Testing Cognitive Hierarchy Assumptions*

Daniel E. Fragiadakis
Texas A&M University

Ada Kovaliukaite
Texas A&M University

David Rojo Arjona
Leicester University

November 1, 2017

JOB MARKET PAPER

Abstract

Camerer et al. [2004] propose a Cognitive Hierarchy model that characterizes an individual as believing her opponents engage in heterogeneous steps of strategic thinking. We are the first paper, to our knowledge, to directly test this assumption. In one treatment, subjects play games. Their behavior is then predicted by different participants in a separate treatment. A subject often believes in a handful of non-strategic individuals who play a naive strategy. She also, however, expects there are players who play the best-response to this strategy as well as types who best respond to these strategic individuals and so on. We thus find that Cognitive Hierarchy's beliefs assumptions clearly pass our test.

1 Introduction

Camerer et al. [2004] develop a Cognitive Hierarchy (CH) model of behavior in one-shot games and use it to characterize non-equilibrium play typically observed in lab experiments.[1] CH has since also been used to estimate behavior in the field.[2] Though used to model aggregate data, CH makes individual-level assumptions. Specifically, it describes a player as a Step k thinker who best-responds to a belief that others are Step 0 through Step $k-1$ thinkers, where Step 0 thinkers play naive strategies.[3] To our knowledge, direct evidence that an individual holds

[*] We are especially grateful to Alex Brown and Catherine Eckel for their impacts on the design of the experiment. We thank Isabelle Brocas, Juan Carillo, Gary Charness, Ryan Oprea, Steven Puller, Tim Salmon, Emmanuel Vespa, Sevgi Yuksel and Georg Weizsäcker for discussions and helpful comments.

[1] See Camerer [2003] for a behavioral game theory overview.

[2] For example, Brown et al. [2013] use CH to more accurately predict moviegoer behavior in comparison to equilibrium. Hortaçu et al. [2017] estimate firms' levels of strategic sophistication using the CH model and find that larger firms engage in more sophisticated reasoning compared to smaller firms.

[3] Stahl and Wilson [1995] also describe an individual as having diverse beliefs over opponents.

such beliefs is nonexistent in the sense that we are unaware of a study that tests whether an individual anticipates thinkers of several steps when asked to forecast the play of a population. Such a test is important, however, if CH is intended to be a descriptive model. In this paper, we conduct such a test using a lab experiment.

In our *Actions* treatment, subjects play a series of two-person Number Selection (NS) games with no feedback. In NS games, players simultaneously select integers, say, between 1 and 14, inclusive. If a player selects an integer, i, she earns i points.[4] Furthermore, a player earns 100 bonus points her number is exactly 3 (in some games, 4) less than her opponent's number and earns 35 points if her number equals her opponent's number.[5] We expect a non-strategic Step 0 thinker to play 14 in this example, since this is the best number to choose if one does not consider how his payoff is affected by his opponent's play.[6] A Step 1 player, then, who assumes her opponents are Step 0 types, is incentivized to choose 11. Step 2 players anticipate Step 1 *and* Step 0 opponents; thus, depending on the particular distribution of beliefs held by a Step 2 player, she may have a best-response of 11 or 8. Continuing this procedure, it can be shown that the predictions from Step k reasoning are confined to $S = \{14, 11, 8, 5, 2\}$. Furthermore, a Step k thinker who anticipates others to play some $s \in S$ should also expect strategies $\{s' | s' > s \text{ and } s' \in S\}$ to be played. This provides us with a clean way to test for Step k beliefs in our next treatment.

In our *Beliefs* treatment, each participant states her beliefs regarding the play of a set of 20 *Actions* subjects in each of the NS games. To state her beliefs for a particular NS game, a *Beliefs* participant constructs a 20-box histogram over a horizontal axis that displays each of the game's pure strategies. For constructing a given histogram, h, a subject earns p points, where p is the number of boxes that would be overlapping were we to place h on top of the histogram of actual behavior by the 20 *Actions* subjects. This renders our mechanism incentive compatible: if a *Beliefs* subject thinks s subjects chose the number n, she is incentivized to place s boxes above n when constructing her histogram. We also record the order in which *Beliefs* participants arrange their boxes in their histograms.[7] Despite having no effect on

[4] In both treatments, earned points are used to pay subjects with binary lotteries (see Roth and Malouf [1979]), incentivizing a subject to maximize her expected number of points, irrespective of her risk preferences.

[5] Our NS games are inspired by others from Arad and Rubinstein [2012], Georganas et al. [2015] and Fragiadakis et al. [2017].

[6] An alternative rule-of-thumb for a Step 0 thinker is to select an integer uniformly at random. Either of these Step 0 specifications give rise to a best-response of guessing 11 in this example, thus, both rules lead Step 1 thinkers to the same number.

[7] The order in which boxes are arranged does not affect payoffs and the instructions make no mention of the order in which boxes are arranged. We view this minimally invasive "belief-tracking" procedure as relating to studies using "eye-tracking" to record where subjects direct their attention (see Wang et al. [2010] as well as studies where subjects must actively "open" boxes to observe payoffs from certain strategy profiles (see Costa-Gomes et al. [2001]).

payoffs, we expect a *Beliefs* participant to place boxes on lower thinking steps earlier.

We find strong support for Step k beliefs: for 48% of our *Beliefs* participants, conditional on anticipating k steps of reasoning, k' steps of reasoning are also expected for all $k' \leq k$. Illustrating this using our previous NS game example, if a *Beliefs* participant expects sufficiently many *Actions* subjects select the number 5, she also believes sufficiently many select 14, 11 and 8. When we consider how a *Beliefs* participant anticipates the relative frequencies of various thinking steps (i.e. how many boxes are placed on 14, 11, etc.), we find that histograms are much better described by normalized Poisson versus Uniform distributions. This resonates with the Poisson structure that Camerer et al. [2004] impose in their estimations. In terms of our "belief-tracking" data, we find that beliefs indeed follow stepwise reasoning: when a box is placed on a certain step of thinking, the subsequent box is likely to be placed on that same step or one step higher.

The remainder of the paper is organized as follows: Section 2 presents the experiment, Section 3 discusses the results, Section 4 mentions some related literature and Section 5 concludes.

2 The Experiment

2.1 Number Selection (NS) Games

We carefully design Number Selection (NS) games for *Actions* participants to play. Before describing NS games in detail, we discuss a design challenge to overcome if we seek to cleanly identify Step k thinking in the *Beliefs* treatment. Suppose *Actions* subjects were to play games where idiosyncratic Step k reasoning would generate a variety of best responses. If a *Beliefs* subject thinks *Actions* participants are idiosyncratic Step k thinkers, she will expect *Actions* data to be thinly scattered. This would make it difficult for us to distinguish such a highly strategic *Beliefs* participant from another who randomly constructs a histogram non-strategically. Using a game with a dominant strategy would solve this issue but would introduce the "opposite" problem since it would merge all Step $k \geq 1$ actions, not allowing us to distinguish a *Beliefs* participant who only anticipates Step 0 and Step 1 *Actions* participants from another who believes *Actions* subjects engage in additional thinking steps.

We thus design NS games so that Step k predictions are neither too diluted nor too concentrated. They satisfy the following property: a Step k thinker–for any risk attitudes and any beliefs of how Step 0 through $k-1$ types are distributed–should *only* select actions consistent with the Level k model from Stahl and Wilson [1994] and Nagel [1995].[8] (See Observation 4.)

[8] This assumes a Level 0 and Step 0 action of choosing the upper bound in our NS games, which we justify in the ensuing paragraphs.

The Level k action in a game is calculated by taking the best-response to the Level $k-1$ action; Level 0 players are nonstrategic like Step 0 thinkers. This ensures that if a *Beliefs* subject thinks that *Actions* participants are all highly heterogeneous Step k reasoners, she should only expect *Actions* subjects to select Level k actions. The Level 0 through 4 actions in the particular NS games we use are all distinct.[9]

In a generic NS game, g, a player i and her opponent simultaneously select integers, n_i and n_{-i}, respectively from a common range $R_g = \{1, 2, \ldots, UB_g\}$, where UB_g is the game's upper bound. Player i earns n_i points automatically for selecting n_i. If n_i is exactly D_g less than n_{-i}, where D_g is g's commonly known undercutting distance, then i earns $B_g > UB_g \times D_g$ *additional* points.[10] If $n_i = n_{-i}$, then player i earns $b_g \in (UB_g - 1, B_g - D_g)$ *additional* points.[11] This payoff function is shown in Equation 1.

$$\pi_i^g(n_i, n_{-i}) = n_i + \begin{cases} B_g \text{ if } n_i = n_{-i} - D_g \\ b_g \text{ if } n_i = n_{-i} \end{cases} \tag{1}$$

The Level k model provides clear behavioral predictions in NS games. We assign UB_g as the Level 0 (and Step 0) action,[12] giving rise to the Level k actions stated in Observation 1.

Observation 1. *In a Number Selection game, g, the Level k strategy is* $\max\{UB_g - k \times D_g, \operatorname{mod}(UB_g, D_g)\}$, *where* $\operatorname{mod}(x, y)$ *is the remainder from* $x \div y$.

Proof. *The Observation states that when undercutting a Level $k-1$ player is not possible, the Level k action coincides with the Level $k-1$ strategy. Thus, assume that an opponent selects $n_{-i} \leq D_g$ (so that undercutting is not possible). Then, $b_g > UB_g - 1$ implies $\pi_i^g(n_{-i}, n_{-i}) = n_{-i} + b_g > n_{-i} + UB_g - 1 \geq UB_g = \pi_i^g(UB_g, n_{-i})$. The Observation also states that when undercutting a Level $k-1$ player is possible, the Level k best response is to undercut. Assume an opponent selects $n_{-i} > D_g$ (so that undercutting is possible). Given that $b_g < B_g - D_g$, we have $\pi_i^g(n_{-i}, n_{-i}) = n_{-i} + b_g < n_{-i} + B_g - D_g = \pi_i^g(n_{-i} - D_g, n_{-i})$.*

While the Level k predictions in Observation 1 assume an upper bound Level 0, there may be alternative strategies that naive subjects would implement. For instance, a player with

[9] While we could have expanded our games such that the Level 0 through k actions in each NS game are all distinct for some $k > 4$, we focus on Levels 0 through 4 because prior evidence shows that play at higher levels drops off rather quickly. For example, Crawford and Costa-Gomes [2006] and Fragiadakis et al. [2016] classify substantially more Level 1 and Level 2 players in comparison to Level 3.

[10] The restriction that $B_g > UB_g \times D_g$ is needed for Observation 2.

[11] The restrictions that $b_g < B_g - D_g$ and $b_g > UB_g - 1$ are needed for Observations 1 and 3.

[12] Arad and Rubinstein [2012], who have a similar game, also designate the upper bound in their games as Level 0 since it is the action that maximizes a player's payoff if he does not form beliefs about his opponent and acknowledges only that his payoff function grants him n points for selecting n.

absolutely no understanding of the NS games may play uniformly at random. This Level 0 specification is actually quite common. We design our experiment such that this alternative Level 0 specification does not alter the Observation 1 predictions for $k \geq 1$, minimizing concerns about the explanatory power of Level k being driven by idiosyncratic Level 0 specifications (see Hargreaves Heap et al. [2014]).

Observation 2. *For each Number Selection game, g, selecting $UB_g - D_g$ is the unique action that maximizes one's expected points against an opponent who plays uniformly at random.*

Proof. *Against a uniform random opponent, player i's points in game g are:*

$$\Pi_i^g(n_i) \equiv \frac{1}{UB_g}\sum_{j=1}^{UB_g}\pi_i^g(n_i,j) = n_i + \frac{b_g}{UB_g} + \begin{cases} B_g/UB_g \text{ if } n_i \leq UB_g - D_g \\ 0 \text{ if } n_i > UB_g - D_g \end{cases} \tag{2}$$

From Equation 2, we see that UB_g and $UB_g - D_g$ are the only local maxima. Given that $B_g > UB_g \times D_g$, we have $B_g/UB_g > D_g$, which implies that

$$(UB_g - D_g + b_g/UB_g) + B_g/UB_g > (UB_g - D_g + b_g/UB_g) + D_g.$$

Therefore, $\Pi_i^g(UB_g - D_g) > \Pi_i^g(UB_g)$, rendering the Level 1 prediction (and hence all Level $k \geq 1$ actions) from Observation 1 unchanged.

In Observation 1, we reason that if a player cannot undercut his opponent, her best response is to match the other player. This proves Observation 3.

Observation 3. *A Number Selection game with undercutting distance D_g has D_g pure strategy Nash equilibria: $\{(1,1),(2,2),\ldots,(D_g,D_g)\}$.*

Before moving to the Results section, we formally define the Step k model used in this paper, which is similar to those from Stahl and Wilson [1995] and Camerer et al. [2004]. Step 0 and Step 1 players are identical to Level 0 and Level 1 types, respectively. For $k > 1$, a Step k thinker anticipates Step 0 through Step 1 players (with any proportions). Observation 4 outlines the behavioral predictions of Step k thinkers in NS games.

Observation 4. *Suppose a player i (in a Number Selection game) believes the chance of facing a Level k action is $\alpha_k \in [0,1]$ for all $k \geq 0$. If $\sum_{h=0}^{\infty}\alpha_h = 1$, her best response (given her beliefs) is some Level k action.*

Proof. *Suppose player i selects some <u>non</u>-Level k action, n_i, in an NS game, g. Her payoff will be n_i since there will be no way of receiving the b_g or B_g bonuses. As a result, deviating to $UB_g > n_i$ is profitable.*

2.2 Box Arrangement (BA) Tasks

We create Box Arrangement (BA) tasks to elicit beliefs. The objective in a BA task t_g that corresponds to a game g (with finite strategy set S) is to distribute N_g boxes across a horizontal axis made up of the pure strategies in g. Let h denote the histogram generated, where h_s is the number of boxes that placed on pure strategy s. Let l denote the histogram of actual behavior by some population of N_g individuals in game g, where l_s is the number of times s was chosen. The points earned in t_g are equal to the number of boxes that are overlapping when h and l are superimposed. Equation 3 gives the payoffs function mathematically.

$$\pi(h,l) = \sum_{j=1}^{|S|} \min\{h_j, l_j\} \tag{3}$$

Thus, if h perfectly coincides with l, N_g points are earned. If h does not coincide with l whatsoever, 0 points are earned. Suppose a risk neutral individual i has a belief system α for a NS game, g.[13] In other words, she believes the chance a randomly chosen opponent will select $s \in R_g$ is $\alpha_s \in [0,1]$. We thus assume that $\sum_{h=1}^{UB_g} \alpha_h = 1$. The payoffs in Equation 3 incentivize boxes to be arranged in a manner that approximates the beliefs system α. We illustrate this in Figure 1 with an example where $R_g = \{1,2\}$ and $N_g = 2$. An individual can create h, h' or h''. Let $f \in [0,1]$ denote the fraction of the population that is believed will select **2**. It can be shown that the expected payoffs of building histograms h, h' and h'' (as functions of f) are π, π' and π'', respectively. Furthermore, it can be shown that (i) for $f \leq 1 - \sqrt{2}/2$, $\pi(f) \geq \max\{\pi'(f), \pi''(f)\}$. Thus, for sufficiently small f, one is best off by placing both boxes on **1**. Symmetrically, for sufficiently large f, one's payoff-maximizing histogram involves both boxes placed on **2**; in other words, (ii) $\pi''(f) \geq \max\{\pi(f), \pi'(f)\}$ for $f \geq \sqrt{2}/2$. If f takes on intermediate values, it is optimal to place a box on **1** and a box on **2**; it can be shown that (iii) for $f \in [1 - \sqrt{2}/2, \sqrt{2}/2]$, we have $\pi'(f) \geq \max\{\pi(f), \pi''(f)\}$.

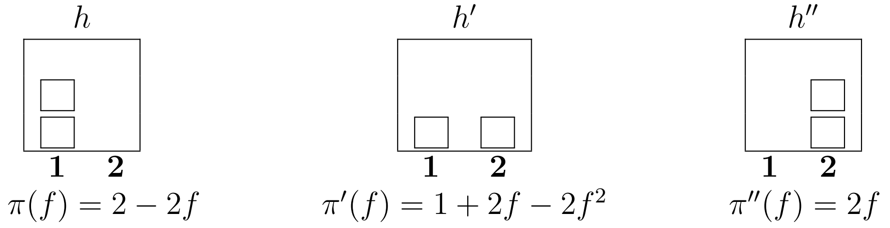


Figure 1.—Histograms and Payoff Functions for $R_g = \{1,2\}$ and $N_g = 2$.

[13] We control for risk attitudes in our experiment, as described in Section 2.3.

2.3 Treatments and Procedures

Experimental subjects are undergraduates and graduate students, recruited using ORSEE (Greiner et al. [2004]). Participants interact via a network of computers linked by z-Tree (Fischbacher [2007]) at the Economics Research Laboratory in TAMU's Department of Economics. Two 20-subject sessions, each lasting approximately 1 hour, make up the *Actions* treatment. The 81 participants in the *Beliefs* treatment are spread across 5 sessions of 14, 13, 18, 18 and 18 subjects, each taking roughly 2 hours. Average earnings are \$28.39 and \$54.07 in the *Actions* and *Beliefs* treatments, respectively, including a \$5.00 show-up payment.

Subjects in the *Actions* treatment play the 11 Number Selection games shown in Table 1, where each game has $B_g = 100$ and $b_g = 35$. The games we select are determined using a list of design criteria. First, each game's lower bound is 1 and upper bounds is no more than 32 due to spatial constraints on subjects' computer screens. Second, we impose that $D_g \in \{3, 4\}$ so that the Level k model only captures a fraction of the strategy space. Third, we make it so that $\{UB_g - k \times D_g\} \Cap \{1, D_g\} = \emptyset$ for all k, which prevents the Nash equilibrium reached via the Level k model from coinciding with the "lower bound equilibrium" of (1,1) or the "efficient equilibrium" of (D_g, D_g). Subjects may play "1" as a rule of thumb or as a focal final result of repeated (but not iteratively performed) undercutting. Participants who are "true" equilibrium players that do *NOT* arrive at equilibrium via converged Level k reasoning may be expected to coordinate on (D_g, D_g), the most efficient equilibrium. Lastly, $UB_g \geq 14$ ensures that Levels 0 through 4 are all distinct (though Level 4 may coincide with an equilibrium).

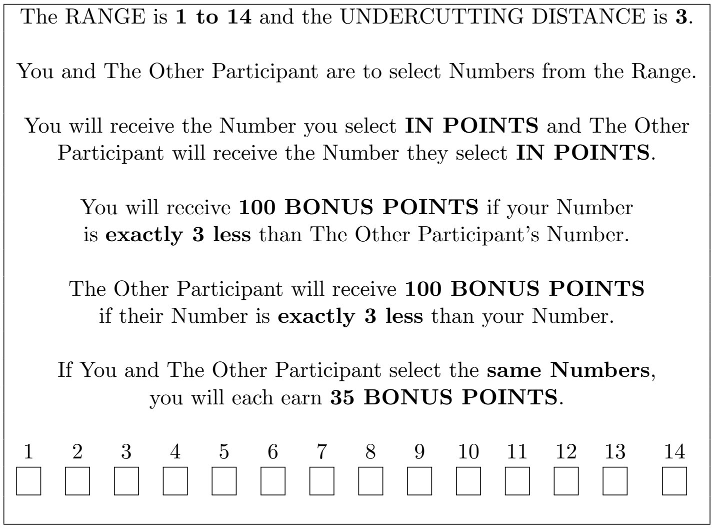
Game Number	1	2	3	4	5	6	7	8	9	10	11
UB_g	14	17	20	23	26	29	32	18	19	22	23
D_g	3							4			

Table 1.—The 11 Number Selection Games used in the experiment

Actions Participants in a session are paired randomly and anonymously at the beginning of their session; matchings are fixed for all 11 games. NS games are presented to participants as in Figure 2. Each subject in the *Actions* treatment views a NS game from the same perspective: a subject is addressed as "You" and her opponent is referred to as "The Other Participant". In the *Beliefs* treatment, participants are shown the same 11 NS games from Table 1, presented as in Figure 2, except that the two players are referred to as "Jack" and "Jill".[14]

For a game, g, a *Beliefs* subject performs a BA task, t_g. When performing a BA task, a subject's screen initially shows a large, empty, rectangular area that has the game's range

[14] Importantly, *Beliefs* participants are not shown any decisions made by *Actions* subjects.



The RANGE is **1 to 14** and the UNDERCUTTING DISTANCE is **3**.

You and The Other Participant are to select Numbers from the Range.

You will receive the Number you select **IN POINTS** and The Other Participant will receive the Number they select **IN POINTS**.

You will receive **100 BONUS POINTS** if your Number is **exactly 3 less** than The Other Participant's Number.

The Other Participant will receive **100 BONUS POINTS** if their Number is **exactly 3 less** than your Number.

If You and The Other Participant select the **same Numbers**, you will each earn **35 BONUS POINTS**.

1	2	3	4	5	6	7	8	9	10	11	12	13	14
□	□	□	□	□	□	□	□	□	□	□	□	□	□

Figure 2.—How subjects view Number Selection Games

of numbers, $R_g = \{1, 2, \ldots, UB_g\}$, strung along its lower horizontal edge. A green, upwards-pointing, arrow button rests underneath each number in R_g. Clicking the green arrow button under a number $n \in R_g$ adds a blue box above n. When one or more blue boxes sit above n, a red, downwards-pointing, arrow button shows below n's green arrow button. Clicking n's corresponding red arrow button removes the top-most blue box that sits above n. A subject can click green and red arrow buttons without any restrictions, allowing her to freely build (and revise) her histogram until she is ready to submit it. Between the game's range and the green arrow buttons, a counter tracks how many boxes sit above each number in R_g.

Subjects click the arrow buttons to allocate their 20 blue boxes across the strategies in R_g to express how they believe the 20 participants from a previously run *Actions* session made their choices in g. For instance, if a *Beliefs* participant believes that two *Actions* subjects chose $1 \in \{R_g\}$, she would place two blue boxes above the number "1". Figure 3 shows the example Box Arrangement Task (generated at random) that we gave subjects in the *Beliefs* treatment instructions. To explain the incentives in Equation 3, we also provide subjects with a Dot Arrangement of hypothetical play by 20 *Actions* subjects. The Dot Arrangement is printed on standard white paper while the example Box Arrangement from Figure 3 is printed on a plastic transparency. This allows subjects to overlay the two arrangements to generate an image that looks like that shown in Figure 4. Subjects are then told that they earn p points for having p boxes that overlap with dots. While the points earned do not depend on the order in which boxes are arranged, we record the order in which *Beliefs* participants arrange the boxes that make up their final histograms.

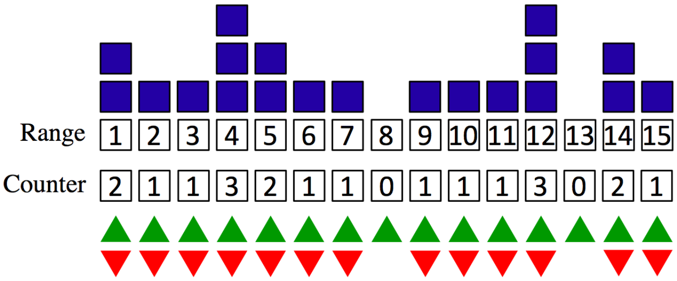


FIGURE 3.—a histogram for a fictitious Box Arrangement Task

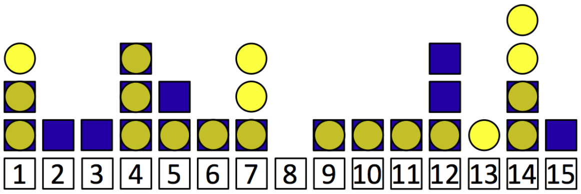


FIGURE 4.—a fictitious Dot Arrangement overlaid with the histogram from Figure 3

In both treatments, instructions are read out loud and subject comprehension is reinforced using on-screen understandings tests. A subject cannot proceed past a question until it is answered correctly.[15] The incentivized portion of each treatment is initiated only once all subjects complete the treatment's understandings test. Subjects receive no feedback whatsoever between NS games and BA tasks until all experimental decisions have been made.

To mitigate concerns about risk attitudes influencing decision-making, the point payoffs in each NS game and BA task are converted to money (at the end of the experiment) using separate and independently run binary lotteries (Roth and Malouf [1979]). If a subject earns p points in a NS game, the corresponding lottery pays \$5 with probability $p/150$ and \$1 with probability $1-p/150$.[16] If a subject earns p points in a BA task, she earns \$5 with probability $p/20$ and \$1 otherwise.

[15] Incorrect answers are met with a prompt asking the subject to try again. If a subject is stuck, she can quietly ask for individual assistance from the experimenter.

[16] There are two main reasons for using binary lotteries in NS games. First, doing so gives rise to Observation 2. Second, a linear exchange rate of points to dollars is problematic for subject payments since a participant may earn over one hundred times as many points as another in a NS game.

2.3.1 Relative Performance (RP) Questions

After a *Beliefs* subject i completes all 11 BA tasks, she performs 11 corresponding Relative Performance (RP) questions. For the RP question q_g corresponding to BA task t_g and NS game g, subject i is shown g *as well as* the histogram she constructed in t_g. Participants are not shown any histograms that were made by any other *Beliefs* subjects. Subject i is informed of the number of subjects in her lab session and is asked to estimate how many participants in her session she believes earned *strictly more* points than she did in t_g. For q_g, subject i earns \$5 for a correct answer and \$1 otherwise. The RP questions are intended to estimate subjects' levels of confidence in the histograms created in the BA tasks.

2.3.2 Bomb Risk (BR) Decisions

After subjects in the *Beliefs* treatment perform their RP questions, they make a Bomb Risk (BR) decision (adapted from Crosetto and Filippin [2013]) as a quick measure of their risk attitudes. The BR decision is very straightforward. There are 100 treasures chests, one of which contains a bomb. The subject chooses, m, the number of (randomly picked) chests it would like the computer to open. If the bomb is in an opened chest, the subject earns nothing (which occurs with a $m/100$ chance). Otherwise, the bomb-filled chest is *not* opened and the subject earns $m/10$ dollars.

3 Results

3.1 Aggregate Behavior

3.1.1 *Actions* treatment

Before analyzing data from the *Beliefs* treatment, we briefly summarize *Actions* behavior. Observation 4 states that, for a wide range of beliefs, individuals' best responses in Number Selection (NS) games coincide with the Level k actions. We thus expect a substantial proportion of behavior in NS games to fall on Level k actions; our data confirm this (see Result 1 and Table 2).

Value of k	0	1	2	3	4	total
# of choices	60	129	89	28	4	310
% of choices	13.6	29.3	20.2	6.4	0.9	70.4

TABLE 2.—Level k action frequencies in *Actions* treatments

Result 1. *In the* Actions *treatment, 70.4% of decisions coincide with the Level 0 through 4 actions; this high frequency is expected given Observation 4.*

3.1.2 *Beliefs* treatment

Roughly thirty percent of *Actions* behavior did not overlap with Levels 0 through 4 (Result 1). This provides a rough estimate on the proportion of non-belief-based decision-making in the population from which we draw our experimental subjects.[17] We thus expect a baseline rate of roughly thirty percent of *Beliefs* data to not overlap with Levels 0 through 4. There is another source of noise in the *Beliefs* data, however. Suppose that a participant i believes that opponents guess uniformly at random. If i participates in the *Actions* treatment, she will guess Level 1, the most commonly selected Level k strategy in the *Actions* treatment. In the *Beliefs* treatment, however, she will spread out her boxes. Taken altogether, we expect *Beliefs* data to be more diluted than *Actions* behavior. Table 3 and Result 2) confirm this hypothesis.

Value of k	0	1	2	3	4	total
# of choices	2305	2631	1635	884	597	8052
% of choices	12.9	14.8	9.2	5.0	3.4	45.2

Table 3.—Level k action frequencies in *Beliefs* treatments

Result 2. *In the* Beliefs *treatment, 45.2% of decisions coincide with the Level 0 through 4 actions. (This percentage is higher in the* Actions *treatment, but this is to be expected.)*

3.1.3 Stepwise reasoning is more than an "as if" theory of beliefs

In addition to obtaining *Beliefs* participants' choices, we record the order in which they arranged their boxes in the BA tasks, allowing us to examine how individuals transition between Level k and k' actions for $k, k' \in \{0, 1, 2, 3, 4\}$. Because the instructions focus on explaining the payment method for overlapping boxes and the order of the arrangement of boxes is not incentivized, recording the order is minimally invasive. We consider all 25 (5×5) pairs of transitions, determining whether a box placed in one category predicts where the next box will be placed.

To do so, we construct several 5×5 transition matrices. The first is the *empirical* transition matrix, E, whose (i, j) entry corresponds to the number times two consecutive boxes are placed first on the Level $i-1$ strategy and then on the Level $j-1$ action. The second is an analogously

[17] It is possible that some of the non-Level k choices in the *Actions* treatment are best responses to alternative beliefs. Similarly, some of the Level k choices may arise by chance.

defined *random* matrix, R, that assumes *Beliefs* participants arrange their boxes in random orders (but still produce the same final empirical histograms). The third is the *normalized* transition matrix, N, where $N = E - R$.[18]

In Table 4 shows the matrix N. Below element $N(i, j)$ we report the p-values of a one-tailed binomial distribution test of the null hypothesis that $N(i, j) = 0$ given the final empirical distribution. Cells are colored if $N(i, j) > 0$ and has $p < 0.01$. First, we note that the diagonal cells (in blue) are all positive and significant, indicating that a box placed in a category is likely to be preceded by a box in that same category (Result 3).

	L0	*L1*	*L2*	*L3*	*L4*
L0	1010.9 (0.000)	171.4 (0.000)	-145.5 (0.000)	-106.6 (0.000)	-50.9 (0.000)
L1	-70.6 (0.000)	1032.8 (0.000)	97.5 (0.000)	-108.5 (0.000)	-54.2 (0.000)
L2	-109.3 (0.000)	-124.3 (0.000)	490.1 (0.000)	130.5 (0.000)	-30.9 (0.000)
L3	-75.8 (0.000)	-99.3 (0.000)	-21.5 (0.005)	170.7 (0.000)	86.5 (0.000)
L4	-43.2 (0.000)	-56.3 (0.000)	-27.2 (0.000)	6.7 (0.089)	64.4 (0.000)

TABLE 4.—Level k action frequencies in *Actions* treatments

Result 3. *For all nine categories, if a box is placed in a category, C, the following box is likely to be placed in C. (See the blue cells in Table 4.)*

If Step k thinking is more than an "as if" representation of behavior, we would expect individuals to express Level k beliefs before they express Level $k + 1$ beliefs. The yellow cells in Table 4 show that transitions where a Level $k + 1$ box is placed immediately after a Level k are significant for $k \in \{0, 1, 2, 3\}$ (Result 4).

Result 4. *For $k \in \{0, 1, 2, 3\}$, if a box is placed on a Level k action, the following box is likely to be placed on the Level $k + 1$ action. (See the yellow cells in Table 4.)*

[18] A subject is allowed to add and remove boxes without any restrictions. When we record the order in which boxes are placed, we only consider the 20 boxes that make up the histogram that is finally submitted.

3.2 Individual Analysis

3.2.1 Classification of *Beliefs* participants as Step k thinkers

We begin investigating Step k thinking in the *Beliefs* treatment by determining the Level k action(s) for $k \in \{0, 1, 2, 3, 4\}$ that a *Beliefs* subject anticipates being played by *Actions* participants. For a subject, i, we compute a vector, $v_i = (v_i^1, \dots, v_i^9)$, where (i) v_i^{2k+1} is the total number of boxes that i places on the Level k action and for $k \in \{0, 1, 2, 3, 4\}$ and (ii) v_i^{2k} is the total number of boxes that i places *between*[19] the Level $k-1$ and Level k action and for $k \in \{1, 2, 3, 4\}$. For a participant to forecast a Level k action, we first require that she place significantly more boxes on the Level k strategy in comparison to a subject playing uniformly at random (Condition 1).

Condition 1. *A necessary condition for identifying a* Beliefs *subject i as believing* Actions *participants play the Level k action (for any $k \in \{0, 1, 2, 3, 4\}$) is $v_i^{2k+1} \geq 16$.*[20]

We then require that Level k actions are forecasted sufficiently more than numbers in their vicinities.[21] Specifically, we create a weighted vector, $w_i = (w_i^1, \dots, w_i^9)$, where $w_i^{2k+1} = v_i^{2k+1}/11$ for $k \in \{0, 1, 2, 3, 4\}$ and $w_i^{2k} = v_i^{2k}/26$ for $k \in \{1, 2, 3, 4\}$. These normalizations account for the fact that while there are only 11 strategies across the 11 games corresponding to a particular Level k action (for $k \in \{0, 1, 2, 3, 4\}$), there are 26 strategies across the 11 games between the Level $k-1$ and Level k actions (for $k \in \{1, 2, 3, 4\}$).[22]

Condition 2. *A necessary condition for documenting a* Beliefs *subject i as believing* Actions *participants play the Level k action is*

- $w_i^1 \geq 2w_i^2$ *if* $k = 0$,
- $w_i^{2k+1} \geq w_i^{2k} + w_i^{2(k+1)}$ *if* $k \in \{1, 2, 3\}$, *and*
- $w_i^9 \geq 2w_i^8$ *if* $k = 4$.[23]

[19] For instance, in NS game 1, where $D_1 = 3$ and $UB_1 = 14$, the Level 0 and Level 1 actions are 14 and 11, respectively; thus, the numbers *between* the Level 0 and Level 1 actions are 13 and 12.

[20] The threshold of 16 is reached via simulations. We generate 10,000 artificial *Beliefs* subjects who allocate their boxes uniformly at random and find that 95% of subjects place fewer than 16 total boxes in total across the upper bounds of the 11 games.

[21] Consider, for instance, a subject who places 2 boxes on each of the 10 largest numbers in each NS game. It is not clear that this subject is a Step k thinker, yet he will have 22 boxes on the Level 0, 1 and 2 actions and hence meet Condition 1 for these three strategies.

[22] $D_g = 3$ for $g \leq 7$ and $D_g = 4$ for $g > 7$. Since there are 2 and 3 numbers between the Level $k-1$ and Level k actions for $k \in \{1, 2, 3, 4\}$ in games with $D_g = 3$ and $D_g = 4$, respectively, this yields $7 \times 2 + 4 \times 3 = 26$ total numbers across all histograms.

[23] We do not require that the mass placed on Level 4 is greater than the sum of (i) the mass placed between

Definition 1 states that if a subject's stated beliefs meet Conditions 1 and 2 for a Level k strategy, then she believes that strategy is played by *Actions* participants. For example, for subject i to believe in the Level 3 action, it is necessary that i place $b \geq 16$ boxes on the Level 3 action (totaled over all her histograms) *and* that $b/11 \geq s/26$, where s is the number of boxes placed between the Level 2 and Level 3 actions *and* between the Level 3 and Level 4 actions (totaled over all her histograms).

Definition 1. *For any $k \in \{0, 1, 2, 3, 4\}$, a* Beliefs *participant i* ***believes*** Actions *participants play the Level k strategy if Conditions 1 and 2 are met for that Level k strategy.*

Using Definition 1, we can check whether *Beliefs* participants build histograms that are consistent with the predictions of the CH model. In other words, we can define Step k thinking (Definition 2) and check for *Beliefs* participants who employ it (Result 5).

Definition 2. *A* Beliefs *participant is a* ***Step k*** *thinker, for some $k \in \{1, 2, 3, 4, 5\}$, if she believes* Actions *participants play the Level k' strategies (according to Definition 1) for all $k' \in \{0, \ldots, k-1\}$ and there is no larger $k \in \{2, 3, 4, 5\}$ satisfying this property.*

Result 5. *We classify 40.7% of* Beliefs *participants as Step k thinkers for $k \in \{2, 3, 4, 5\}$, supporting the assumptions in Stahl and Wilson [1995] and Camerer et al. [2004].*

Value of k	1	2	3	4	5	total
# of participants	6	15	12	3	3	39
% of subjects	7.4	18.5	14.8	3.7	3.7	48.1

TABLE 5.—Classification of types in *Beliefs* treatments

To see Step k thinking more strikingly, we plot Step k thinkers' w_i vectors. Figure 5 does this for the three Step 5 thinkers and three Step 4 players classified in Table 5. Notice that a panel has nine categories on the horizontal axis, representing the nine coordinates of w_i. In a panel constructed by a subject, i, the black bars denote the actions anticipated by i. For example, subject 1 is a Step 5 player, and hence, Level 0 (L0) through Level 4 (L4) are black. Bars on Level k actions that are *not* anticipated are in gray, such as the L4 bar in subject 5's panel. White bars represent the mass *between* Level k actions.

Examining the panels in Figure 5, we see that the black bars meet Condition 1: each black bar exceeds $16/11 \approx 1.45$. (We see that the gray bars do not meet this condition.) We can

Level 3 and 4 and (ii) the mass placed between Level 4 and 5 because, in some games, there are different numbers of pure strategies in the regions described by (i) and (ii). For example, in the first NS game, the actions 3 and 4 form the region described by (i) while the number 1 is the only action in the region described by (ii).

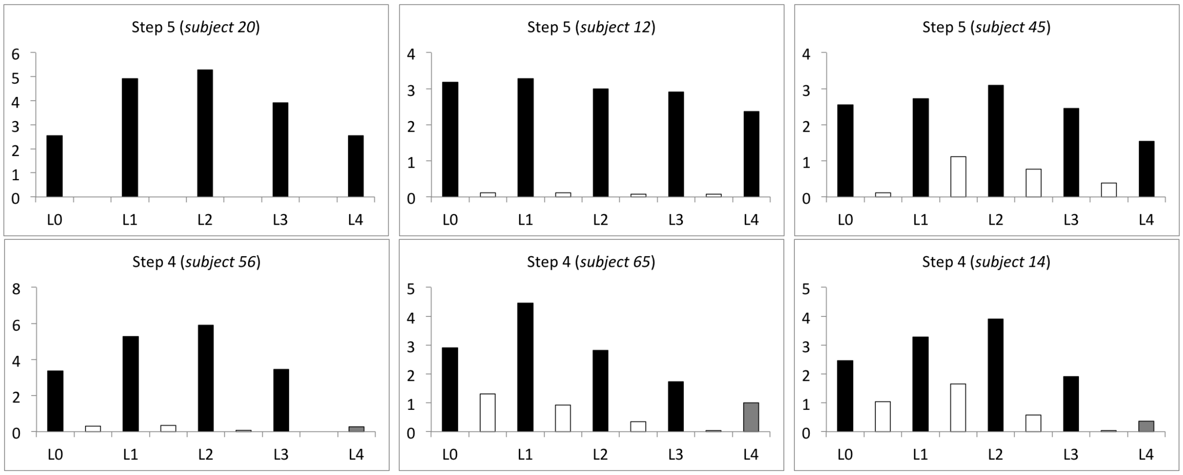


FIGURE 5.—The three **Step 5** thinkers (top) and three **Step 4** players (bottom) classified in the *Beliefs* treatment.

also readily see that Condition 2 is met based on how disproportionately taller the black bars are than the white bars. Figure 6 plots the w_i vectors for a Step 3, 2 and 1 thinker.

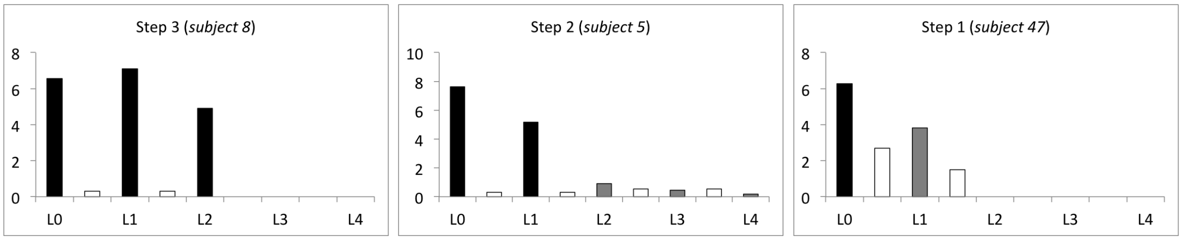


FIGURE 6.—Plots of a **Step 3** thinker, a **Step 2** player and a participant classified as **Step 1**.

3.2.2 How do Step k thinkers anticipate the relative frequencies of lower types?

To estimate data sets using CH, Camerer et al. [2004] find the value of τ such that $f(k) = e^{-\tau}\tau^k/k!$ best approximates the proportion of Step k thinkers in the population. This Poisson specification leads to a simple expression of *relative* proportions of Step k thinkers: $f(k+1)/f(k) = \tau/(k+1)$. Camerer et al. [2004] model a Step k thinker as having these same relative beliefs over Step 0 through Step $k-1$ players. With this assumption, we could estimate a τ for each *Beliefs* participant. This, however, would require us to make assumptions about a *Beliefs* subject's beliefs over not only the actions of *Actions* subjects (which we obtain), but also over the beliefs held by *Actions* participants (which we do not obtain). Inferring beliefs over beliefs is difficult. For example, consider a *Beliefs* participant predicting play in Game

Number 1 (whose Level k actions are 14, 11, 8, 5 and 2). If her histogram consists of ten boxes on 14 and ten on 11, we do not know if she solely anticipates Level 0 and Level 1 players or if she also envisions some Step 2 thinkers who play 11 as a best response to their beliefs.[24]

Accordingly, we restrict our analysis to what we observe: *Beliefs* participants' predictions of the play by *Actions* subjects. We adapt the Poisson specification from Camerer et al. [2004] to our environment and check whether *Beliefs* subjects expect the various thinking steps in the *Actions* treatment to be distributed according to normalized Poisson distributions. To approximate the histogram of a Step k subject i with a normalized Poisson distribution, we find the $\widehat{\lambda_i}$ that minimizes the Euclidean distance between her empirical relative beliefs over adjacent thinking steps and the corresponding ratios given by the Poisson specification. This Poisson Distance (P^D) is given in Equation 4, where $v_i = (v_i^1, \ldots, v_i^k)$ and v_i^t is the total number of boxes placed on Level $t-1$ actions across all Box Arrangement tasks.

$$\widehat{\lambda_i} = argmin_{\lambda_i} P^D(v_i, \lambda_i), \text{ where } P^D(v_i, \lambda_i) = \sqrt{\sum_{t=1}^{k-1} \left(\frac{v_i^{t+1}}{v_i^t} - \frac{\lambda_i}{t} \right)^2} \tag{4}$$

Once we calculate $\widehat{\lambda_i}$ for a subject, we would like to get a sense of how well it approximates her stated beliefs. To do so, we normalize the vector v_i to $w_i = v_i / (\sum_{t=1}^k v_i^t)$ and define a vector $f_i = (f_i^1, \ldots, f_i^k)$ where $f_i^t = e^{-\widehat{\lambda_i}} \widehat{\lambda_i}^{t-1}/(t-1)!$ for $t = 1, \ldots, k$. We then normalize f_i to $g_i = f_i / (\sum_{t=1}^k f_i^t)$ and compute the mean-squared error of between w_i and g_i. This is given by the Poisson Error (P^E) in Equation 5.

$$P^E(w_i, \widehat{\lambda_i}) = \sum_{t=1}^{k} (w_i^t - g_i^t)^2 \tag{5}$$

As a point of reference, we compute an analogous Uniform Error (U^E) in Equation 6.

$$U^E(w_i, \widehat{\lambda_i}) = \sum_{t=1}^{k} (w_i^t - \bar{w_i})^2 \text{ where } \bar{w_i} = \frac{\sum_{t'=1}^k w_i^{t'}}{k} \tag{6}$$

Table 6 lists the values of $P^E(w_i, \widehat{\lambda_i})$ and $U^E(w_i, \widehat{\lambda_i})$ for the 18 participants classified as Step k thinkers for $k \geq 3$. Subjects are ordered as their values of $[P^E(w_i, \widehat{\lambda_i})]/[P^E(w_i, \widehat{\lambda_i}) + U^E(w_i, \widehat{\lambda_i})]$ increase. This ratio is less than 1/2 for all but the last two subjects in the table, meaning that $P^E(w_i, \widehat{\lambda_i}) < U^E(w_i, \widehat{\lambda_i})$ for all but these last two participants (Result 6).

Result 6. *Of the 18 Step k thinkers having $k \geq 3$, 16 are better described by the Poisson model*

[24] In fact, a Step 2 player who expects Step 0 and 1 players to occur with frequencies α and $1-\alpha$ have a best response of 11 if $\alpha \geq 62/165 \approx .38$.

Subject i	λ_i	$P^E(w_i, \widehat{\lambda}_i)$	$U^E(w_i, \widehat{\lambda}_i)$	$\frac{P^E(w_i,\widehat{\lambda}_i)}{P^E(w_i,\widehat{\lambda}_i)+U^E(w_i,\widehat{\lambda}_i)}$	Step k
49	1.2604	0.0000	0.0106	0.0026	3
67	0.9816	0.0002	0.0319	0.0057	3
37	1.02066	0.0003	0.0213	0.0121	3
6	1.1298	0.0004	0.0223	0.0188	3
65	1.5146	0.0016	0.0266	0.0558	4
20	2.1097	0.0011	0.0178	0.0606	5
8	1.1838	0.0013	0.0075	0.1479	3
51	0.8733	0.0161	0.0916	0.1489	3
56	1.7852	0.0028	0.0153	0.1556	4
75	1.3277	0.0085	0.0373	0.1866	3
43	1.5109	0.0014	0.0059	0.1939	3
38	1.7383	0.0036	0.0137	0.2094	3
68	1.8376	0.0140	0.0376	0.2709	3
14	1.6452	0.0069	0.0175	0.2837	4
73	1.5230	0.0284	0.0617	0.3157	3
29	2.4775	0.0621	0.0737	0.4573	3
45	1.7412	0.0148	0.0086	0.6333	5
12	1.7892	0.0235	0.0023	0.9096	5

Table 6.—Poisson and Uniform Errors for Step 3, 4 and 5 participants

in comparison to a Uniform specification.

We visually display the relative success of the Poisson specification by plotting the w_i and g_i vectors along with $\bar{w}_i$ vectors for the 18 Step k thinkers having $k \geq 3$ in Figure 7. In the plots, w_i, g_i and $\bar{w}_i$ are labelled A for Actual, P for Poisson and U for Uniform, respectively. Subjects are ordered in Figure 7 as they are listed in Table 6.

3.2.3 Analyzing unclassified participants

In addition to our 39 Classified (C) participants, we have 42 Unclassified (U) subjects in the *Beliefs* treatment. Given that we find a non-negligible number of U participants, we would like to devote some attention to understanding their behavior. We begin by asking whether they anticipate deterministic types that we do not understand with existing models. Several findings suggest that this is unlikely. If we partition the columns built in the Box Arrangement (BA) tasks by height, we find that the taller a column is, the more likely it is built on a Level k action. This is shown in Figure 8. The x-axis indicates column height; the y-axis shows the percentage of columns (of that height) that are constructed on a Level k action. In other words, the blue dot corresponding to 8 along the x-axis has a height of just over 80%, meaning

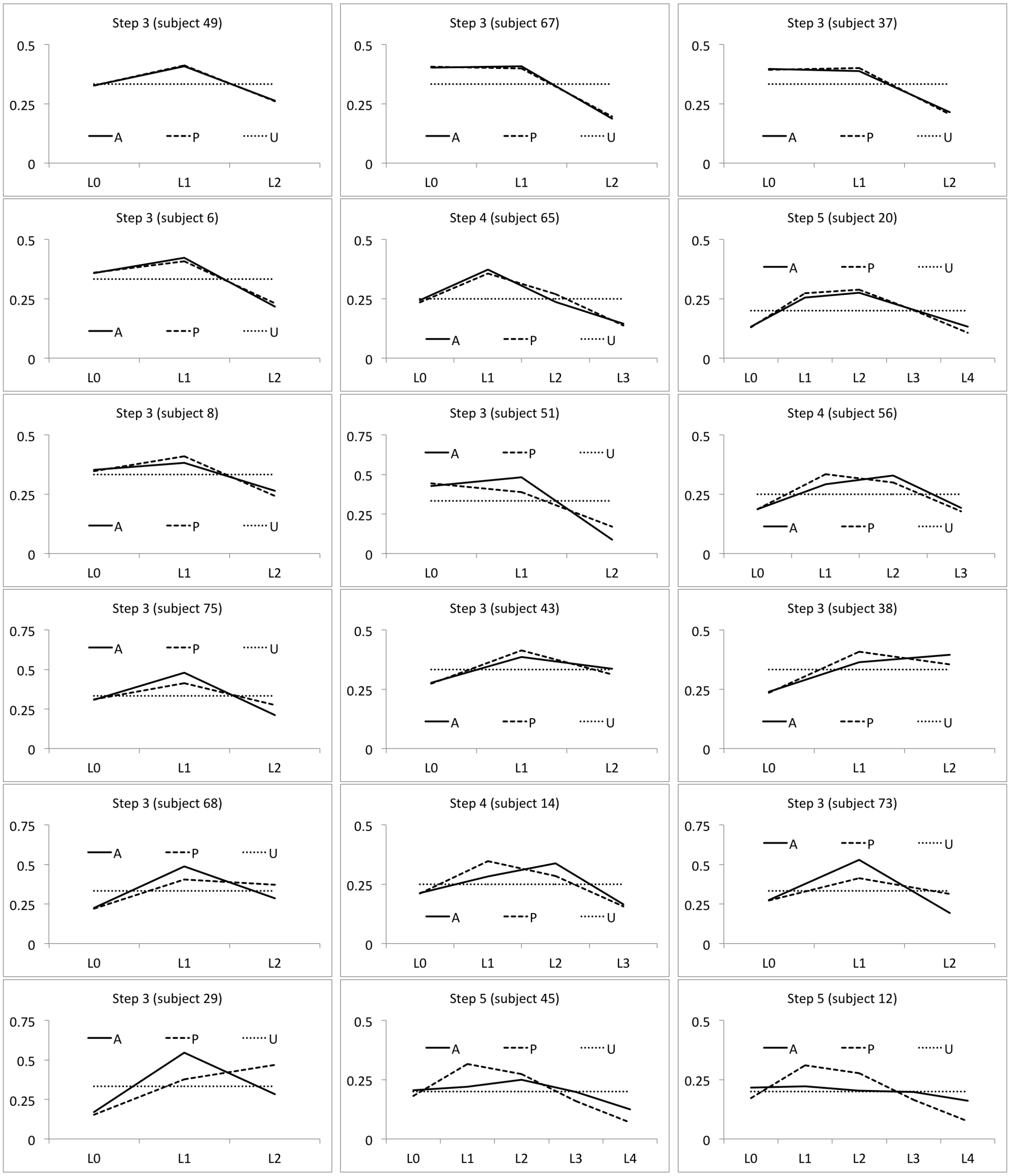


FIGURE 7.—Plots of w_i, g_i and $\bar{w}_i$, labelled A for Actual, P for Poisson and U for Uniform, respectively. The plots are for the 18 Step k thinkers having $k \geq 3$.

that roughly 80% of the columns that are exactly 8 boxes tall are built on Level k actions. Notice that each dot with an x-value of 10 or greater has y-value Level k frequency of 100% (Result 7). It thus appears unlikely that U participants are anticipating deterministic types that our existing models fail to capture.

Result 7. *If a column from a histogram in the* Beliefs *treatment consists of 10 or more boxes, that column rests on top of a Level k strategy.*

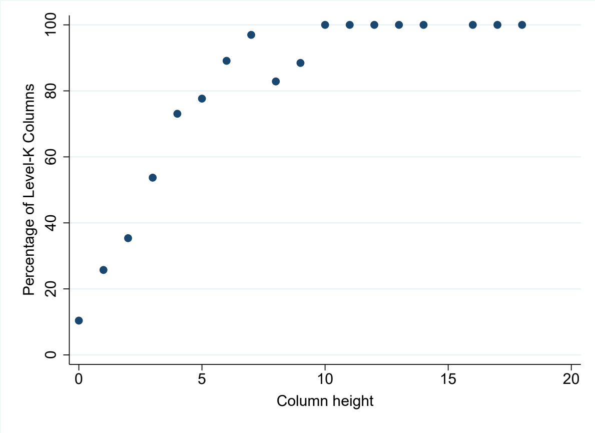


FIGURE 8.—Percentage of Level k beliefs according to belief strength

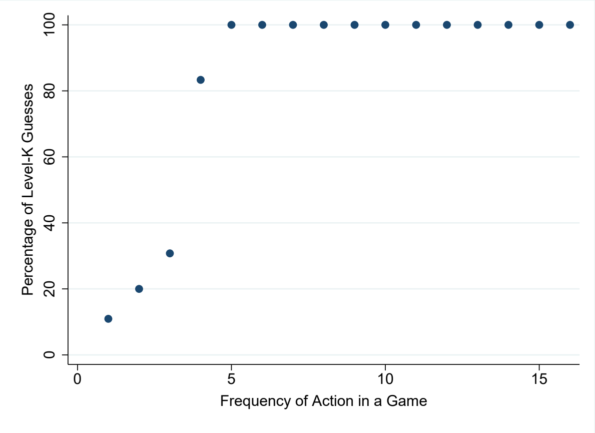


FIGURE 9.—Percentage of Level k actions vs. action frequency

We obtain similar results when we consider behavior in the *Actions* treatment. Figure 9 organizes *Actions* behavior according to how many participants play the same strategy in a given NS game (x-axis). Empirically, we find that there are at most 16 subjects (out of 40) who select the same action in an NS game, g. The y-axis in Figure 9 shows the percentage of strategies that are constructed on Level k actions. The blue dot corresponding to 4 along the x-axis has a height of just over 80%, meaning that, when considering strategies that are played in a certain game by exactly 4 subjects, just over 80% of such strategies are Level k actions. Each dot with an x-value of 5 or greater has a y-value Level k frequency of 100% (Result 8). This suggests that *Beliefs* participants are not "missing" any types when they express their beliefs.

Result 8. *If 5 or more* Actions *treatment participants select a given action in a Number Selection Game, g, then that action is a Level k strategy.*

Another reason we expect U types do not believe in unknown deterministic types is that, on average, they distribute their boxes more in BA tasks compared to C subjects. For each

Beliefs subject i, we compute s_i, the sum the squared heights of each column built.[25] The s_i values of C participants are greater than those for U subjects (Result 9).

Result 9. *The mean sum of squared column heights for C and U participants are 875 and 444, respectively. This difference is highly significant using a two-sided t-test (*$p < 0.0001$*).*

Despite the aforementioned results, it may be the case that U participants anticipate deterministic behavior, yet, may not express it because they are less confident in their beliefs and more risk averse in comparison to C subjects. We thus consider the responses by U and C participants in the Relative Performance (RP) questions and Bomb Risk (BR) decision. For a *Beliefs* subject i, we define her confidence level as the average fraction of subjects in i's session believed to have performed weakly worse than i in the BA tasks. (This is computed using the responses to the RP questions.) Thus, if i believes that, in each BA task, she has weakly more overlapping boxes than all others in her session, her confidence level is 1. If she believes that all others have strictly more overlapping boxes than does she in each BA task, her average confidence level is 0. We do not find any evidence that C participants are more confident in their constructed histograms in comparison to U subjects (Result 10).

Result 10. *The average confidence levels of C and U participants are .700 and .667, respectively. This difference is insignificant using a two-sided t-test (*$p = 0.2112$*).*

In both treatments, subjects' points are converted to monetary payoffs using binary lotteries (Roth and Malouf [1979]). This decision was to incentivize participants to maximize their expected number of points irrespective of their risk preferences. Binary lotteries do not always work in practice (see Selten et al. [1999]), but they seem to have served their purpose in our study. In the BR decision, if we consider the average number of "treasure chests" opened by a *Beliefs* participant, we find that U subjects do not open significantly fewer chests than C participants. In other words, U subjects are not significantly more risk averse than C participants (Result 11).[26]

Result 11. *The average numbers of boxes opened by C and U participants in the BR decisions are 49 and 46, respectively. This difference is insignificant using a two-sided t-test (*$p = 0.4206$*).*

[25] Using this measure has its advantages over "simpler" alternative statistics. For instance, if we were to simply count the number of columns subjects construct or the average height of each constructed column, we would not be able to differentiate between a participant who always places 19 boxes on one action and 1 on another from a subject who always places 10 boxes on each of two actions.

[26] More generally, we see that the binary lotteries used in the BA tasks seem to have effectively neutralized risk attitudes among all *Beliefs* participants. An OLS regression of the sum the squared heights of each column built on the number of treasure chests opened in the BR yields a positive slope coefficient with p-value of 0.544.

Given the findings up to this point, the next question to address is whether U subjects are creating flatter histograms than C participants because they (i) believe *Actions* participants are noisy or (ii) are noisy themselves. If (ii) is a more accurate explanation, it seems reasonable to expect U participants to be less confident in their choices compared to C subjects (who we know are not noisy). We know, however, from Result 10, that U subjects are no less confident in their histograms than C participants. It thus seems that (i) is more plausible than (ii). In fact, simply looking at the histograms made by U subjects suggests more evidence supporting (i): it appears that U subjects arrange their boxes more towards the upper bounds of NS games in comparison to random allocations.

To investigate this more rigorously, we compute the following "front-load factor" f_i^h for subject i's histogram h. Specifically, if h corresponds to a game with upper bound UB, then $f_i^h = sum_{j=1}^{UB} h_j$ where h_j is the total number of boxes placed across actions $UB - j + 1$ through UB in h. Thus, if h has all 20 boxes on the upper bound, $f_i^h = 20 \times UB$. If h has 19 boxes on UB and 1 box on $UB - 1$, then $f_i^h = 19 + 20 \times (UB - 1)$. The smallest f_i^h can be is if h has all 20 boxes placed on its lower bound, leaving $f_i^h = 20$. For *Beliefs* participant i, we average her front-load factors over all 11 BA tasks to compute her "front-load statistic", $F_i = (f_i^{h_1} + \cdots + f_i^{h_{11}})/11$. We find that U participants front-load their histograms more than would random subjects (Result 12). Furthermore, C subjects front-load their histograms more than U participants (Result 13).

Result 12. *The mean front-load statistic for U participants is 254 which is significantly different than 221 ($p = 0.0126$), the expected front-load statistic reached via random box allocation.*

Result 13. *The average front-load statistics for C and U participants are 323 and 254, respectively. This difference is significant using a two-sided t-test ($p < 0.0001$).*

4 Related Literature

The Number Selection (NS) games are inspired by a variety of existing games. They most closely resemble the Generalized Centipede (GC) games from Fragiadakis et al. [2017]. In GC games, players select integers from (possibly different) guessing ranges. As with our NS games, guessing x guarantees a player earns at least x. In addition, there are bonuses that can be attained for undercutting as well as matching one's opponent. The GC games are less constrained than ours. We impose more restrictions on ours for parsimony, ease of explanation to experimental subjects and to facilitate our analysis.

The GC games were preceded by the 11-20 Money Request (MR) game studied in Arad and Rubinstein [2012]. In the two-person MR game, each player simultaneously selects an integer between 11 and 20 (inclusive). Guessing x earns a player x, unless x is exactly 1 less than the opponent's guess, in which case the player earns $x + 20$. The Level 0 prediction is 20, the upper bound. Then, Level 1 is 19, Level 2 is 18 etc. The MR game has inspired others as well. For instance, Georganas et al. [2015] and Goeree et al. [2013] and Alaoui and Penta [2016] study games that are very similar to the 11-20, game, except for some changes in the payoff structures.

While we design our NS games, a more prominent feature of our design is our method for belief elicitation. A variety of *scoring rules* exist for eliciting beliefs; see Selten [1998] for an overview. The Quadratic Scoring Rule (QSR), for instance, is quite common.[27] As an example of QSR, consider a game with three pure strategies, a, b and c. When an individual reports A, B and C for the likelihoods that her opponent will play a, b and c, respectively, her payoffs are computed as $P - (A - \mathbb{1}\{s = a\})^2 - (B - \mathbb{1}\{s = b\})^2 - (C - \mathbb{1}\{s = c\})^2$, where $P > 0$ is a prize and $\mathbb{1}\{s = x\}$ is an indicator function that equals 1 if and only if $s = x$ and equals 0 otherwise. It can be shown that a risk neutral agent is incentivized to truthfully express her beliefs to this mechanism.

Though QSR has been widely implemented,[28] data suggests it may have some shortcomings in practice. For example, Palfrey and Wang [2009] find that QSR elicits more extreme beliefs than an alternative (logarithmic) scoring rule that is also incentive compatible. Interestingly, Huck and Weizsäcker [2002] find beliefs to be biased towards 50-50. Armantier and Treich [2013] highlight some additional drawbacks of incentive compatible scoring rules, namely, that stakes, incentives and hedging opportunities can substantially distort reported probabilities.

We believe there are two features of QSR that may pose some difficulty for real-world participants. First, it does not seem to convey its incentive properties transparently. We believe the design of our Box Arrangement tasks clearly do by making the incentives more visual via overlapping histograms. After the design of our experiment, we discovered that Carpenter et al. [2013] also use histograms to elicit beliefs, perhaps because they held similar concerns regarding a subject's understanding of the belief-elicitation mechanism.

Second, QSR asks a subject to think about how likely it is for a single individual to take various actions, akin to asking an individual for the probability that it will rain today. We

[27] The quadratic scoring rule was first studied by Brier [1950] and Good [1952].

[28] Costa-Gomes and Weizsäcker [2008] use it to investigate whether a subject's actions in a normal form game is a best response to her stated beliefs of how others will play that game. Dufwenberg and Gneezy [2000] elicit beliefs in a Lost Wallet game to investigate how beliefs affect trust dilemmas. Dominitz and Hung [2009] explore how beliefs are updated as new pieces of information are publicly announced. See Palfrey and Wang [2009] for a discussion of additional papers that have used QSR.

believe an easier question would be: "how many days this week do you expect it will rain?" This is why we ask subjects to predict how many others choose different strategies as oppose to matching an individual with another and asking her for the likelihood that her opponent will choose different actions. Huck and Weizsäcker [2002] also ask subjects to forecast quantities of other participants. In addition to the potential difficulty in formulating a probabilistic belief over a single occurrence (i.e., whether it will rain today), we think that asking an individual to predict the behavior of a single participant may trigger one to think extremely. In other words, if asked for the beliefs of whether an opponent will select pure strategy x or y, one may simply report the action that she believes her opponent is more likely to choose, which can explain the prior instances where QSR has recorded extreme responses.

5 Conclusion

This paper contributes to the behavioral game theory literature by testing whether the Cognitive Hierarchy (CH) model developed by Camerer et al. [2004] is more than simply an "as if" theory of behavior. CH describes an individual as believing her opponents engage in heterogeneous steps of strategic thinking. To test for such beliefs, we first have subjects in one treatment play a series of games. Their behavior is then predicted by separate participants (in a different treatment). As a participant builds a histogram to express her beliefs, she first anticipates a number of non-strategic individuals who play a naive strategy. Then, she believes that there are players who play the best-response to this strategy, followed by players who best respond to these strategic individuals and so on. Furthermore, we find that the *shape* of such beliefs resonates with Camerer et al. [2004] who use Poisson CH specifications to estimate aggregate data: we find that normalized Poisson distributions do better than Uniform supports in modeling how an individual expects various thinking steps to occur.

Our data show that CH cleanly passes the test of its beliefs assumptions, which may shed some light on a few existing puzzles in behavioral game theory. First, even in games whose most commonly selected actions are Level k strategies, a substantial portion of behavior is often unexplained by Level k; such behavior may, however, be rationalized by a Step k thinker best responding to some distribution of lower step thinkers. Second, the classification of an individual as a Level k player in a certain type of game has limited predictability as far as their behavior in other types of games.[29] A Step 2 thinker who believes $\alpha \in (0,1)$ and $1-\alpha$ of the population are Step 0 and Step 1 thinkers, respectively, may have a best response that

[29] See Georganas et al. [2015]. Similar results are found in Fragiadakis et al. [2017].

coincides with the Level 1 strategy in certain games but with the Level 2 strategy in others.[30]

Lastly, our relatively "clean" results from our *Beliefs* treatment provide a proof-of-concept that our belief-elicitation method is not only theoretically appealing, but also practically successful. We hope future researchers interested in recording beliefs seriously consider our method; we expect this would help them to similarly obtain minimally noisy responses.

[30] Furthermore, she may adjust her beliefs of α depending on her perception of games' degrees of complexity.

References

Larbi Alaoui and Antonio Penta. Endogenous depth of reasoning. *The Review of Economic Studies*, 83(4):1297–1333, 2016. 22

Ayala Arad and Ariel Rubinstein. The 11–20 money request game: a level-k reasoning study. *The American Economic Review*, 102(7):3561–3573, 2012. 2, 4, 22

Olivier Armantier and Nicolas Treich. Eliciting beliefs: Proper scoring rules, incentives, stakes and hedging. *European Economic Review*, 62:17–40, 2013. 22

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950. 22

Alexander L Brown, Colin F Camerer, and Dan Lovallo. Estimating structural models of equilibrium and cognitive hierarchy thinking in the field: The case of withheld movie critic reviews. *Management Science*, 59(3):733–747, 2013. 1

Colin Camerer. *Behavioral game theory: Experiments in strategic interaction.* Princeton University Press, 2003. 1

Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898, 2004. 1, 3, 5, 14, 15, 16, 23

Jeffrey Carpenter, Michael Graham, and Jesse Wolf. Cognitive ability and strategic sophistication. *Games and Economic Behavior*, 80:115–130, 2013. 22

Miguel Costa-Gomes, Vincent P Crawford, and Bruno Broseta. Cognition and behavior in normal-form games: An experimental study. *Econometrica*, 69(5):1193–1235, 2001. 2

Miguel A Costa-Gomes and Georg Weizsäcker. Stated beliefs and play in normal-form games. *The Review of Economic Studies*, 75(3):729–762, 2008. 22

Vincent Crawford and Miguel Costa-Gomes. Cognition and behavior in two-person guessing games: An experimental study. *American Economic Review*, 96(5):1737–1768, 2006. 4

Paolo Crosetto and Antonio Filippin. The "bomb" risk elicitation task. *Journal of Risk and Uncertainty*, 47(1):31–65, 2013. 10

Jeff Dominitz and Angela A Hung. Empirical models of discrete choice and belief updating in observational learning experiments. *Journal of Economic Behavior & Organization*, 69(2): 94–109, 2009. 22

Martin Dufwenberg and Uri Gneezy. Measuring beliefs in an experimental lost wallet game. *Games and economic Behavior*, 30(2):163–182, 2000. 22

Urs Fischbacher. z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, 10(2):171–178, 2007. 7

Daniel E Fragiadakis, Daniel T Knoepfle, and Muriel Niederle. Who is strategic? 2016. 4

Daniel E Fragiadakis, Daniel T Knoepfle, and Muriel Niederle. Do individuals employ the same decision rules across strategic environments? 2017. 2, 21, 23

Sotiris Georganas, Paul J Healy, and Roberto A Weber. On the persistence of strategic sophistication. *Journal of Economic Theory*, 159:369–400, 2015. 2, 22, 23

Jacob K Goeree, Philippos Louis, and Jingjing Zhang. Noisy introspection in the "11–20" game. Technical report, Working paper, 2013. 22

Irving John Good. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 107–114, 1952. 22

Ben Greiner et al. The online recruitment system orsee 2.0-a guide for the organization of experiments in economics. *University of Cologne, Working paper series in economics*, 10 (23):63–104, 2004. 7

Shaun Hargreaves Heap, David Rojo Arjona, and Robert Sugden. How portable is level-0 behavior? a test of level-k theory in games with non-neutral frames. *Econometrica*, 82(3): 1133–1151, 2014. 5

Ali Hortaçu, Fernando Luco, Steven Puller, and Dongni Zhu. Does strategic ability affect efficiency? evidence from electricity markets. 2017. 1

Steffen Huck and Georg Weizsäcker. Do players correctly estimate what others do?: Evidence of conservatism in beliefs. *Journal of Economic Behavior & Organization*, 47(1):71–85, 2002. 22, 23

Rosemarie Nagel. Unraveling in guessing games: An experimental study. *The American Economic Review*, 85(5):1313–1326, 1995. 3

Thomas R Palfrey and Stephanie W Wang. On eliciting beliefs in strategic games. *Journal of Economic Behavior & Organization*, 71(2):98–109, 2009. 22

Alvin E Roth and Michael W Malouf. Game-theoretic models and the role of information in bargaining. *Psychological review*, 86(6):574, 1979. 2, 9, 20

Reinhard Selten. Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1(1):43–62, 1998. 22

Reinhard Selten, Abdolkarim Sadrieh, and Klaus Abbink. Money does not induce risk neutral behavior, but binary lotteries do even worse. *Theory and Decision*, 46(3):213–252, 1999. 20

Dale O Stahl and Paul W Wilson. Experimental evidence on players' models of other players. *Journal of economic behavior & organization*, 25(3):309–327, 1994. 3

Dale O Stahl and Paul W Wilson. On players models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1):218–254, 1995. 1, 5, 14

Joseph Tao-yi Wang, Michael Spezio, and Colin F Camerer. Pinocchio's pupil: using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. *The American Economic Review*, 100(3):984–1007, 2010. 2