Noisy Introspection in the "11–20" Game

Jacob K. Goeree, Philippos Louis, and Jingjing Zhang*

November 13, 2013

Abstract

Arad and Rubinstein (*American Economic Review*, 102(7), 2012, 3561–3573) recently proposed a simple money-request game designed to trigger level-k reasoning. In an experiment that explores three variants of the game, they find evidence for the level-k model with observed levels of strategic thinking consistently ranging from 0 to 3. Our baseline treatment uses the basic version of the money-request game and replicates their results. We apply the noisy introspection model developed by Goeree and Holt (*Games and Economic Behavior*, 46, 2004, 365–382) to the baseline-treatment data and use this to predict *behavior and beliefs* in five other treatments that employ games with a very similar structure. The data from these additional treatments clearly refute the level-k model, which predicts no better than the Nash equilibrium in these games. Our data provide striking evidence that the assumption of best-response behavior underlying the level-k model is untenable. The noisy introspection model, which instead assumes "common knowledge of noise," predicts behavior remarkably well.

Keywords: *Noisy introspection, level-k, QRE, Nash, money-request game*

*ESEI Center for Market Design, University of Zürich, Blümlisalpstrasse 10, CH-8006 Zürich, Switzerland, and Economics Discipline Group, University of Technology Sydney, PO Box 123, Broadway NSW 2007 Sydney, Australia. We benefitted from helpful comments made by seminar participants at University of Vienna, University of Technology, Sydney, University of Sydney, the 8th Alhambra Experimental Workshop in Rome, the 8th Annual Australia New Zealand Workshop in Brisbane, the 2013 Annual Meeting of the Association of Southern European Economic Theorists in BilBao and the 2013 Economic Science Association World Meetings in Zurich. We would like to thank the Swiss National Science Foundation (SNSF 135135) and the European Research Council (ERC Advanced Investigator Grant, ESEI-249433) for financial support.

1. Introduction

Behavior in one-shot games often differs substantially from Nash equilibrium predictions (e.g. Goeree and Holt, 2001[?]), which has led to the development of alternative models. These alternatives relax either the assumption of correct beliefs or the assumption of perfectly maximizing behavior. The leading candidate in the latter category is McKelvey and Palfrey's (1994)[?] quantal response equilibrium (QRE), which subsumes that decision making is noisy but that beliefs are correct on average. An important strength of QRE is that it is "context-free," i.e. it can be applied uniformly to data sets from different experiments without having to be adapted to the specifics of the experimental context. In repeated-game experiments where behavior has a chance to converge, QRE typically does a good job at predicting final-period averages as well as comparative statics across treatments. For one-shot games, however, the assumption that beliefs are correct on average is generally not realistic. Moreover, the basic QRE model corresponds to a symmetric Bayes-Nash equilibrium that predicts homogenous behavior.

Observed behavior, in contrast, typically appears quite heterogenous. This has stirred interest in theories that allow for different levels of strategic sophistication, or different levels of thinking. In this category, the leading candidate is the level-k model ([?, ?, ?]) that employs a potentially infinite hierarchy of strategic thinking: level-0 chooses naively or randomly, level-1 best responds to level-0, level-2 best responds to level-1, etc. Given that the behavior of higher levels is fixed by that of the lowest level, the specification of level-0 behavior is crucially important. Initially, level-0 behavior was simply modeled to be uniform, resulting in a context free model that can be generally applied. Recently, more elaborate specifications of level-0 behavior that take into account details of the environment have been proposed in order to improve fit. Without generally applicable rules for how to map certain game (or other) variables into level-0 behavior, however, this approach has the flavor of "doing theory with a dummy variable."

Unless, of course, the environment dictates an obvious and unique choice for the non-strategic level-0. Arad and Rubinstein (2012)[?] propose such an environment: the "11-20" game where two players can ask for any integer amount between (and including) 11 and 20 and receive what they ask for. This is the non-strategic part of the game and since even a level-0 understands that "more is better," the obvious choice for level-0 is to ask for 20. The strategic part of the game specifies that an additional bonus of 20 is rewarded to a player whose ask amount is 1 less than that of the other player. A level-1 player would therefore ask for 19, level-2 for 18, etc. In three variations of the "11-20" game, Arad and Rubinstein (2012) find that the inferred levels of thinking consistently range from 0 to 3. Arad and Rubinstein (2012) thus accomplish two important goals: (i) they design a game for which level-k type thinking is natural and for which the level-0 choice is obvious, (ii) they report data that support the level-k model and corroborate results from previous experiments.

That is not to say that their data are inconsistent with alternative models such as QRE. Given observed choice frequencies, requesting an amount of 17, 18, or 19 (attributed to levels 3, 2, and 1 respectively) yields an expected payoff above 20 and QRE thus also predicts these numbers are likely to be chosen.[1] To better separate the different models we consider variations of the "11-20" game that leave intact the obvious level-0 choice and the best-response structure of the game but that change the payoffs associated with different levels of thinking. We do this by assigning the numbers 11 to 20 to ten boxes arranged on a line, always reserving the rightmost box for 20. Subjects receive the number in the box they choose plus a reward if their chosen box is immediately to the left of that chosen by the other subject. The standard "11-20" game corresponds to arranging numbers in increasing order (from left to right) but in other variations the sequence is not monotone. For example, in an "extreme" variation, numbers decline from 19 to 11 ending, as usual, with 20. This reshuffling of numbers does not affect the logic underlying the level-k model: level-0 chooses the rightmost box with 20, level-1 the box next to it, level-2 the box next to that, etc. In other words, the level-k model predicts behavior in these variations to be *identical* to that in the standard game.

Observed behavior in these variations differs markedly from level-k predictions, however. Subjects submit a high request, say 19, irrespective of whether this corresponds to a level-1 choice in the standard game or to a level-9 choice in the extreme variation. While not predicted by the level-k model, a choice of 19 is actually quite intuitive in that it costs only 1 and potentially rewards 20. When others' behavior is noisy and dispersed, all request amounts have some chance of yielding the bonus and those for which the cost is low will naturally be explored. Importantly, this argument requires "common knowledge of noise," i.e. not only *is* behavior noisy but subjects *expect* it to be noisy and act accordingly. Importantly, common knowledge of noise results in drastically different predictions than simply adding noise to the level-k model: the latter would simply predict a dispersion of the levels observed in the baseline game and cannot explain why a substantial fraction of the subjects acts as if they are of level 9 in the extreme variation of the game.

The noisy introspection model introduced by Goeree and Holt (2004)[**?**] naturally captures the notion of common knowledge of noise. Like the level-k model it assumes heterogeneity

[1] In Arad and Rubinstein's (2012) experiment the choice frequencies for amounts of 20, 19, 18, and 17 are 6%, 12%, 30% and 32% resulting in expected payoffs of 20, 20.2, 20.4, and 23 respectively.

in levels of thinking but it replaces strict best responses with noisy best responses. In other words, level-1 makes a noisy best response to level-0, level-2 makes a noisy best response to level-1, etc. We put the noisy introspection model to the test as follows. We first replicate Arad and Rubinstein's (2012) baseline treatment and use this to identify the percentages of noisy level-k thinkers, for $k = 0, 1, 2, \ldots$, as well as a common noise parameter. These are then used to out-of-sample predict behavior *and* beliefs in the five variations of the "11-20" game we study. As detailed below, the noisy introspection model predicts choices and beliefs strikingly well across all game variations unlike the level-k model, which predicts no better than Nash.

This paper is organized as follows. The next section explains the noisy introspection model. Section 3 details the experimental design and Section 4 discusses the experimental results. Section 5 concludes and the Appendix contains the experimental instructions.

2. Noisy Introspection

In the noisy introspection model, players apply a process of iterated reasoning about what the other will choose, what the other thinks the player will choose, what the other thinks the player thinks the other will choose, etc. It is natural to assume that this thought process becomes increasingly complex with every additional iteration, which can be neatly captured by considering a sequence of noisy responses with non-decreasing noise parameters. To formalize, consider a two-player, symmetric game with a finite set of actions, A.[2] The expected payoff $\pi^e(a, q)$ of choosing $a \in A$ depends on a player's beliefs, q, which is a probability distribution over A. Adopting the familiar logit formulation we can define the "better response" mapping $\phi_\mu : [0,1]^{|A|} \to [0,1]^{|A|}$ with components

$$\phi^a_\mu(q) \;=\; \frac{\exp(\pi^e(a,q)/\mu)}{\sum_{a' \in A} \exp(\pi^e(a',q)/\mu)} \qquad \forall a \in A \tag{1}$$

The noise parameter, μ, determines how sensitive the response function is with respect to expected payoffs: $\mu = 0$ results in a best response and $\mu = \infty$ in uniform randomization.

The unique noisy introspection prediction, ϕ, can be defined as the limit sequence

$$\phi \;=\; \lim_{n \to \infty} \phi_{\mu_0} \circ \phi_{\mu_1} \circ \ldots \circ \phi_{\mu_n}(q) \tag{2}$$

where $\mu_0 \leq \mu_1 \leq \ldots \leq \mu_\infty = \infty$, which guarantees that ϕ is independent of the belief q used as a starting point for the iterated thought process. Besides the monotonicity and limit con-

[2] Symmetry allows us to avoid player specific subscripts.

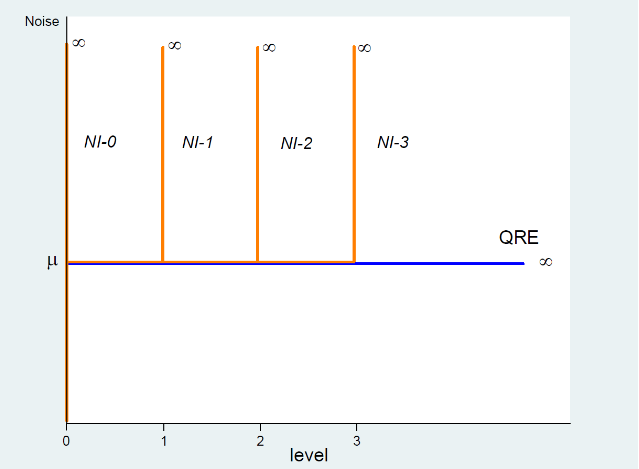


Figure 1: Various levels of noisy thinking in the NI-k model. Each line corresponds to a different sequence of error parameters $\mu_0 \leq \mu_1 \leq \ldots \leq \mu_\infty = \infty$. For example, the line labeled NI-0 corresponds to completely random decision making, which occurs if $\mu_0 = \infty$. The next level NI-1 reflects a noisy best response to uniform beliefs, which occurs if $\mu_0 = \mu$ and $\mu_1 = \infty$. Similarly, NI-k for higher k simply corresponds to the case $\mu_0 = \mu_1 = \ldots = \mu_{k-1} = \mu$ and $\mu_k = \infty$.

ditions, the noisy introspection model imposes no further restrictions on the sequence of noise parameters thereby allowing for various special cases to be included. Goeree and Holt (2004), for instance, consider a homogeneous noisy introspection model where all players are characterized by the same geometrically increasing sequence of noise parameters. A parsimonious model that exhibits heterogeneity follows by considering different levels of noisy thinking, NI-k for $k = 0, 1, 2, \ldots$, where the sequence of noise parameters for NI-k is given by

$$\mu_{\hat{k}} = \begin{cases} \mu & \hat{k} < k \\ \infty & \hat{k} \geq k \end{cases} \tag{3}$$

So level-0 randomizes, level-1 makes a noisy best response to uniform beliefs, level-2 makes a noisy best response to a noisy best response to uniform beliefs, etc. Figure 1 illustrates the noise sequences of the various levels.

An appealing feature of the NI model presented in (3) is that it includes other popular models as special cases. For instance, when $\mu = 0$ the noisy introspection model reduces to the level-k model that employs strict best responses.[3] As another example, suppose all players have

[3] One potential difference is that level-0 corresponds to random behavior in the noisy introspection model but not necessarily in the level-k model. Recent versions have allowed the definition of level-0 to depend on the specifics of the game. For example, for the "11-20" game, Arad and Rubinstein (2012) argue that level-0 play is more adequately described by a choice of 20. When we apply level-k to the data we follow their suggestion (although our results would not change by specifying level-0 behavior to be random). For the noisy introspection model we always assume that level-0 chooses randomly.

infinite levels of noisy thinking so that the sequence of noise parameters is constant at μ. Then (2) converges to a quantal response equilibrium, if it converges at all, as the limit sequence satisfies $\phi_\mu(\phi) = \phi$. This limit is illustrated by the horizontal line in Figure 1. Finally, in those cases where (2) converges with constant noise parameters even when $\mu = 0$, the outcome converges to a Nash equilibrium. So all the familiar models, level-k, QRE, and Nash, are potentially nested.

3. Experimental Design

The experiment used variations of Arad and Rubinstein's (2012) money request game, which were described as follows:[4]

> *You and another participant in the experiment are randomly matched to play the following game. On your screen you see 10 boxes in line, containing different amounts. Each player requests an amount of points by selecting one of the 10 boxes. Each participant will receive the amount in the box he/she selected. A participant will receive an additional amount of R points if the selected amount is exactly 'one to the left' of the amount that the other participant chooses. Which box do you select?*

Subjects were in one of two treatments. In the "11–20" treatment the amounts in the boxes ranged from 11 to 20 experimental points and the bonus was $R = 20$ points. In the "1–10" treatment the amounts ranged from 1 to 10 points and the bonus was $R = 8$ points. The exchange rate from experimental points to Swiss Francs was adjusted accordingly so that a choice of the highest number in the rightmost box would equal 5 Swiss Francs in either treatment.

Within a treatment there were three parts. Subjects were given separate instructions at the start of each part and received no feedback about their payoffs until the end of the experiment. In part 1, subjects played three versions of the game against a random opponent. Each game has a different arrangement of the amounts in the boxes, see Figure 2, with the highest amount always located on the far right. In the baseline version the numbers are arranged in increasing order from left to right. In the extreme (E) version the numbers are arranged in decreasing order except that the rightmost box again contains the highest number. Finally, in the moderate (M) version, the second to highest amount is put in the middle. To control for order effects, subjects were randomly assigned (in equal proportions) to one of 6 possible orderings of the three game

[4] The complete set of instructions can be found in the Appendix.

Treatment "11–20"

B	11	12	13	14	15	16	17	18	19	20
E	19	18	17	16	15	14	13	12	11	20
M	14	13	12	11	19	18	17	16	15	20

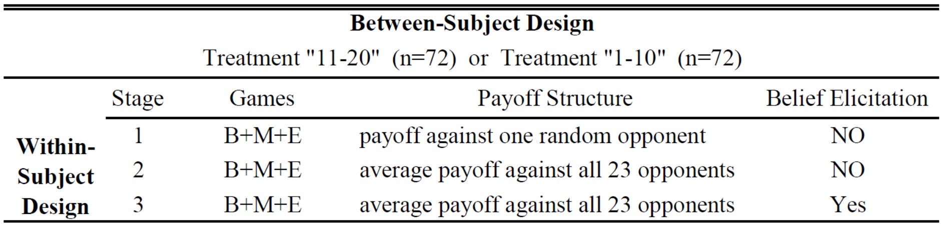
Treatment "1–10"

B	1	2	3	4	5	6	7	8	9	10
E	9	8	7	6	5	4	3	2	1	10
M	4	3	2	1	9	8	7	6	5	10

Figure 2: In one treatment, subjects played the three versions of the "11-20" game shown in the top panel. The baseline (B) version corresponds to Arad and Rubinstein's (2012) basic version while the moderate (M) and extreme (E) games reorder the positions of the 10 numbers and place 19 in the middle and in leftmost node respectively. The other treatment consists of three parallel versions of the "1-10" game where the request amounts range from 1 to 10 and the bonus is $R = 8$.

variations. In part 2, subjects played the games in the same order as they had in part 1, but now a subject's payoff was equal to the average payoff resulting from all possible matches (each session had 24 subjects so there were 23 possible matches). Part 3 also used population payoffs but now play was preceded by a belief-elicitation stage: subjects were asked to guess how many of the other 23 participants would choose each of the amounts. Subjects were rewarded for their guesses using a quadratic scoring rule. Figure 3 provides a summary of the experimental design, which has both between-subjects ("11-20" or "1-10" game) and within-subjects elements (three variations of the game played with standard payoffs, population payoffs, and population payoffs plus belief elicitation).

To determine subjects' earnings from the experiment, one game was randomly chosen from each part and subjects received their payoff in that game, plus the payoff from the belief



Between-Subject Design

Treatment "11-20" (n=72) or Treatment "1-10" (n=72)

	Stage	Games	Payoff Structure	Belief Elicitation
Within-Subject Design	1	B+M+E	payoff against one random opponent	NO
	2	B+M+E	average payoff against all 23 opponents	NO
	3	B+M+E	average payoff against all 23 opponents	Yes

Figure 3: Experimental Design

elicitation process corresponding to the game picked from part 3, and a show-up fee of 10 Swiss Francs. This resulted in average earnings of 28.91 Swiss Francs.

A total of 144 subjects participated in 6 experimental sessions, 24 in each. We conducted three sessions for both treatments. Subjects were recruited among undergraduate students at ETH Zurich and the University of Zurich using ORSEE (Greiner, 2003)[**?**]. The experiment was conducted in the Experimental Economics Lab of the University of Zurich using Z-Tree (Fischbacher, 2007[**?**]).

4. Experimental Results

Arad and Rubinstein (2012) describe several features of the "11-20" money request game that make it ideal for studying level-k type reasoning. The two most relevant features, which remain true in the variations of the "11-20" game we study, pertain to the level-0 choice and the best-response structure of the game. First, the level-0 choice is intuitively obvious since selecting the number in the rightmost box gives the highest certain payoff. Second, best responding to any level-k strategy is straightforward because given level-0 choosing the rightmost box, level-1 best responds by choosing the box to the left of it, level-2 best responds by choosing the box to the left of that, and so on. In other words, only the positions of the boxes, not the numbers they contain, matter for the level-k best-response process. As a result, the level-k model predicts identical behavior across the different treatments.

4.1. Replicating Arad and Rubinstein (2012)

The top-left panel of Figure 4 shows the distribution of choices made by 72 subjects in our baseline "11-20" treatment, which replicates Arad and Rubinstein's (2012) main findings: 77% of the choices correspond to levels 1-3 (choosing amounts between 19 and 17), which is not different from the 74% reported by Arad and Rubinstein (2012), and few choices (12.5%)

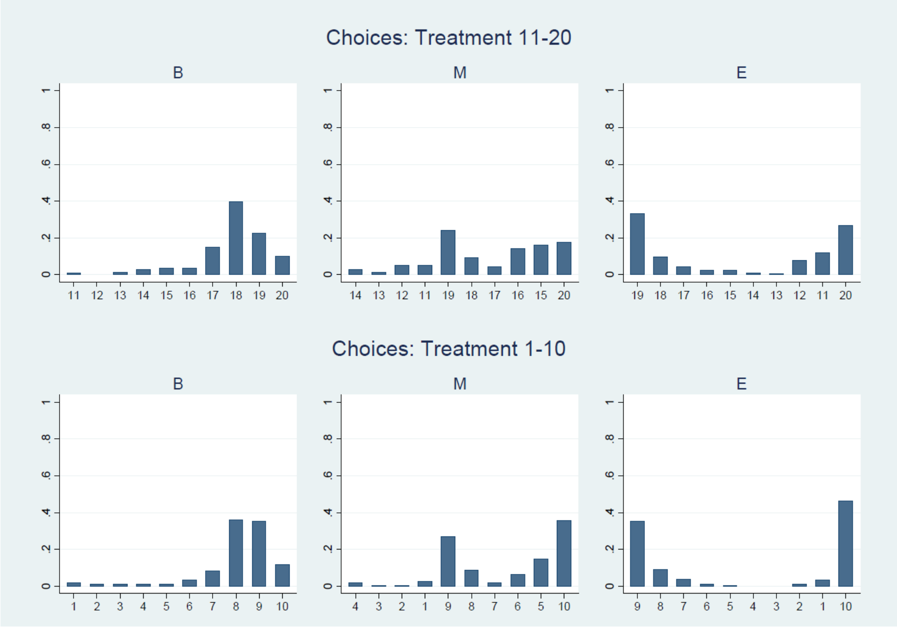


Figure 4: Observed distribution of choices by game and treatment.

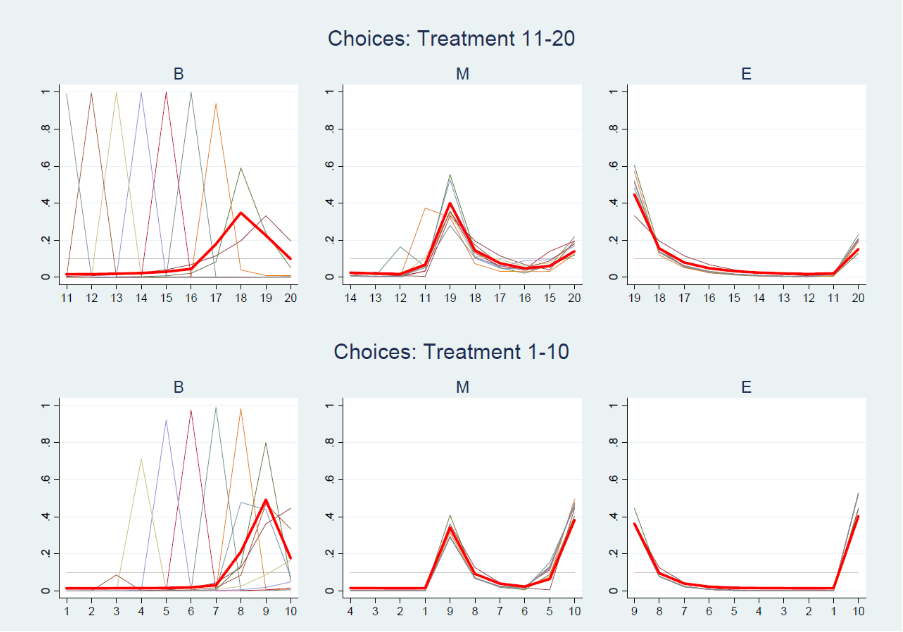


Figure 5: Predicted distribution of choices under Noisy Introspection Model.

reflect a level higher than three.

Finding 1. *Our baseline "11-20" treatment replicates Arad and Rubinstein's (2012) main finding that the bulk of the choices reflect up to three levels of reasoning.*

Following Arad and Rubinstein (2012) we might conclude that the baseline data provides strong support for the level-k model. However, as noted in the Introduction, other models (e.g. QRE) can also rationalize observed behavior in the baseline treatment. Furthermore, the level-k model predicts identical behavior in the other variations of the "11-20" game, which is clearly not borne out by the experimental data, see the top-middle and top-right panels of Figure 4. The fraction of choices that can be attributed to levels 1-3 is only 35% and 20% in the moderate and extreme variations respectively. Roughly half of the choices correspond to levels higher than three in both variations. The most frequent choice in the moderate game is the middle box containing number 19 (24%) and in the extreme game it is the leftmost box containing 19 (33%). Under the level-k model these choices would correspond to previously undocumented high levels of thinking (levels 5 and 9 respectively). More realistically, these choices reflect that subjects react to expected payoffs, which are maximized at 19 in both the moderate and extreme variations.[5]

4.2. Out-of-sample predictions of the Noisy Introspection model

We next use the data from the baseline "11-20" game to classify subjects according to the noisy introspection model. Recall that each subject plays the baseline game three times, once in each part. Assuming a common error parameter, μ_{NI}, we determine a subject's level as the one that maximizes the likelihood of the three observed choices in the baseline "11-20" game. The error parameter in turn is chosen to maximize the overall likelihood of all 216 subjects' choices. This estimation procedure yields an estimate for the error parameter of $\mu_{NI} = 1.89(0.07)$ and the following distribution of levels: out of 72 subjects, 10 are NI-0, 22 are NI-1, 33 are NI-2, and 7 are NI-3.

Remark 1. *Relatively few (26.4%) subjects choose three times the same amount when they play the baseline game in the different parts of the experiment. This may be problematic for the level-k model but not for the noisy introspection model where some degree of trembling is expected.*

Remark 2. *The noisy introspection model identifies levels up to three even though occasion-*

[5] Subjects' ex post explanations confirm that a choice of 19 is rarely the result of level-5 or level-9 type reasoning.

ally amounts lower than 17 are chosen. The reason is that subjects may choose low amounts occasionally but never do so three times in a row. In some sense, the noisy introspection model "validates" Arad and Rubinstein's (2012) claim that choices less than 17 reflect noise not higher levels of thinking.

We can use the estimated error parameter to out-of-sample predict the choice distributions of the various levels in all game variations. These predictions are shown by the thin lines in the bottom part of Figure 4. For example, in the top-left panel of the bottom part, the thin horizontal line at height 0.1 shows the choice distribution of NI-0. The line with a peak at 19 corresponds to NI-1, the line with a peak at 18 to NI-2, etc. The thick line is the aggregate choice distribution calculated by weighting the thin lines with the empirical frequencies of the different levels: 10/72 for NI-0, 22/72 for NI-0, 33/72 for NI-0, and 7/72 for NI-0, while the weights are 0 for NI-k with $k > 3$. Notice that the aggregate choice distribution predicted by the noisy introspection model nicely reproduces the actual choice distribution shown in the top-left panel of the top part of Figure 4.

More importantly, this procedure gives predictions for the five game variations as well. Note that in the moderate and extreme variations the various levels (different from 0) are predicted to behave quite similarly. As a result, the noisy introspection model yields very tight predictions for the aggregate choice distribution.

The top panel of Figure 4 shows the distribution of choices made by 72 subjects in the three variations of the "11-20" game and the bottom panel shows choices for another 72 subjects in three parallel variations of the "1-10" game.[6] For each game, we pool the choices from the three different parts of the experiment.[7] The baseline game (B) of the "11-20" treatment replicates the A&R's results: 77% of the choices correspond to levels 1-3 (choosing amounts between 19 and 17), which is not different from the 74% reported by A&R. Only 12.5% of the choices reflect a level higher than three.

[6] Comparing treatments "11-20" and "1-10," the distributions are significantly different according to a chi-square test ($p < 0.05$ for each game), which is mainly driven by the higher percentage of level-0 and level-1 choices in "1-10." The percentage of level-0 choices increases from 18% to 36% in game M and from 27% to 46% in game E. In game B, the biggest difference is in the level-1 choices (23% to 35%) whereas level-0 choices are almost the same (11% in "1-10" and 10% in "11-20").

[7] Recall that each experimental session consists of three parts that differ in the payment rule and whether or not beliefs were elicited. In each part, participants made decisions in games B, M, and E. There are six possible ways to order the three games and we randomly assigned 4 participants to each of the six orderings (for a total of 24 subjects per session). Two-sided chi-square tests regarding the equality of choice distributions indicate no significant order effects within each part and no significant differences across the three parts (for all three games and in both treatments). In the analyses reported below we therefore pool data from all three parts, unless otherwise stated.

Finding 1. *The baseline of our "11-20" treatment replicates Arad and Rubinstein's (2012) main result: the vast majority of choices reflect up to three levels of reasoning only and higher levels are rare.*

Behavior is substantially different in games M and E where the fraction of choices that can be attributed to levels 1-3 is only 35% and 20% respectively. Roughly half of the choices correspond to levels higher than three in both games. The most frequent choice in game M is the middle box containing number 19 (24%) and in game E it is the leftmost box containing 19 (33%). Under the level-k model these choices would correspond to previously undocumented high levels of thinking (levels 5 and 9 respectively). More realistically, these choices reflect that subjects react to expected payoffs, which are maximized at 19 in both the M and E variations.[8]

A similar effect of the second-highest number is observed in the "1-10" treatment. In games M and E, number 9 attracts 27% and 35% of the choices respectively. Behavior in the game E is even more of a challenge for the level-k model since the best response to a choice of 10 is 10 since the bonus of 8 is not sufficient to compensate the loss of certain payoff. As a result, all levels are predicted to choose the rightmost box with 10.

Finding 2. *The position of the second-highest number has a significant effect on behavior. In treatment "11-20," the position of 19 shifts a considerable fraction of the choices from levels 1-3 in game B to level-5 in game M and to level-9 in game E. In treatment "1-10," the position of 9 shifts choices to level-*5 *in game M and to choices that do not correspond to any level in game E. These results clearly refute the level-k model that predicts identical choices across all games.*

The relevance of the second-highest amount in explaining the mode of the choice distribution suggests that subjects trade off the costs and benefits of obtaining the bonus. Compared to getting 20 for sure, the cost of choosing 19 is only 1 while it may yield a bonus of 20 when others' behavior is sufficiently dispersed. This tradeoff is naturally captured by the noisy introspection model of Section 2. Using data from the baseline variation of the "11-20" game only we estimate the noisy introspection model,

This section presents aggregate outcomes in our experiment and compares the performance of different models (Noisy Introspection, QRE, Level-k and Nash) in explaining the observed data pattern. In order to measure the predictive power of the NI model, we first apply NI to the baseline game (B) in the "11-20" treatment and then use the estimated parameters to

[8] Subjects' ex post explanations confirm that a choice of 19 is rarely the result of level-5 or level-9 type reasoning.

predict behavior and beliefs in the five other games (M, E games in the treatment "11-20" and B, M, E games in the treatment "1-10").

The model is estimated using the maximum likelihood techniques. To construct a log-likelihood function, let P_a denote the observed probability of requesting number a and $p_a(\mu)$ denote the corresponding predicted probability from the noisy introspection model. These introspection probabilities are calculated by taking the limit of the composition of the logit best response functions $\phi_{\mu_0} \circ \phi_{\mu_1} \circ ... \circ \phi_{\mu_n}(p_0)$ as the number of iterations goes to infinity. We approximate this by using only 10 iterations. We further impose the restriction that $\mu_0 = \mu_1 = \mu_2 ... = \mu_k = \mu_{NI}$ and $\mu_\infty = \infty$. The log-likelihood function is then

$$\log L(\mu_{NI}) = \sum_{a \in A} NP_a log(p_a(\mu_{NI})) \tag{4}$$

where $N = 3$ is the number of decisions made by each subject. Based on 216 choices made by 72 subjects who played game B in three parts of the "11-20" treatment, we obtain the estimate of $\mu_{NI} = 1.9$ (0.065) and the distribution of $NI - k$ levels is: $NI - 0$: 10, $NI - 1$: 22, $NI - 2$: 33, $NI - 3$: 7.

Replacing the introspection probabilities with the logit equilibrium probabilities in the log-likelihood function given by (3), we estimate the QRE model on the same data set. This yields the μ_{QRE} estimate of 3.1 (0.25). [9] Given the small standard errors in the NI estimate, we can reject that NI is a special case of Nash ($\mu = 0$) and QRE.

To measure the performance of different models in capturing the observed behavior, we compute the mean of the squared distances (MSD) between the predictions and data averages for all six games. Figure 8 shows MSD between predicted and actual percentages based on choice data (left panel), belief data (middle panel) under the four models we consider. We postpone the discussion about the right panel to the next section. Clearly, NI works better than Level-k and Nash in tracking the observed data. QRE seems to perform as well as NI on the aggregate level. We will see later that QRE performs much worse than NI on the indivdual level.

Finding 3. *Noisy introspection model captures the different choice patterns across games better than Level-k model.*

[9] For a given μ_{QRE} we can calculate the logit equilibrium probabilities by equating the probabilities that enter into the expected payoff function on the right side of equation 1 to the probabilities that come out on the left.

4.3. Beliefs

In part 3 of the experiment we elicited subjects' beliefs before play. The aggregate distribution (average across subjects) for each game is shown in Figure 6. An important point that must be made is that this aggregate distribution is not a collection of point predictions made by the different subjects. Each of the 144 subjects made 3 guesses about the choices of the subjects in the session, one for each game. Only in 3.2% of cases did a subject make a point prediction, that is, assigning a probability of 1 to the opponent choosing a particular amount. Even when adopting a looser definition for a point prediction, such as a subject assigning more than 50% probability to the opponent choosing a particular amount, this happens in only 27% of cases. We interpret this as clear evidence of a "common knowledge of noise". This lies in stark contrast to the assumption made in the level-k model about players believing that others best respond given their beliefs.

Finding 4. *In contrast to the assumptions of the level-k model, we find "common knowledge of noise" among subjects.*

Figure 7 shows the prediction of the NI model about subjects' beliefs. It should be noted that these are out-of-sample predictions, since the model is estimated using only the choices made by subjects in the first game ("11-20" Baseline). Comparing these predictions to the aggregate distribution of reported beliefs it is remarkable to notice how well the noisy introspection model captures the shape of these distributions.

Finding 5. *The NI model out-of-sample predicts beliefs in all six games very well.*

As we did with choices in the previous section, we compare the performance of the different models in predicting subjects' beliefs by calculating the mean square disstance between the predicted distribution of beliefs in each game and the aggregate distribution of reported beliefs. The results are reported in the right panel of Figure 9. The noisy introspection model performs better than all other models, while QRE is a close second. Level-k performs better than Nash equilibrium but still poorly compared to the other models.

Finding 6. *Beliefs predicted by Noisy Introspection reflect aggregate reported beliefs better than the ones predicted by alternative models.*

It is interesting to notice that subjects' guesses about how others play are remarkably accurate. This explains the very good performance of QRE in predicting beliefs. Remember that QRE is an equilibrium concept and therefore assumes beliefs are correct. In the next section we show that QRE underperforms compared to noisy introspection when it comes to individual behavior.

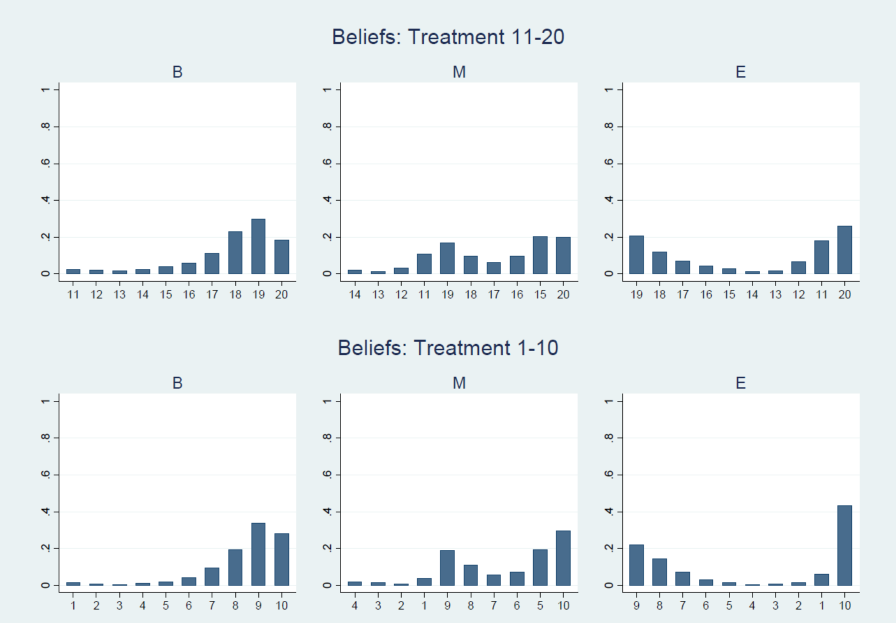


Figure 6: Realized distribution of beliefs.

4.4. Individual behavior

The discussion of results up to this point has focused mainly on subjects' aggregate behavior. We now turn our attention to indiviual choices. We observe significant heterogeneity in choices not only across subjects, but within subjects as well. That is, we find that subjects often switch to different choices when playing the same game again in different parts of the experiment. This behavior contradicts the predictions of the level-k model. According to these, subjects should consistently make the same choice when playing a particular game. On the other hand, both QRE and noisy introspection allow for "noisy behavior".

To compare the different models we calculate the expected number of times a particular subject will make the same choice in a particular game across all three parts (every time, only twice, never) based on the estimated models and compare it to the actual data. The results are shown in Figure **??**. The first thing one notices is how the level-k model fails to capture the within subject heterogeneity observed in the data. QRE tends to overestimate the number of times a subject does not repeat the same choice and underestimate the times a subject consistently repeats the same choice in all 3 parts. Noisy introspection comes closer to the actual frequencies observed in the data across all games.

Finding 7. *Behavior at the individual level is best captured by the Noisy Introspection model.*

Noisy introspection outperforms QRE because it in addition to the heterogeneity within subjects it also captures the across subjects heterogeneity. The latter is absent in the QRE model.

5. Conclusions

Arad and Rubinstein propose the 11-20 game as a tool to study level-k reasoning. We take on this task and introduce modifications in to the original game to study the possibilities of level-k to predict behavior. What we find is that the model's performance in five modified versions of the 11-20 game is no better than Nash equilibrium.

A moment's thinking on how individuals may play the 11-20 game, both in the modified versions as well as in its original form, point to an important element missing in the level-k model: payoff-dependent noise. Individuals' decisions are affected by it and individuals know that others are affected by it. To put it differently, there is "common knowledge of noise". Once this element becomes apparent, it becomes less surprising to see that the noisy introspection model that incorporates "common knowledge of noise" performs that well in predicting both behavior and beliefs out-of-sample.

Although we do not agree with Arad and Rubinstein about level-k's suitability for describing subjects' behavior, we do think that the 11-20 game they propose is very well suited to study strategic thinking. Small modifications, as the ones we do here, can allow experimenters to study a series of questions related to depth of reasoning, belief formation or learning in games.

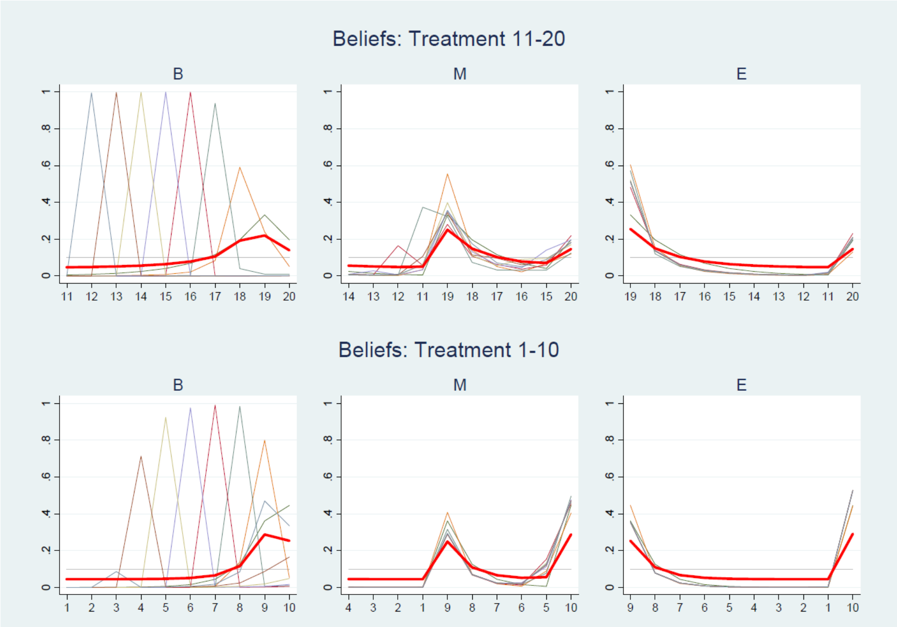


Figure 7: Predicted distribution of beliefs under Noisy Introspection Model.

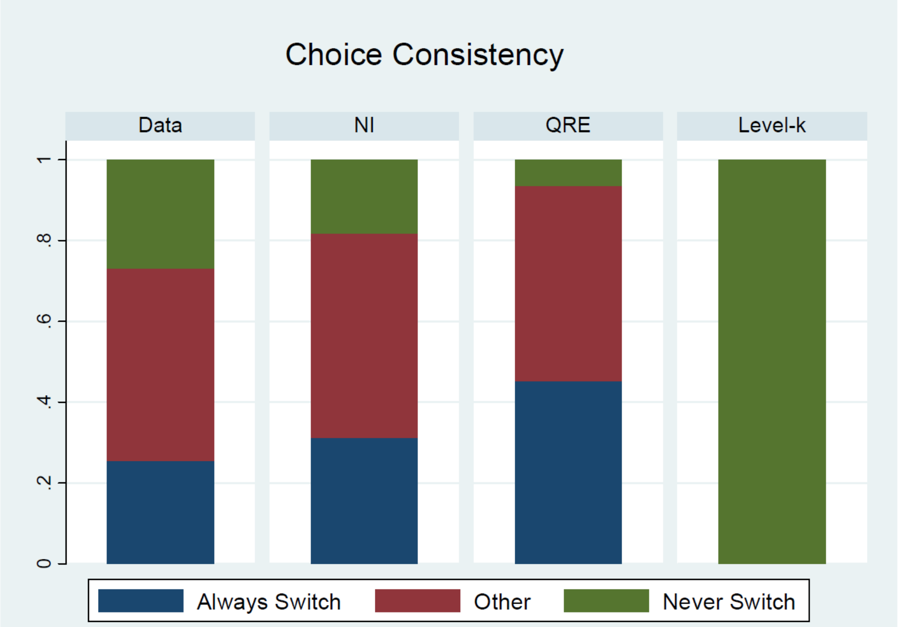


Figure 8: Observed distribution of choices.

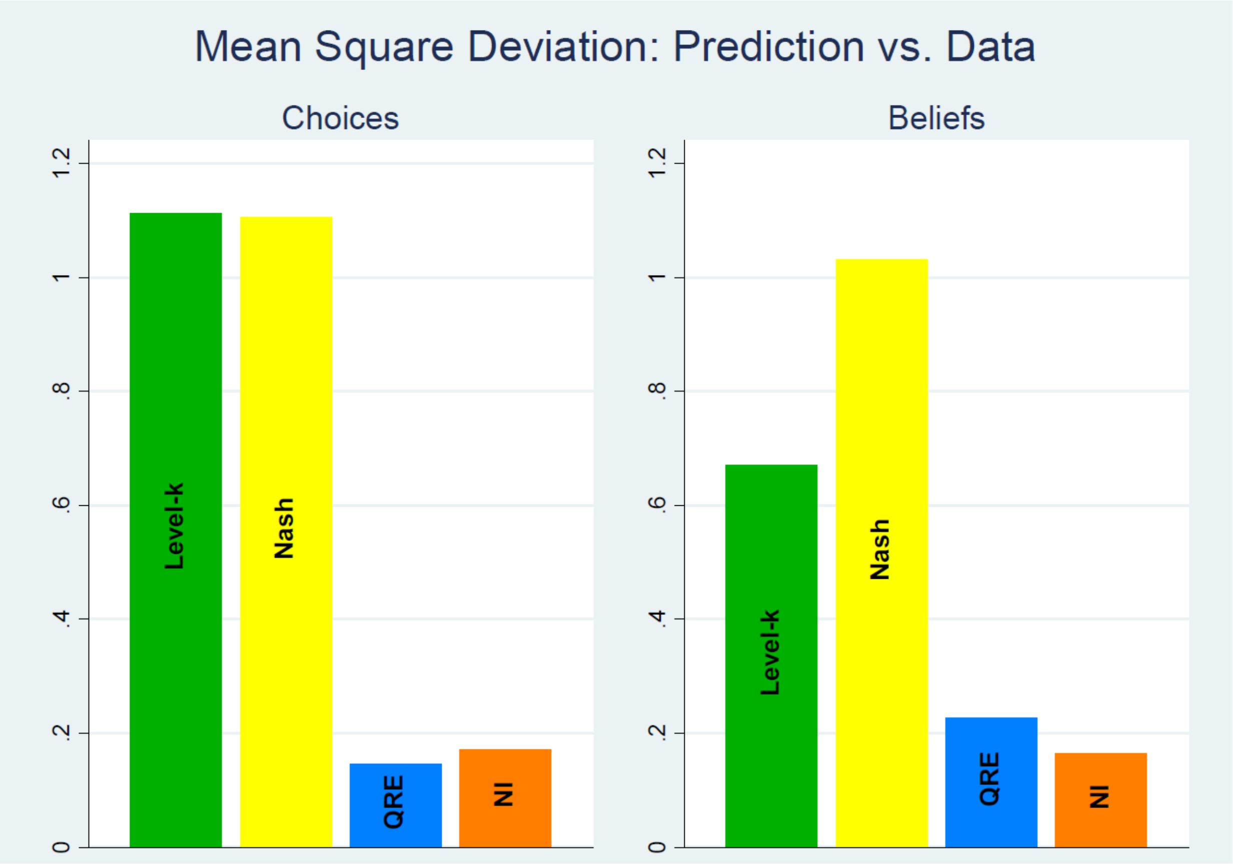


Figure 9: Mean squared distances between realized and predicted distribution.