for Class 3. This slightly improves matters, with 30.4% of subjects weakly consistent with either level-1 or level-2 in all three classes. Only a single subject is strongly consistent with either level-1 or level-2 in all three classes.

Classes 1 – 3 are designed to provide a direct check for consistency that does not rely on an econometric test. We see little evidence that subjects are consistently level-1 or level-2. This is *not* due to a lack of reaction in to changing the payoff parameters. As Table 3 makes clear, on aggregate Classes 1 – 3 yield large changes in the expected directions in response to shifts in the

payoff parameters. The problem is that individual behavior is *not* in line with a consistent depth of reasoning.¹⁴

Table 5 provides an additional illustration of the lack of consistency. It displays the distribution of experimental subjects along two dimensions of consistency. The rows give the number of reactions to changing a subject's *own* payoff parameter (α_i) that are consistent with level-1 across Classes 1 – 3. The columns give a subject's number of reactions to changing their rival's payoff parameter (α_j) that are consistent with level-2 across Classes 1 – 3. A subject has two opportunities to be consistent with level-1 and two opportunities to be consistent with level-2 in each class, so these numbers range between 0 and 6. A subject who falls in the lower left corner, (row 6, column 0), was *always* classified as responding to a change in α_i in a manner consistent with level-1 and *never* classified as responding to a change in α_j in a manner consistent with level-2. A subject who falls in the upper right corner, (row 0, column 6), was *never* classified as responding to a change in α_i in a manner consistent with level-1 and *always* classified as responding to a change in α_j in a manner consistent with level-2. We report the percentage of experimental subjects falling into each cell.

[Insert Table 5 here]

If subjects used a consistent depth of reasoning across all three classes, we should observe the distribution concentrating in two regions: bottom-left for level-1 and upper-right for level-2. Table 5 shows little data in these regions, with the bulk of the observations concentrated in the lower center of the table. A rectangle (highlighted in yellow) with subjects who have 3 – 6 shifts consistent with level-1 and 2 – 4 shifts consistent with level-2 contains slightly more than two-thirds of the subjects. The data does not suggest that subjects select strategies randomly, given the strong aggregate patterns, nor does it suggest that subjects use a consistent depth of reasoning. Instead, subjects appear to mix between levels.

Result 3: Only about a fifth of subjects are strongly consistent with a specific level within classes of games for Classes 1 – 3. Virtually no subjects are consistent with a specific level across all three classes of games.

¹⁴ We learn little from analyzing consistency for Classes 4 and 5 since, as predicted, there is little response to shifts in the payoff parameters. For the sake of completeness, 79% and 93% (68% and 72%) are classified as weakly consistent with level-1 (level-2) in Classes 4 and 5 respectively. These figures drop to 4% and 2% (2% and 1%) for strong consistency with level 1 (level-2). The high frequency of weak consistency is due to the large fraction of subjects who do not change their action when the payoff parameters shift.

4. Econometric Models: The previous section provides descriptive evidence that most subjects' choices are not in line with a consistent depth of reasoning. However, the descriptive approach relies upon deterministic model predictions, ignoring noise in subjects' decision making. This leaves us with a natural question: can the lack of consistency be explained by noise in subjects' decisions?

To address this question, we formulate and estimate a wide variety of structural models. Fitting these models has multiple purposes. First, the models incorporate noise into subjects' decision-making processes, making possible to distinguish between inconsistency and stochastic choice. Second, we consider a large number of alternative models, including a number of variations suggested by the literature, and show that our consistency results are robust across different model specifications. Third, we examine the effects of subjects' cognitive ability and the payoff premium for greater depth of reasoning. Finally, we demonstrate the ability of the model to predict out of sample and discuss what this implies for the interpretation of our results.

This section provides a summary of how the model is constructed and the main results of various fitting exercises. A full description of the technical details and additional results for models beyond those discussed in this section can be found in Appendix C.

4.1 Baseline Model: The econometric models described in this section are finite mixture models, meaning we estimate the distribution of "types" (e.g. consistent level-0, consistent level-1, etc.) in the population but do not attempt to identify the type of any specific individual. In addition to "consistent" types who use a fixed depth of reasoning across all games, the models include "inconsistent" types who randomize ("mix") across different depths of reasoning. Critically, all types optimize with noise. The probability that an action is chosen is an increasing function of its expected payoff based on a subject's beliefs, but all actions are chosen with positive probability. The econometric exercise asks whether the data is more likely to have been generated by a model where all subjects use a fixed depth of reasoning or a model where some subjects are inconsistent. In the former case, the inconsistency documented in Section 3 may be explained solely by noise in the optimization process while in the latter case it reflects inconsistent depth of reasoning.

An alternative approach is to estimate a fixed depth of reasoning (level-0, level-1, etc.) for each individual subject for each class of games. A subject is identified as using an inconsistent

depth of reasoning if his *estimated* depth of reasoning varies between classes of games. However, this alternative approach can erroneously identify a subject as an inconsistent type *if there is a mistake in estimating his depth of reasoning for one of the classes*. To get a sense of how severe this problem can be, consider a population where all subjects have a fixed depth of reasoning (i.e. no inconsistency) split equally between level-1 and level-2. Suppose an econometric model is used to identify each subject's depth of reasoning for each of the five classes of games. Imagine that identification is correct with 90% probability for each class, which is quite good. We would conclude that 41% of the subjects use an inconsistent depth of reasoning when in reality there is no inconsistency!¹⁵

The baseline model allows for five types of subjects:

- Level-0: Subjects make choices consistent with a fixed probability distribution p_0 across actions, where $p_0(x)$ is the probability that a level-0 type chooses action $x \in \{110,120, \dots, 200\}$. Distribution p_0 is predetermined, and does not change across different games either within or between classes. The distribution across actions is uniform in the baseline model ($p_0(x) = 1/10$ for all $x \in \{110,120, \dots, 200\}$). The effect of using a different distribution of actions for level-0 types is covered in Section 4.3, our discussion of alternative specifications.
- Level-1: Subjects make choices based on beliefs that all other individuals are level-0 types. A level-1 type's expected payoffs depend on level-0 types' choice probabilities p_0 , the class of games being played ($C \in \{1,2,3,4,5\}$), and his own payoff parameter (α_i). The resulting probability that a level-1 type chooses action x in class C with own payoff parameter α_i is $p_1^C(x/\alpha_i)$.
- Level-2: Subjects make choices based on beliefs that all other individuals are level-1 types. A level-2 type's expected payoffs depend on level-1 types' choice probabilities $p_1^C(x/\alpha_j)$,¹⁶ the class of games being played ($C \in \{1,2,3,4,5\}$), and her own payoff parameter (α_i). The

¹⁵ Assuming that errors in identification are independent across classes of games, the probability that an individual's depth of reasoning is correctly identified in all five classes is $.9^5 = .59$. This yields the 41% figure in the text.

¹⁶ Note that the level-1 types' choice probabilities are conditioned on her rival's payoff parameter, α_j , rather than her own payoff parameter, α_i , since, from the point of view of a level-2 type, the behavior of a level-1 type depends on *her rival's* payoff parameter.

resulting probability that a level-1 type chooses action x in class C with own payoff parameter α_i and rival payoff parameter α_j is $p_2^C(x|\alpha_i, \alpha_j)$.

- **Pure-Mixing Type (Type M):** Subjects randomize (“mix”) across different levels. Pure-mixing types act as a level-1 type with probability θ_1 , a level-2 with probability θ_2 , and a level-0 with probability $1 - \theta_1 - \theta_2$. Both θ_1 and θ_2 are parameters estimated from the data. *Critically, a pure-mixing type is assumed to draw a new level for every game.* In other words, a pure-mixing type’s choices reflect twenty independent draws of one of the basic types (level-0, level-1, or level-2).
- **Semi-Mixing Type (Type S)** This type is identical to a pure-mixing type with one important exception. *Rather than drawing a new level for every game, a semi-mixing type draws a new level for every class, but uses the same level for all games within a class.* A pure-mixing type does not exhibit a consistent depth of reasoning either within a class of games or between classes of games. A semi-mixing type is consistent within a class of games, but is generally not consistent between classes. A semi-mixing type’s choices reflect five independent draws, one per class, of one of the basic types (level-0, level-1, or level-2). For simplicity, we constrain the mixing weights θ_1 and θ_2 to be the same for pure and semi-mixing types.¹⁷

We incorporate noise into subjects’ decision making. Except for level-0 types, whose choices are uniformly distributed over actions, all types use a logit rule. Define $E\pi_l^C(x|\alpha_i, \alpha_j)$ as a subject’s expected payoff from action $x \in \{110, 120, \dots, 200\}$ given his level $l \in \{1, 2\}$ and the game as defined by the class, $C \in \{1, 2, 3, 4, 5\}$ and his own and rival’s payoff parameters (α_i and α_j). His probability of choosing action x , $p_l^C(x|\alpha_i, \alpha_j)$, is given by Equation 7. The parameter λ , giving the sensitivity of subjects to differences in expected payoffs, governs the amount of noise in subjects’ decisions. If $\lambda = 0$, subjects’ choices are uniformly distributed over the ten available options. As λ increases, choices become more sensitive to differences in expected payoffs. As $\lambda \rightarrow \infty$, the distribution of choices converges to deterministic expected payoff maximization. For the baseline model, the value of λ is assumed to be the same for all types.

¹⁷ See Appendix C for a variant of the baseline model that allows for different mixing weights. This has little impacts on the results.

$$(7) \quad p_i^C(x|\alpha_i, \alpha_j) = \frac{e^{\lambda E \pi_i^C(x|\alpha_i, \alpha_j)}}{\sum_{k \in \{110, 120, \dots, 200\}} e^{\lambda E \pi_i^C(k|\alpha_i, \alpha_j)}}$$

Parameters w_1 , w_2 , w_M , and w_S assign weights (probabilities) in the mixture model to level-1, level-2, pure-mixing, and semi-mixing types respectively. A subject's chance of being a level-0 type equals $w_0 = 1 - w_1 - w_2 - w_M - w_S$.

For each subject, we observe a sequence of 20 choices, one for each game played. We construct the likelihood of observing each 20-tuple by first calculating the likelihood for each type, based on the choice probabilities described above, and then using w_0 , w_1 , w_2 , w_M , and w_S to calculate a weighted average of the likelihoods. Note that the unit of observation is a subject's 20-tuple, not each individual choice in a game by a subject. Our 224 subjects yield 224 independent observations, not $20 \times 224 = 4480$ independent observations. Observing the sequence of choices allows us to separately identify the weights of the pure-mixing type and the semi-mixing type. Although they have the same choice distribution ex ante for any specific game, they face different distributions over a *sequence* of actions for a class of games. We estimate the model parameters using a maximum likelihood approach.

4.2 Estimation Results, Baseline Model: Table 6 presents estimation results for the baseline model described above as well as two restricted versions of the baseline model. Standard errors are reported in parentheses below the parameter estimates. In addition to reporting the log-likelihood as a measure of goodness of fit, we also report the Akaike information criterion (AIC) and Bayesian information criterion (BIC). These measure the goodness of fit with a penalty for the number of parameters, with BIC imposing a larger penalty than AIC. Lower AIC/BIC indicates better fit after accounting for the number of parameters.

[Insert Table 6 here]

Model 1 is the baseline model. Looking at the results of the baseline model, the vast majority of the population is identified as belonging to one of the two “inconsistent” types that randomize over levels ($w_M + w_S = 0.893$). This resembles our descriptive analysis, but the formal econometric model picks up a much higher rate of consistency within classes of games. After accounting for stochastic choice, almost half of the subjects are estimated to be consistent within classes but mixing their depth of reasoning across classes (i.e. semi-mixing types).

Model 2 does not include either of the inconsistent types ($w_M = w_S = 0$). Comparing Models 1 and 2 allows us to see how the estimation results are affected by imposing consistency (i.e. subjects have a fixed depth of reasoning for all games). Allowing for the two inconsistent types improves the fit even after accounting for the four additional parameters in Model 1. The estimated distribution of levels is similar for Models 1 and 2. Model 2, which does not allow for inconsistency, estimates 73.5% of the population are level-1 types and 12.1% are level-2 types, while Model 1 implies that 59.7% of the population plays as a level-1 and 13.7% plays as a level-2 in any given game.¹⁸ Model 2 puts more weight on level-1 but the difference is not dramatic. The major difference between the two models are the estimated values of λ , the parameter governing the amount of noise in decision making. The value of λ is more than halved in Model 2 as compared to Model 1, implying much more noise in subjects' decisions. Model 1 has no mechanism to directly account for subjects' inconsistent depth of reasoning, so it attributes the effects of inconsistency to noise.

Model 3 only includes the two mixing types ($w_0 = w_I = w_2 = 0$). Note that $w_S = 1 - w_M$ and therefore is not reported for Model 3. Given the history of the literature, it is natural to think of the level-k model with only consistent types (Model 2) as the default, but it is equally plausible to think of a model with only inconsistent types (Model 3) as the default. Comparing Models 1 and 3 lets us see if allowing for consistent types improves the fit. The log-likelihood is improved by adding consistent types but it is not clear that this is worth the cost of adding three parameters to the model given that the BIC is larger for Model 1 than Model 3. The implied probability of playing as a level-1 or level-2 is barely affected by inclusion of consistent types,¹⁹ and the noise parameter λ is almost identical for Models 1 and 3. To a surprising extent, adding consistent types to the model has minimal effect on its explanatory power.

4.3 Estimation Results, Alternative Specifications Table 7 examines three plausible alternative specifications to the baseline model, all of which have a basis in the existing literature. More than goodness of fit, we are interested in whether alternative specifications change our main qualitative

¹⁸ The probability of playing as a level-1 in any given game for Model 1 is given by the probability of being a consistent level-1 type (w_I) plus the probability of being an inconsistent type multiplied by the probability of playing as a level-1 conditional on being an inconsistent type ($(w_M + w_S)\theta_1$). The probability of playing as a level-2 type is given by an analogous calculation.

¹⁹ The implied probability of being a level-1 is 59.7% in Model 1 vs. 59.8% in Model 3. For level-2, these figures are 13.7% and 13.5%.

conclusions: the vast majority of subjects are mixing types with a large proportion of both pure and semi-mixing types. Our general approach to modifying the baseline model is to add one feature at a time rather than fitting a kitchen sink model that adds every possible feature. This lets us see the effect of added features in isolation, limits the danger of overfitting the data through use of a huge number of parameters, and reduces the computational demands of fitting the models.

[Insert Table 7 here]

The first column of Table 7 repeats the baseline model (Model 1) as a point of comparison. Model 4 (“Non-Uniform Level 0”) uses an alternative specification for the choice probabilities of level-0 types. A reasonable interpretation of level-0 types is that their choices are driven by non-strategic considerations. The most common way of specifying the choice probabilities of level-0 types is to assume a uniform distribution over actions as in the baseline model, but the level-k model does not require this restriction.²⁰ Model 4 allows for the possibility that level-0 play puts extra weight on other natural non-strategic concepts for how to play the games.

All the games we consider have a “cooperative” choice, defined as the choice that maximizes a player’s payoffs *subject to both players making identical choices*. For example, the cooperative choice is 200 in the imperfect price competition games. The cooperative choice is *not* consistent with a Nash equilibrium for most of the classes and is often *not* efficient (in the sense of maximizing total payoffs across the two players) for the asymmetric games.²¹ All of the games also have a “safe” choice, defined as the maximin choice. In the minimum coordination game, for example, choice of 110 maximizes the minimum possible payoff. Model 4 lets level-0 types put extra weight on the cooperative and safe choices as natural non-strategic options.²² The parameter γ_{Coop} gives the added weight that level-0 types put on the cooperative choice and the parameter γ_{Safe} gives the added weight on the safe choice. With probability $1 - \gamma_{Coop} - \gamma_{Safe}$, level-0 types choose using a uniform distribution. The results indicate that level-0 types significantly

²⁰ This is a central point of Arad and Rubinstein’s analysis of the 11-20 game.

²¹For instance, it is easily confirmed that the cooperative choice, mutual choice of 200, is neither a Nash equilibrium nor surplus maximizing for the asymmetric imperfect price competition games.

²² The cooperative choice is 200 in Classes 1 – 4 and 110 in Class 5. The safe choice is 110 in Classes 1, 2, 3 and 5 and 200 in Class 4. In practice, putting extra weight on the cooperative and safe choices amounts to putting extra weight on the tails of the distribution. We could do this by directly fitting a distribution (i.e. a discretized beta distribution) over the actions, but this runs into problems that using safe and cooperative choices avoids. Specifically, if we mechanically put more weight on the two tails it implies more weight on choice of 110 in the 11 – 20 games and choice of 200 in the all-pay auctions. Both of these choices are strictly dominated and virtually never chosen, causing the model to put artificially little weight on the tails.

overweight the safe choice, but not the cooperative choice. Relaxing the uniform distribution assumption improves the model’s fit but has little impact on the model’s main qualitative feature as the percentages of pure and semi-mixing types are little changed from the baseline model.

Model 5 implements a variant of the cognitive hierarchy (CH) model of Camerer, Ho, and Chong (2004). In a standard level-k model, a level-k type assumes that the rest of the population consists of individuals who are one level lower ($k - 1$). Therefore, a level-2 type assumes that all other individuals are level-1 types. In our version of CH model, level 2 types take into account that both level-1 types and level-0 types exist, and use Bayes rule to generate beliefs about the likelihood of being matched with a level-1 type: $\sigma_1 = \frac{w_1 + (w_M + w_S)\theta_1}{(1 - w_2 - w_M - w_S) + (w_M + w_S)(1 - \theta_2)}$. This is slightly different from Camerer et al’s model as we are using rational expectations to generate beliefs rather than applying a Poisson distribution. The CH model yields a slightly better fit to the data. The overall fraction of inconsistent types is a bit higher than in the baseline model and the distribution is shifted toward the semi-mixing type. The overall interpretation changes little: almost all subjects have an inconsistent depth of reasoning and both pure-mixing and semi-mixing types are common.

To keep the baseline model simple, we only allowed for three depths of reasoning: level-0, level-1, and level-2. There is ample evidence of higher depth of reasoning from other papers (e.g. Kneeland, 2015). Allowing for higher depth of reasoning should have little effect on our descriptive analysis of consistency. As noted previously, the pattern of shifts in response to changing payoff parameters should be the same for a level-3 as a level-1, the same for a level-4 as a level-2, etc. However, adding higher level types should improve our ability to fit the data. Model 6 tests this conjecture by adding level-3 types. Two parameters are added to the baseline model: w_3 is the weight of consistent level-3 types and θ_3 is the probability the two mixing types put on playing as a level-3 type. Adding level-3 types to the model improves the fit, as expected. The model detects no consistent level-3 types in the population, but the weight inconsistent types put on level-3 is both statistically and economically significant. The fraction of inconsistent types ($w_M + w_S$) increases slightly relatively to the baseline model, but the distribution between pure-mixing and semi-mixing types is almost unchanged.

In summary, all of the alternative models find that a high frequency of inconsistent types is a robust feature of our empirical setting. Appendix C includes results on additional alternative specifications, including models that vary the mixing probabilities (i.e. θ_1 and θ_2) between the two

inconsistent types, models that vary the mixing probabilities between different classes of games, models with fewer inconsistent types, and models with more inconsistent types. Our main finding is robust as inconsistent types are predominant in all specifications.

4.4 Determinants of the Distribution of Types: The models shown in Table 8 examine the determinants of subjects' types. The first column once again repeats the baseline model (Model 1) as a point of comparison.

[Insert Table 8 here]

We have presented ample evidence that most subjects use an inconsistent depth of reasoning, but this inconsistency needs not imply that the depth of reasoning is arbitrary. Thinking more deeply about a game presumably requires effort, and subjects should be more willing to expend effort when the potential reward is larger. We therefore expect a shift to higher levels when the benefits of a greater depth of reasoning are increased. Alaoui and Penta (2016) present a formal model that captures this intuition as well as experimental evidence that depth of reasoning is sensitive to incentives. Model 7 modifies the baseline model to see if the distribution over levels used by inconsistent types responds to incentives to reason more deeply about the games.

To capture the effects of incentives, we first calculate the expected payoff for each level of reasoning (level-0, level-1, and level-2) in each game. Specifically, the model generates a distribution over own actions as a function of the game being played and a subject's depth of reasoning (level-0, level-1, or level-2).²³ The population's observed distribution of choices is used to generate a distribution over their rival's actions. Combining these, we calculate expected payoffs for each level. This is done by game for pure-mixing types and by class for semi-mixing types. We then calculate the payoff premium for being a level-1 (expected payoff for level-1 minus the expected payoff for level-0) and the payoff premium for being a level-2 (expected payoff for level-2 minus the expected payoff for level-1). Conditional on being an inconsistent type, the mixing weight of each possible depth of reasoning is a linear function of the payoff premium. Abusing notation, $\theta_1 = \bar{\theta}_1 + \mu_1 \cdot (E\pi_1 - E\pi_0)$ where $\bar{\theta}_1$ and μ_1 are parameters estimated from the data. Likewise, $\theta_2 = \bar{\theta}_2 + \mu_2 \cdot (E\pi_2 - E\pi_1)$.²⁴

²³ This distribution is a function of λ , but not the other estimated parameters.

²⁴ To make sure that $\theta_1 + \theta_2$ is between 0 and 1, we use a logit transformation. It follows that the mixing weights are *not* linear in payoff premiums. For more details of Models 7 - 9 and logit transformations, see Appendix C.

Comparing Model 7 and Model 1, the fit is improved by allowing the distribution of levels to depend on the payoff premiums. The estimates for μ_1 and μ_2 are both positive, indicating that greater payoff premiums are associated with greater depth of reasoning. This does not imply that mixing disappears if the model accounts for the payoff premiums. The distribution over levels generated by the fitted model puts substantial weight on multiple levels in all classes and all games. The weights on the various levels change as the optimization premium varies, but never approaches either 0% or 100%.²⁵ The likelihood of being either inconsistent type is little changed. In line with the results of Alaoui and Penta, there is a systematic relationship between depth of reasoning and incentives. Inconsistent depth of reasoning does not imply arbitrary depth of reasoning.

Models 8 and 9 examine the relationship between subjects' reasoning ability, as measured by their scores on the Raven's Progressive Matrices (RPM) test, and their depth of reasoning. We expected a positive relationship based on the results of Gill and Prowse (2016). The two models examine this issue in slightly different ways. Model 8 allows the mixing probabilities (θ_1 and θ_2) for the inconsistent types to vary with the RPM score. Model 9 lets the weight on the two inconsistent types (w_1 and w_2) vary with the RPM score, but does not allow the mixing probabilities (θ_1 and θ_2) to depend on the subject's RPM score.

In Model 8, θ_1 and θ_2 are linear functions of the subject's RPM score. Abusing notation, probability of an inconsistent type being level-1 is $\theta_1 = \bar{\theta}_1 + \mu_1 \cdot RPM$ where $\bar{\theta}_1$ and μ_1 are parameters estimated from the data. Likewise, probability of being level-2 is $\theta_2 = \bar{\theta}_2 + \mu_2 \cdot RPM$.²⁶ Comparing Model 8 and Model 1, the fit is improved by allowing the mixing probabilities to depend on the RPM score. The estimates for μ_1 and μ_2 are both positive, indicating that greater reasoning ability is associated with greater depth of reasoning (albeit not significantly in the case of μ_2). Once again, this indicates that subjects' levels of reasoning, while inconsistent, vary in a sensible and systematic fashion. The probability of being either inconsistent type changes little.

Model 9 makes the weight on being an inconsistent type a linear function of the subject's RPM score. Abusing notation, $w_M + w_S = \phi + \beta \cdot RPM$ where ϕ and β are parameters fit from the data. Subject to being a mixing type, δ_M is the probability of being a pure-mixing type. This implies that $w_M = \delta_M \cdot (\phi + \beta \cdot RPM)$. Subject to *not* being a mixing type, δ_1 and δ_2 are the

²⁵ To give a sense of how the weight on levels varies with as the optimization premiums change, the implied weight on level-1 (averaging across games) is 44%, 44%, 63%, 81%, and 59% for Classes 1 – 5 respectively.

²⁶ As in Model 7, we use a logit transformation of mixing probabilities in Model 8.

probabilities of being a consistent level-1 and level-2 type respectively. Letting the weight on being a mixing type vary with the RPM score does *not* improve the model's ability to fit the data relative to the baseline model after penalizing the model for using more parameters: both the AIC and BIC are *higher* in the modified model. Looking at the parameter estimates, the estimate for β is small and does not approach statistical significance. Going from the 10th percentile ($RPM = 9$) to the 90th percentile ($RPM = 13$) implies a decrease in the probability of being an inconsistent type from 94% to 87%.²⁷ The likelihood of having an inconsistent depth of reasoning is unresponsive to RPM scores. Taken together, Models 8 and 9 indicate that cognitive ability affects the depth of reasoning used by inconsistent types, but does not affect the probability of being an inconsistent type.

To summarize, almost all of our subjects are inconsistent, using different depths of reasoning in different games, but this inconsistency is *not* pure noise. The distribution over depths of reasoning varies in a sensible way in response to incentives and subjects' reasoning ability.

4.4 Out-of-Sample Prediction: We have demonstrated that inconsistent depth of reasoning is a robust feature of the level-k model. This raises an obvious question: What are the implications of this finding? Our discussion of Model 2 on Table 6 addressed the effects of failing to account for inconsistency on estimates of the distribution over levels. This subsection considers the model's ability to predict out of sample.

[Insert Table 9 here]

To do this, we fit Model 1 (the baseline model) and Model 2 (the level-k model without inconsistency) from Table 6 to data from Classes 2 – 5 and then use the estimated parameters to predict choices in Class 1 (imperfect price competition).²⁸ Specifically, after fitting the models to Classes 2 – 5, we use the implied distribution over choices to generate a predicted mean and standard deviation for each game. These are reported in Table 9 as well as the mean and standard deviation for the observed data. Model 1 fits the data better than Model 2, but both models do a

²⁷ Analogous to Model 7 and 8, a logit transformation of type weights is used in Model 9. It follows that the type weights are not linear in RPM scores.

²⁸ The choice of Class 1 as the predicted class was arbitrary, but the qualitative conclusions do not depend on which class we try to predict.

credible job of predicting data for Class 1. Both models have some differences from the observed data, but allowing for inconsistency does not meaningfully reduce these differences.

The ability of Model 2 to predict out of sample should not be surprising. Model 1 fits the data better than Model 2, because it accounts for individuals changing their depth of reasoning. This enables Model 1 to better capture the pattern of changes between games *at the individual level*. However, the prediction exercise reported in Table 9 is not concerned with the pattern of changes between games *at the individual level*. It doesn't matter if the same subjects remain at a specific level across the different games. All that matters is the fraction of subjects at any particular level for a specific game. Both models predict well because the distribution across levels is fairly stable across our five classes of game.²⁹

This leads to a more general point. The level-k model that does *not* allow for inconsistency will generally work fine if all we care about is the distribution over depths of reasoning for a specific game. For example, consider the Crawford and Iriberri (2007) application of a level-k model to the winner's curse. The predictions of their model are driven by the presence of players with different depths of reasoning. Consistency across games does not matter in their model. We should not change our interpretation of their model or experimental work just because we know that subjects' types are not consistent across games.

To summarize, inconsistent depth of reasoning is a major feature of our experimental data. This inconsistency is important both because it informs us about the nature of subjects' decision making processes and also affects our ability to estimate parameters for a level-k model. It does *not* imply that applications of level-k model are necessarily flawed, or that level-k models cannot predict out of sample. These exercises only run into difficulty if they are impacted directly by inconsistency or if the inconsistency implies not only that subjects change their depths of reasoning but also that the distribution over levels changes across games (or classes of games).

5. Conclusions: The primary purpose of this paper is to explore whether or not subjects employ a consistent depth of reasoning when playing games. Subjects play a series of games where consistency implies a specific pattern of responses to changes in payoff parameters. We observe little evidence of consistency. This is true whether we use descriptive analysis that does not rely

²⁹ This implies that if we studied classes of games where the distribution of levels varied more between classes, the ability of the model to predict out of sample would be diminished.

on any specific econometric model or take an econometric approach that accounts for noise in subjects' decision making. The lack of consistency is quite robust to a wide variety of alternative model specifications. Many subjects are consistent within classes of games and only vary their depth of reasoning between classes of games, but a large fraction of the population also vary their depth of reasoning within classes of games. This is particularly damaging to the assumption that subjects use a consistent depth of reasoning, as it cannot easily be explained away as subjects thinking differently (or expecting others to think differently) about different types of games.

The robust inconsistency we observe implies neither that subjects' depth of reasoning is purely random nor that level-k models are not useful tools. Depth of reasoning responds systematically to changes in incentives and the cognitive ability of subjects. Even without accounting for inconsistency, a level-k model does well at predicting out of sample for our dataset. Accounting for the inconsistency is obviously important, especially when estimating parameters for a level-k model, but for applications that focus on the aggregate distribution of behavior, inconsistency at the *individual* level does not play a central role.

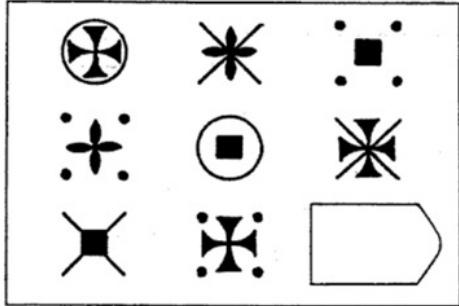
Ultimately, we argue that inconsistency is important not only because it affects the estimation of level-k model parameters, but also because it tells us something about the nature of subjects' decision-making processes. Experimenters and theorists have accepted the idea of stochastic choice in individual decision making (e.g. Agranov and Ortoleva, 2017). Depth of reasoning is just another individual choice. It isn't a big step to say that random choice models are just as applicable here as elsewhere.

References

- AGRANOV, M., AND ORTOLEVA, P. “Stochastic Choice and Preferences for Randomization.” *Journal of Political Economy* 125, no. 1 (2017): 40–68.
- AGRANOV, M., POTAMITES, E., SCHOTTER, A., AND TERGIMAN, C. “Beliefs and Endogenous Cognitive Levels: An Experimental Study.” *Games and Economic Behavior* 75, no. 2 (2012): 449–463.
- ALAOUI, L., AND PENTA, A. “Endogenous Depth of Reasoning.” *The Review of Economic Studies*, 83, no. 4 (2016): 1297–1333.
- ARAD, A., AND RUBINSTEIN, A. “The 11–20 Money Request Game: A Level-k Reasoning Study.” *American Economic Review* 102, no. 7 (2012): 3561–3573.
- CAMERER, C. F., HO, T.-H., AND CHONG, J.-K. “Sophisticated Experience-weighted Attraction Learning and Strategic Teaching in Repeated Games.” *Journal of Economic theory* 104, no. 1 (2002): 137–188.
- . “A Cognitive Hierarchy Model of Games.” *The Quarterly Journal of Economics* 119, no. 3 (2004): 861–898.
- CAPRA, C. M., GOEREE, J. K., GOMEZ, R., AND HOLT, C. A. “Anomalous Behavior in a Traveler’s Dilemma?” *American Economic Review* 89, no. 3 (1999): 678–690.
- CAPRA, M., GOEREE, J. K., GOMEZ, R., AND HOLT, C. A. “Learning and Noisy Equilibrium Behavior in an Experimental Study of Imperfect Price Competition.” *International Economic Review* 43, no. 3 (2002): 613–636.
- COSTA-GOMES, M., AND CRAWFORD, V. P. “Cognition and Behavior in Two-person Guessing games: An Experimental Study.” *American Economic Review* 96, no. 5 (2006): 1737–1768.
- COSTA-GOMES, M., CRAWFORD, V. P., AND BROSETA, B. “Cognition and Behavior in Normal-form Games: An Experimental Study.” *Econometrica* 69, no. 5 (2001): 1193–1235.
- COSTA-GOMES, M. A., AND WEIZSÄCKER, G. “Stated Beliefs and Play in Normal-form Games.” *The Review of Economic Studies* 75, no. 3 (2008): 729–762.
- CRAWFORD, V. P., COSTA-GOMES, M. A., AND IRIBERRI, N. “Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications.” *Journal of Economic Literature* 51, no. 1 (2013): 5–62.
- CRAWFORD, V. P., AND IRIBERRI, N. “Level-k Auctions: Can a Nonequilibrium Model of Strategic Thinking Explain the Winner’s Curse and Overbidding in Private-Value Auctions?” *Econometrica* 75, no. 6 (2007): 1721–1770.
- FISCHBACHER, U. “z-Tree: Zurich toolbox for ready-made economic experiments.” *Experimental Economics* 10, no. 2 (2007): 171–8.


- GEORGANAS, S., HEALY, P. J., AND WEBER, R. A. “On the Persistence of Strategic Sophistication.” *Journal of Economic Theory* 159 (2015): 369–400.
- GILL, D., AND PROWSE, V. “Cognitive Ability, Character Skills, and Learning to Play Equilibrium: A Level-k Analysis.” *Journal of Political Economy* 124, no. 6 (2016): 1619–1676.
- GNEEZY, U., AND SMORODINSKY, R. “All-pay Auctions - An Experimental Study.” *Journal of Economic Behavior & Organization* 61, no. 2 (2006): 255–275.
- GOEREE, J. K., AND HOLT, C. A. “A Model of Noisy Introspection.” *Games and Economic Behavior* 46, no. 2 (2004): 365–382.
- . “An Experimental Study of Costly Coordination.” *Games and Economic Behavior* 51, no. 2 (2005): 349–364.
- HARGREAVES HEAP, S., ROJO ARJONA, D., AND SUGDEN, R. “How Portable Is Level-0 Behavior? A Test of Level-k Theory in Games With Non-Neutral Frames.” *Econometrica* 82, no. 3 (2014): 1133–1151.
- IVANOV, A., LEVIN, D., AND NIEDERLE, M. “Can Relaxation of Beliefs Rationalize the Winner’s Curse?: An Experimental Study.” *Econometrica* 78, no. 4 (2010): 1435–1452.
- KLINE, B. “An Empirical Model of Non-equilibrium Behavior in Games.” *Quantitative Economics*. (2017): *Forthcoming* .
- KNEELAND, T. “Identifying Higher-Order Rationality.” *Econometrica* 83, no. 5 (2015): 2065–2079.
- NAGEL, R. “Unraveling in Guessing Games: An Experimental Study.” *American Economic Review* 85, no. 5 (1995): 1313–1326.
- ÖSTLING, R., WANG, J. T.-Y., CHOU, E. Y., AND CAMERER, C. F. “Testing Game Theory in the Field: Swedish LUPU Lottery Games.” *American Economic Journal: Microeconomics* 3, no. 3 (2011): 1–33.
- STAHL, D. O., AND WILSON, P. W. “On Players’ Models of Other Players: Theory and Experimental Evidence.” *Games and Economic Behavior* 10, no. 1 (1995): 218–254.


Figure 1: SAMPLE QUESTION FROM RPM TEST





Time remaining 30


Question 11. From the lower part, identify the element that is missing from the pattern of shapes in the upper part. You have 30 seconds to answer; if not, this question will be counted as answered incorrectly.

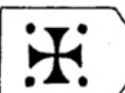
1 

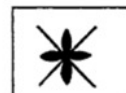
2 

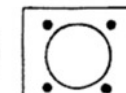
3 

4 

5 

6 

7 

8 

1

2

3

4

5

6

7

8

OK

Figure 2: DISTRIBUTION OF CHOICES IN THE “11-20” GAME

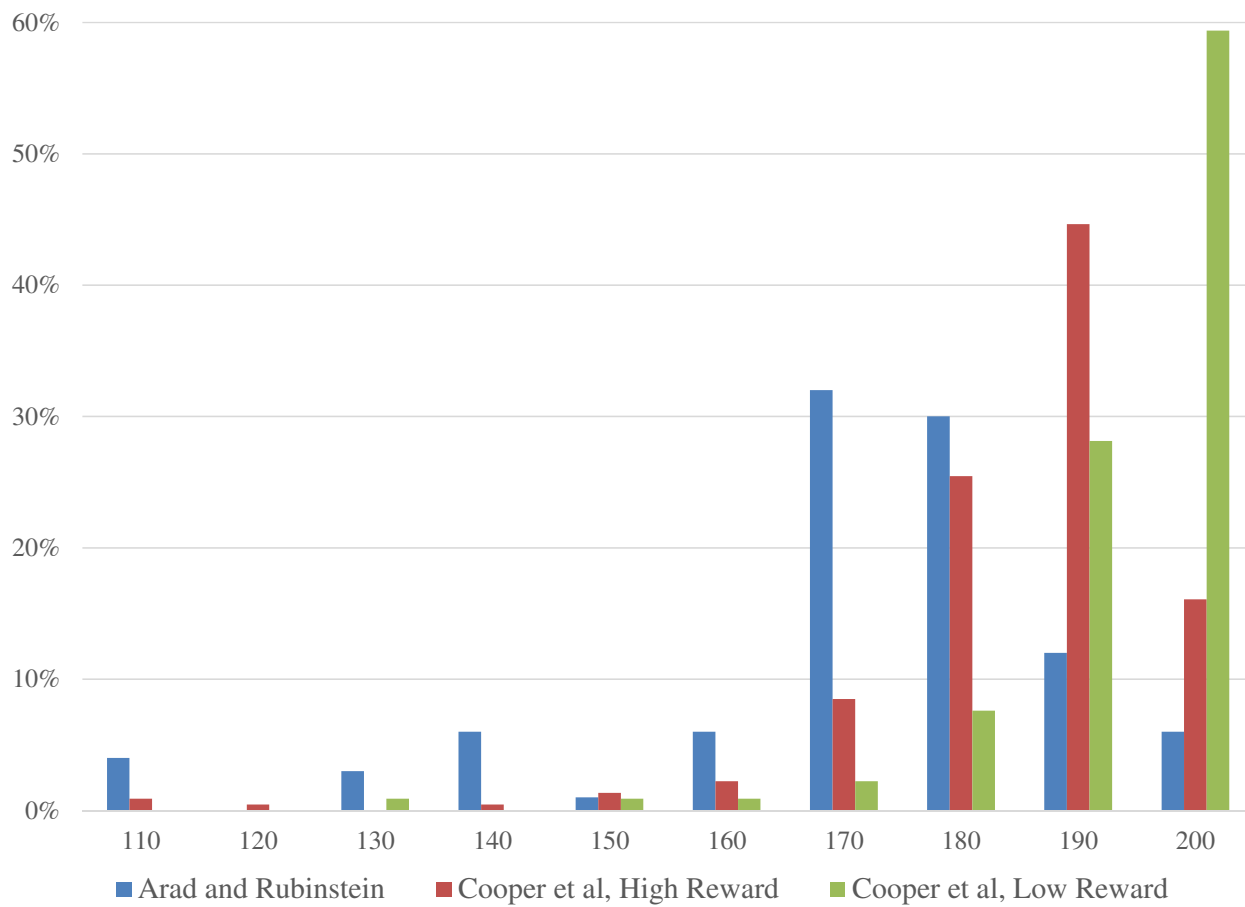


Table 1: LEVEL-K PREDICTIONS

CLASS 1: IMPERFECT PRICE COMPETITION		
	20	80
20	L1: 110 , L2: 110	L1: 110 , L2: 160
80	L1: 170 , L2: 110	L1: 170 , L2: 160

CLASS 2: MINIMUM COORDINATION GAME		
	20	80
20	L1: 180-190, L2: 180-190	L1: 180-190, L2: 120-130
80	L1: 120-130, L2: 180-190	L1: 120-130, L2: 120-130

CLASS 3: TRAVELER'S DILEMMA		
	20	80
20	L1: 160-170, L2: 150-160	L1: 160-170 , L2: 110
80	L1: 110 , L2: 150-160	L1: 110 , L2: 110

CLASS 4: 11-20 GAME		
	20	80
20	L1: 200 , L2: 190	L1: 200 , L2: 190
80	L1: 200 , L2: 190	L1: 200 , L2: 190

CLASS 5: ALL-PAY AUCTION		
	20	80
20	L1: 110 , L2: 110-120	L1: 110 , L2: 110-120
80	L1: 110 , L2: 120	L1: 110 , L2: 120

Table 2: SUMMARY OF SESSIONS

SESSION	# OF SUBJECTS	PLAYER 1	PLAYER 2	ORDER OF CLASSES
1	50	25	25	1/2/3/4/5
2	50	25	25	1/2/3/4/5
3	60	30	30	2/3/1/4/5
4	64	32	32	3/1/2/4/5
TOTAL	224	112	112	

Table 3: AVERAGE CHOICE

CLASS 1: IMPERFECT PRICE COMPETITION		
	20	80
20	131.0	133.3
80	157.2	164.1

CLASS 2: MINIMUM COORDINATION GAME		
	20	80
20	178.1	164.0
80	141.7	135.1

CLASS 3: TRAVELER'S DILEMMA		
	20	80
20	167.7	157.2
80	134.2	129.5

CLASS 4: 11-20 GAME		
	20	80
20	193.6	189.3
80	185.8	184.9

CLASS 5: ALL-PAY AUCTION		
	20	80
20	112.6	112.9
80	121.5	126.3

Table 4: WEAKLY AND STRONGLY CONSISTENT LEVEL-1 AND LEVEL-2

CLASS	LEVEL 1		LEVEL 2	
	STRONG	WEAK	STRONG	WEAK
1	22.32%	56.70%	1.34%	17.86%
2	16.07%	58.93%	0.89%	19.20%
3	20.54%	53.13%	3.57%	16.52%
AVERAGE	19.64%	56.25%	1.93%	17.86%
CLASSES 1 - 3	0.45%	23.66%	0%	0.45%

Table 5: CONSISTENT SHIFTS ACROSS CLASSES 1 - 3

		SHIFTS CONSISTENT WITH LEVEL-2							TOTAL
		0	1	2	3	4	5	6	
SHIFTS CONSISTENT WITH LEVEL-1	0	3.60%	0.40%	0.00%	0.00%	0.00%	0.00%	0.00%	4.00%
	1	0.40%	3.10%	0.00%	0.40%	0.40%	0.00%	0.00%	4.50%
	2	0.00%	1.30%	2.70%	1.30%	0.40%	0.00%	0.00%	5.80%
	3	0.90%	1.30%	6.30%	2.20%	1.30%	0.90%	0.00%	12.90%
	4	0.00%	3.60%	6.30%	5.40%	4.90%	1.30%	0.00%	21.40%
	5	0.00%	4.00%	6.70%	9.40%	5.80%	1.30%	0.40%	27.70%
	6	0.40%	2.70%	6.70%	8.90%	3.60%	1.30%	0.00%	23.70%
TOTAL		5.40%	16.50%	28.60%	27.70%	16.50%	4.90%	0.40%	100.00%

Table 6: ESTIMATION RESULTS OF THE BASELINE MODELS

	MODEL 1	MODEL 2	MODEL 3
	BASELINE	CONSISTENT TYPES ONLY	MIXING TYPES ONLY
w_1	0.097*** (0.036)	0.735*** (0.039)	0 FIXED
w_2	0.000 -	0.121*** (0.034)	0 FIXED
w_M	0.431*** (0.065)	0 FIXED	0.434*** (0.016)
w_S	0.462*** (0.017)	0 FIXED	
θ_1	0.560*** (0.065)		0.598*** (0.019)
θ_2	0.153*** (0.023)		0.135*** (0.016)
λ	0.175*** (0.010)	0.076*** (0.002)	0.172*** (0.010)
LOG LIKELIHOOD	-8,201.187	-8,308.313	-8,206.992
AIC	16,416.375	16,622.626	16,421.984
BIC	16,440.256	16,632.861	16,435.631

NOTES: Standard errors are given in parentheses. Three (***) , two (**), and one (*) stars indicate statistical significance at the 1%, 5%, and 10% respectively.

Table 7: COMPARISON BETWEEN BASELINE AND VARIANT MODELS

	MODEL 1 BASELINE	MODEL 4 NON-UNIFORM LEVEL 0	MODEL 5 COGNITIVE HIERARCHY	MODEL 6 MODEL WITH LEVEL 3
w_1	0.097*** (0.036)	0.075* (0.041)	0.079** (0.035)	0.062** (0.027)
w_2	0.000 -	0.000 -	0.000 -	0.000 -
w_M	0.431*** (0.065)	0.447*** (0.075)	0.283*** (0.062)	0.478*** (0.048)
w_S	0.462*** (0.017)	0.463*** (0.068)	0.632*** (0.025)	0.459*** (0.050)
θ_1	0.560*** (0.065)	0.595*** (0.021)	0.470*** (0.068)	0.476*** (0.022)
θ_2	0.153*** (0.023)	0.117*** (0.015)	0.217*** (0.034)	0.108*** (0.016)
λ	0.175*** (0.010)	0.183*** (0.015)	0.165*** (0.008)	0.195*** (0.010)
γ_{safe}		0.105*** (0.011)		
γ_{coop}		0.000 -		
ρ_3				0.000 -
θ_3				0.222*** (0.024)
LOG LIKELIHOOD	-8,201.187	-8,097.555	-8,192.194	-8,117.500
AIC	16,416.375	16,213.109	16,398.388	16,253.000
BIC	16,440.256	16,243.814	16,422.270	16,283.704

NOTES: Standard errors are given in parentheses. Three (***), two (**), and one (*) stars indicate statistical significance at the 1%, 5%, and 10% respectively.

Table 8: COMPARISON BETWEEN BASELINE AND VARIANT MODELS

	MODEL 1	MODEL 7	MODEL 8	MODEL 9
	BASELINE	EXP. PAYOFF	RAVEN PREDICT MIX. PROB.	RAVEN PREDICT MIX. TYPE
w_1	0.097*** (0.036)	0.035 (0.032)	0.084** (0.035)	
w_2	0.000 -	0.000 -	0.000 -	
w_M	0.431*** (0.065)	0.439*** (0.047)	0.459*** (0.066)	
w_S	0.462*** (0.017)	0.509*** (0.051)	0.453*** (0.066)	
θ_1	0.560*** (0.065)			0.555*** (0.023)
θ_2	0.153*** (0.023)			0.152*** (0.017)
λ	0.175*** (0.010)	0.178*** (0.008)	0.179*** (0.011)	0.173*** (0.010)
θ_1		-0.693*** (0.162)	-1.023** (0.507)	
$\bar{\theta}_2$		-2.340*** (0.300)	-1.797** (0.768)	
μ_1		0.064*** (0.007)	0.144*** (0.043)	
μ_2		0.201*** (0.021)	0.098 (0.065)	
δ_1				0.961*** (0.079)
δ_2				0.000 -
δ_M				0.469*** (0.070)
ϕ				4.605* (2.600)
β				-0.211 (0.208)
LOG LIKELIHOOD	-8,201.187	-8,112.769	-8,195.756	-8,200.671
AIC	16,416.375	16,243.537	16,409.511	16,417.341
BIC	16,440.256	16,274.242	16,440.216	16,444.635

NOTES: Standard errors are given in parentheses. Three (***), two (**), and one (*) stars indicate statistical significance at the 1%, 5%, and 10% respectively.

Table 9: COMPARISON OF DATA DISTRIBUTIONS TO MODEL DISTRIBUTIONS IN CLASS 1

	DATA		CONSISTENT TYPES ONLY		BASELINE MODEL	
	MEAN	S.D.	MEAN	S.D.	MEAN	S.D.
LL	131.03	26.91	131.33	21.02	129.50	23.11
HH	164.06	28.86	163.00	24.42	164.67	22.48
LH	133.35	25.81	133.22	20.69	133.21	22.33
HL	157.23	30.01	161.63	25.27	161.27	25.31

	DATA		CONSISTENT TYPES ONLY		BASELINE MODEL	
	MEAN	S.D.	MEAN	S.D.	MEAN	S.D.
LL-HL	-26.21	36.31	-30.29	33.05	-31.76	34.56
LH-HH	-30.71	33.19	-29.78	32.75	-31.46	33.27
LL-LH	-2.32	28.38	-1.89	27.17	-3.71	28.31
HL-HH	-6.83	30.06	-1.38	34.97	-3.41	33.34

NOTES: The parameters used for each model are estimated using data of Classes 2, 3, 4 and 5, but not including data from Class 1.