Noisy Introspection in the "11–20" Game

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

March 8, 2016

Abstract

Previous experiments based on the 11-20 game have produced evidence for the level-k model with observed levels of strategic thinking consistently ranging from 0 to 3. Our baseline treatment uses the 11-20 game and replicates previous results. We apply four models of strategic thinking 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. Data from these additional treatments refute the level-k model, which predicts no better than Nash. A model of noisy introspection, which assumes "common knowledge of noise," predicts behavior remarkably well.

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

*Goeree: University of Technology Sydney, Economics Discipline Group and Centre for Policy and Market Design, PO Box 123, Broadway NSW 2007 Sydney, Australia; International Faculty, University of Cologne, Albertus-Magnus-Platz, D-50923 Cologne, Germany; Louis: ESEI Center for Market Design, University of Zürich, Schönberggasse 1, CH-8001 Zürich, Switzerland. Zhang: University of Technology Sydney, Economics Discipline Group and Centre for Policy and Market Design, PO Box 123, Broadway NSW 2007 Sydney, Australia. We benefitted from helpful comments made by seminar participants at University of Vienna, WZB-TU Colloquium, University of Technology Sydney, University of Sydney, Shanghai Jiaotong University, 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), the European Research Council (ERC Advanced Investigator Grant, ESEI-249433), and the Australian Research Council (ARC DP150104491) 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 (Stahl and Wilson, 1994, 1995; Nagel, 1995), which 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 level-0, 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 in Arad and Rubinstein's (2012) experiment, requesting amounts of 17, 18, or 19 (attributed to levels 3, 2, and 1 respectively) all yield expected payoffs 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 requests have some chance of yielding the bonus and those for which the loss in requested amount 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 disperse observed levels in the baseline game but 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

[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.

the notion of common knowledge of noise. We adapt the more general model here to allow for heterogeneity in levels of thinking in a way similar to the level-k model, but replacing 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. Allowing for heterogeneity in this form is not crucial for testing the noisy introspection model. We do it for two reasons. First, we find it is a parsimonious way to introduce heterogeneity in a model to be compared to homogeneous models that share similar features like QRE. Second, it facilitates comparison to the level-k model and allows us to pin-down differences in performance to the most salient difference of the model, namely the "common knowledge of noise" aspect.

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 distribution 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 five variations of the "11-20" game. As detailed below, the noisy introspection model predicts choices and beliefs strikingly well across all game variations.

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 Appendices contain additional estimation results and 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

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

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

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 conditions, 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 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.[4]

[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 consider both the possibility that level-0 chooses 20 and that level-0 chooses randomly.

[4] Notice that the term 'nested' here is not meant in the strict econometric sense. In fact, the specifications of the various models we estimate in the paper are non-nested, so that no a priori ranking in terms of fitness is possible.

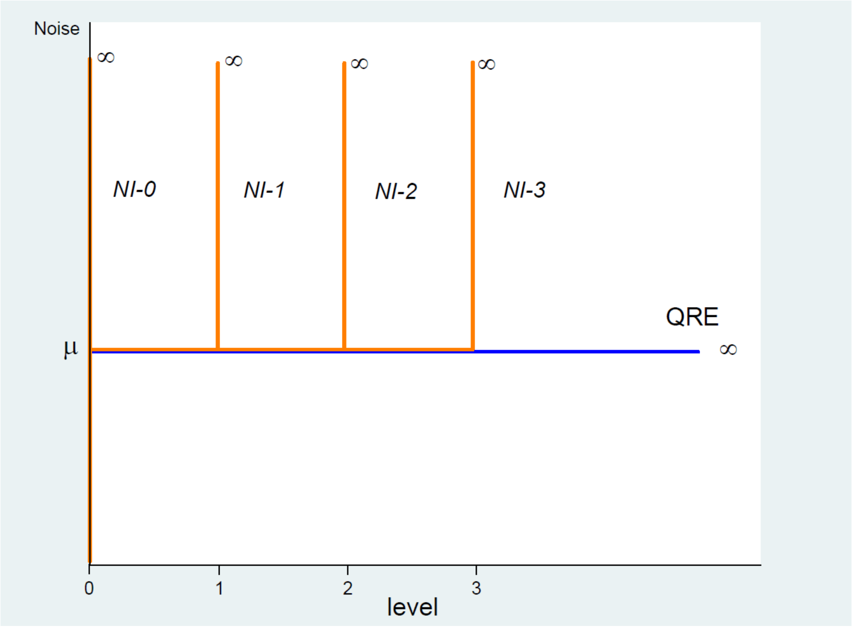


Figure 1: Various levels of noisy thinking in the NI-k model. Each line corresponds to a different sequence of error parameters $\mu_0 \leq \mu_1 \leq \ldots \leq \mu_\infty = \infty$. For example, the line labeled NI-0 corresponds to completely random decision making, which occurs if $\mu_0 = \infty$. 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 $\mu_k = \infty$.

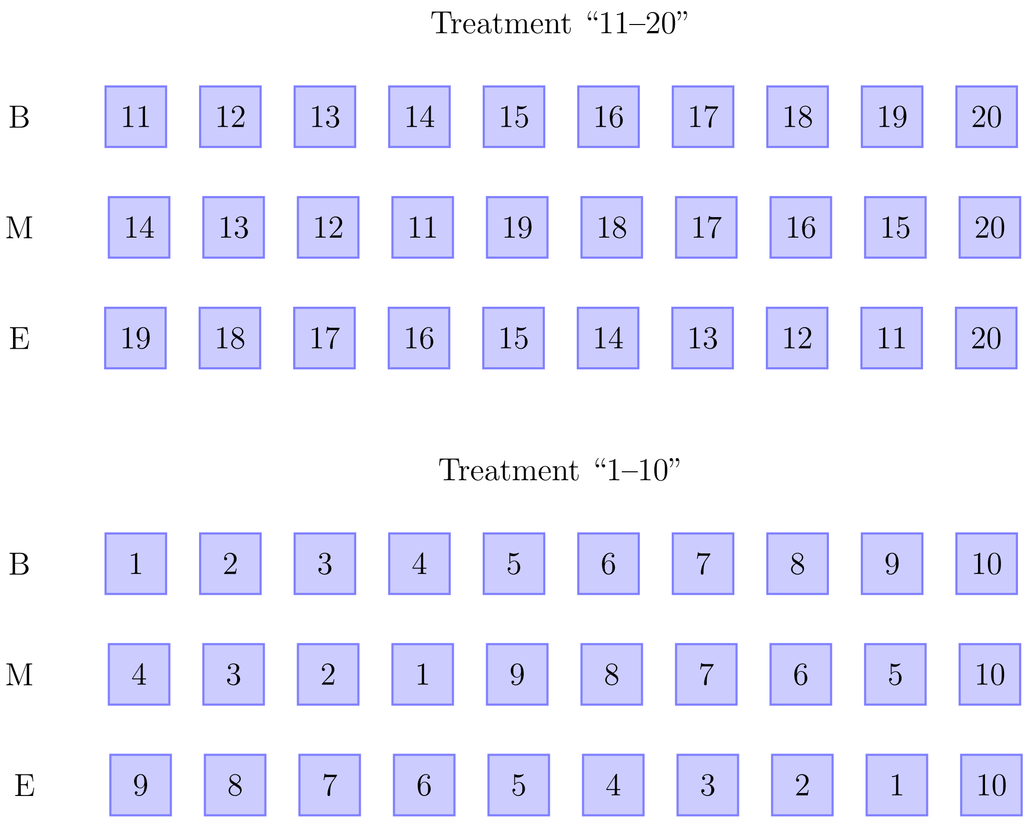
3. Experimental Design

The experiment used variations of Arad and Rubinstein's (2012) money request game, which were described as follows:[5]

> *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

[5] The complete set of instructions can be found in Appendix B.



Treatment "11–20"

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

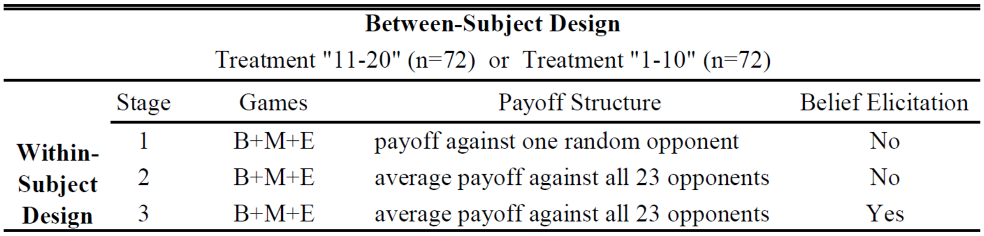
Treatment "1–10"

B	1	2	3	4	5	6	7	8	9	10
M	4	3	2	1	9	8	7	6	5	10
E	9	8	7	6	5	4	3	2	1	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 $R = 8$.

a choice of the highest number in the rightmost box would equal 5 Swiss Francs in either treatment.

Within a treatment there were three stages. Subjects were given separate instructions at the start of each stage and received no feedback about their payoffs until the end of the experiment. In stage 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 variations. In stage 2, subjects played the games in the same order as they had in stage 1, but now a subject's payoff was equal to the average payoff resulting from all possible matches (each



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

Table 1. Experimental Design

session had 24 subjects so there were 23 possible matches). Stage 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. Table 1 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 stage and subjects received their payoff in that game, plus the payoff from the belief elicitation process corresponding to the game picked from stage 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, 2015). The experiment was conducted in the Experimental Economics Lab of the University of Zurich using z-Tree (Fischbacher, 2007).

4. Experimental Results

The top panel of Figure 3 shows the distribution of choices made by the 72 subjects in the three variations of the "11-20" game, and the bottom panel shows choices for the other 72 subjects in three parallel variations of the "1-10" game.[6] For each game, we pool the choices

[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. The difference is significant for both game ($p < 0.05$, proportion test). In game B, the biggest difference is in the level-1 choices (23% to 35%, $p < 0.05$) whereas level-0 choices are almost the same (11% in

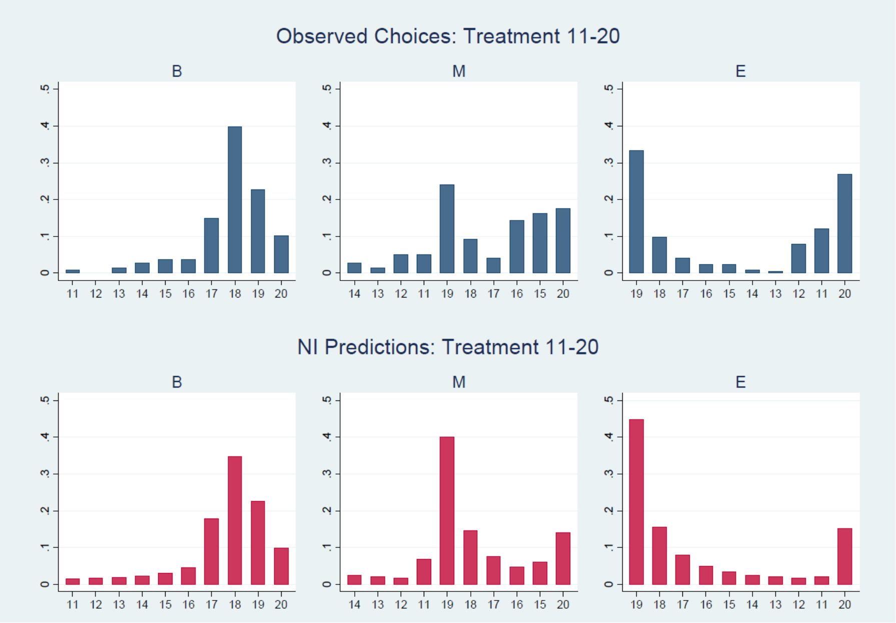


Figure 3: Observed (top) and predicted (bottom) choice distributions by game in the 11-20 treatment.

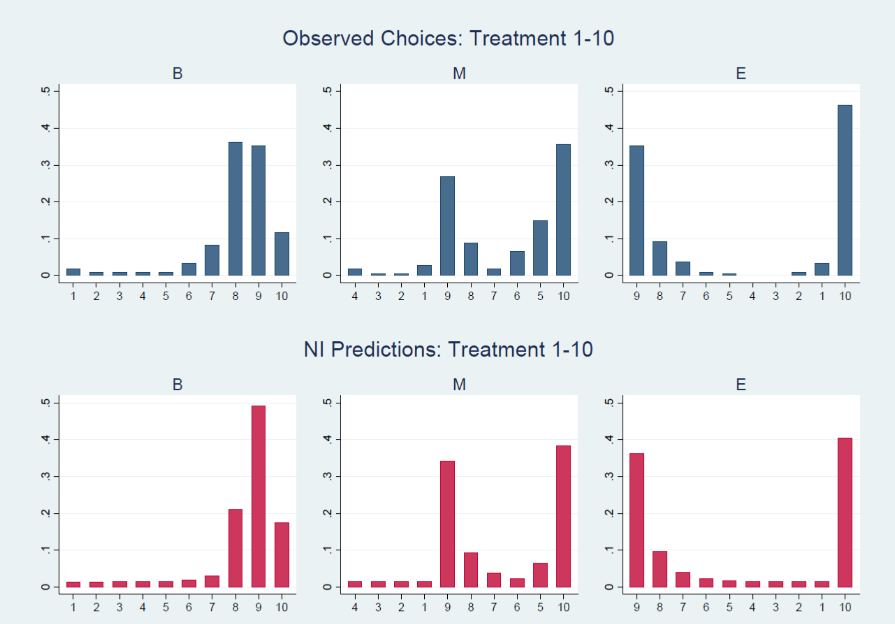


Figure 4: Observed (top) and predicted (bottom) choice distributions by game in the 1-10 treatment.

from the three different stages of the experiment.[7] The baseline game of the "11-20" treatment replicates Arad and Rubinstein's (2012) main findings: 10% of the choices correspond to level zero, 77% of the choices correspond to levels 1-3, and only 12.5% of the choices reflect a level higher than three. These percentages are not different at the 5% level from those reported by Arad and Rubinstein (2012): 6% level zero, 74% levels 1-3, and 20% levels higher than three.[8]

In what follows we study how well our experimental results are captured by the four models of Section 2: Nash, QRE, Level-k, and Noisy Introspection.[9]. First, we apply standard maximum-likelihood techniques to pin down the parameters of these models to fit behaviour in the baseline "11-20" game (B_{11-20}). Then we evaluate the performance of the different models in terms of their out-of-sample predictive power in the other five game variations.

One issue that needs to be addressed upfront is that both Nash and level-k have a "zero-likelihood problem." For instance, for the 11-20 baseline treatment the Nash equilibrium predicts that requests less than 15 should not be observed.[10] We deal with the zero-likelihood problem using two models of "noise" or "error." In one approach, players behave as predicted by the original model with probability $1 - \epsilon$ and with probability ϵ they randomize uniformly over all actions. Such **action trembles** are insensitive to their costs. Alternatively, the logit choice rule in (1) also allows for trembles to occur but such that their likelihood falls with the cost which we refer to as **payoff trembles**. The different error structures for Nash and level-k are shown in Table 2.

Another issue with level-k models is the specification of level-0 behaviour. In the baseline game, as in the original game in Arad & Rubinstein (2012), the best response of a level-1 is to choose 19, irrespectively of whether level-0 chooses 20 or randomises uniformly. This feature

"1-10" and 10% in "11-20"). All p-values reported in this paper are two-sided, unless otherwise stated.

[7] Recall that each experimental session consists of three stages that differ in the payment rule and whether or not beliefs were elicited, see Table 1. In each stage, 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 stage and no significant differences across the three stages (for all three games and in both treatments). In the analyses reported below we therefore pool data from all three stages, unless otherwise stated.

[8] Proportion tests comparing the three percentage pairs yield p-values of 0.208, 0.511, and 0.075 respectively.

[9] Another family of models to consider would be the Cognitive Hierarchy models (see Camerer, Ho and Chong 2004). In these a player of level k believes others to be from a distribution over all levels smaller than k. We omit analysis of such models as their performance is similar to some of the level-k or Nash models we analyse and it would not add much to our discussion.

[10] It is readily verified that there are no pure-strategy Nash equilibria and that any mixed equilibrium includes 20. Indifference between 19 and 20 dictates that 20 is played with probability 0.05. Likewise, indifference between 18 and 20 dictates that 19 is played with probability 0.10. This logic continues for lower request amounts until the choice probabilities add up to 1. In the mixed-strategy Nash equilibrium for the 11-20 baseline game the probabilities of each request amount between 11 and 20 are therefore (0,0,0,0,0.25,0.25,0.20,0.15,0.10,0.05). The Nash equilibria of the other game variations can be computed similarly.

is not preserved in our other games. We therefore also allow for these two different ways of specifying the behaviour of level-0. The second column of Table 2 specifies which level-0 behavior is used in each level-k model estimated

The QRE and NI models also employ the logit choice rule and, hence, they are not prone to a zero-likelihood problem. An important distinction is that in QRE and NI, players are aware that others' choices follow the logit rule, i.e. that their behavior is noisy. In contrast, Nash and level-k retain the best-response assumption and noise is only introduced to explain deviations from the model's predictions.

4.1. Data

A total of 72 subjects played the 11-20 version of games B, M, E, and another 72 subjects played the 1-10 version, see Figure 2. Let G denote the set of all six games. Each subject played all three games (B, M, and E) in each of the three stages of the experiment for a total of 9 choices. Let $x^i_{g,s}$ denote the observed choice of subject $i = 1, \ldots, 144$ in game $g \in G$ played in stage $s \in \{1, 2, 3\}$. Define $x^i_g = \{x^i_{g,1}, x^i_{g,2}, x^i_{g,3}\}$, $x^i = \{x^i_B, x^i_M, x^i_E\}$ and $x_g = \{x^1_g, \ldots, x^{144}_g\}$.

In each of the three games played in stage 3, each subject reported beliefs about the opponent's choices. Subjects reported their beliefs as the number of opponents, out of the 23 in the session, they believed would make one of the ten possible choices in game g. Let $b^i_g = \{b^i_{g,1}, \ldots, b^i_{g,10}\}$ denote the reported beliefs for subject i in game g, where each entry is a non-negative integer and the entries sum to 23, and define $b^i = \{b^i_B, b^i_M, b^i_E\}$.

4.2. Estimation using 11-20 baseline data only

We apply maximum-likelihood techniques to estimate parameter values for the different models using only the 11-20 baseline treatment. For Nash we estimate an error parameter $\epsilon \in [0, 1]$, which corresponds to trembles in actions, or a logit error parameter $\mu \geq 0$, which corresponds to cost-sensitive errors. For QRE we only estimate the latter. The level-k and NI models allow for heterogeneity among subjects. We model the distribution of types for each such model as a Poisson distribution (truncated at 9, the highest level type we can distinguish). This is characterised by a single parameter τ. Both parameters, τ and the common error parameter (ϵ or μ), are estimated in a finite mixture model. Let θ_M represent the set of parameters corresponding to model $M \in \{\text{Nash}, \text{QRE}, \text{level-}k, \text{NI}\}$, e.g. $\theta_{\text{QRE}} = \{\mu\}$ while $\theta_{\text{NI}} = \{\mu, \tau\}$.

Given a game g and parameter values θ_M, each model generates a probability distribution $p_M(a \,|\, \theta_M, g)$ over the set of possible actions $a \in A$. For example, for QRE this distribution

Baseline model	level-0	heterogeneity	payoff trembles	common knowledge of noise	error distribution	estimation's log-likelihood	error parameter	level distibution parameter	choice prediction performance	belief prediction performance
Nash					uniform	-465.59	$\epsilon = 0.228$ $_{(0.061)}$		25.3%	44.3%*
			√		logit	-431.90	$\mu = 2.298$ $_{(0.373)}$		62.7%	44.3%*
QRE			√	√	logit	-386.27	$\mu = 3.097$ $_{(0.249)}$		79.4%	65.8%
Level-k	20	√			uniform	-386.81	$\epsilon = 0.355$ $_{(0.034)}$	$\tau = 1.845$ $_{(0.187)}$	16.9%	-5.7%
	20	√	√		logit	-373.58	$\mu = 7.14$ $_{(0.507)}$	$\tau = 1.597$ $_{(0.240)}$	60.1%	-5.0%
	uniform	√			uniform	-396.18	$\epsilon = 0.332$ $_{(0.042)}$	$\tau = 1.613$ $_{(0.208)}$	39.4%	-2.0%
	uniform	√	√		logit	-372.72	$\mu = 1.295$ $_{(0.155)}$	$\tau = 0.969$ $_{(0.046)}$	48.5%	-0.9%
NI		√	√	√	logit	- 360.07	$\mu = 1.862$ $_{(0.152)}$	$\tau = 1.393$ $_{(0.174)}$	81.3%	63.7%

Table 2: Overview of the different models' specifications, estimation results and predictive performance. Estimation of the model's parameters and the corresponding log-likelihood is done using the choice data from the B_{11-20} game. Predictive performance is based on the estimated models' predictions for choices in the remaining five games and for beliefs in all games, including the B_{11-20} game. For performance we impose *strong consistency* of behavior across games.

* In calculating the likelihood score for beliefs in the Nash equilibrium we again encounter the zero-likelihood problem. We therefore assume that under this model reported beliefs are Nash with some probability $1-\lambda$ and purely random with probability λ. The likelihood score is maximised when $\lambda = .56$ and the resulting maximum score is reported here. It should therefore be interpreted as an upper bound for the performance of Nash in predicting beliefs and not be compared to the level-k models that perform worst than that.

follows from the fixed-point condition

$$p_M(a \,|\, \mu, g) \;=\; \phi_\mu(p_M(a \,|\, \mu, g)) \quad \forall a \in A$$

and is the same for all players, i.e. behavior is homogeneous. In contrast, in the NI model we allow for different types

$$p_M(a \,|\, \mu, k, g) \;=\; \overbrace{\phi_\mu(\phi_\mu(\cdots\phi_\mu}^{k \text{ times}}(\phi_\infty(a)))) \quad \forall a \in A$$

where $\phi_\infty(a) = 1/10$ for all $a \in A$, i.e. uniform randomization.

An individual's likelihood function evaluated at θ_M given the observed choices, x^i, in the set of games G for the homogeneous models is given by:[11]

$$L^i_M(\theta_M \,|\, x^i, G) \;=\; \prod_{\substack{g\,\in\,G\\ s\,=\,1\ldots3}} p_M(x^i_{g,s} \,|\, \theta_M, g)$$

and for models with heterogeneity by

$$L^i_M(\theta_M \,|\, x^i, G) \;=\; \sum_{k\,=\,0}^{9} f(k;\,\tau) \prod_{\substack{g\,\in\,G\\ s\,=\,1\ldots3}} p_M(x^i_{g,s} \,|\, \theta_M, g)$$

where $f(k;\,\tau) \;=\; \frac{e^{-\tau}\tau^k}{k!}/(\sum_{\ell=0}^{9}\frac{e^{-\tau}\tau^\ell}{\ell!})$ is the truncated Poisson distribution. The log-likelihood function evaluated at θ_M given the observed choices, x_g, in game $g \in G$ is then:

$$\mathrm{logL}(\theta_M \,|\, x_g, g) \;=\; \sum_{i\,=\,1}^{144} \log(L^i_M(\theta_M \,|\, x^i_g, g))$$

We obtain parameter estimates by maximizing the log-likelihood function, using data from the 11-20 baseline game only:

$$\theta^*_M \;=\; \underset{\theta_M}{\mathrm{argmax}} \left(\mathrm{logL}(\theta_M \,|\, x_g, g = B_{\text{11-20}})\right)$$

The parameter estimated values are summarized in Table 2.

[11] Subjects played variants of the 11-20 game or the 1-10 game, but not both. To keep the notation simple we use the convention that $p_M(x^i_{g,s} \,|\, \theta_M, g) \;=\; 1$ if subject i did *not* play a certain game g.

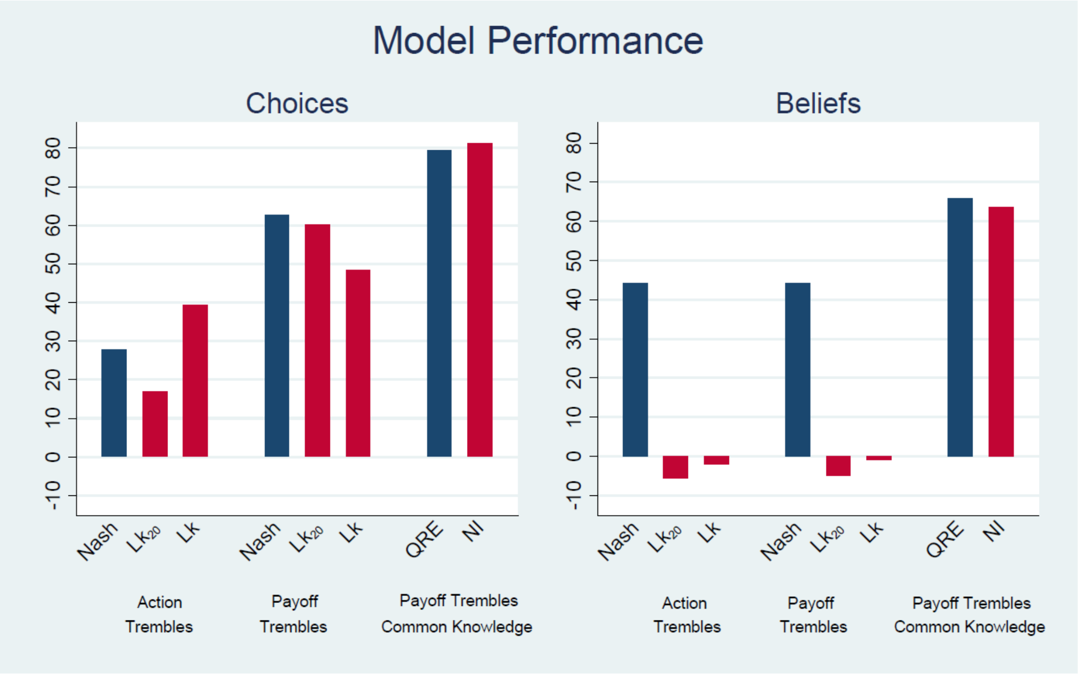


Figure 5: Predictive performance of the models against benchmarks.

4.3. Out-of-sample performance: choices

We next evaluate the out-of-sample performance of the various models. For this we use all games, including the 1-10 games, except for the B_{11-20} game that was used to estimate the models' parameters. We denote this set of games as $G' = G \setminus \{B_{11-20}\}$. The subjects that played the 1-10 games are different from the subjects whose 11-20 baseline choices were used to estimate model parameters. Still, there is no reason to suspect that there are systematic differences between the *pool* of 72 subjects that played the 11-20 game and the *pool* of 72 subjects that played the 1-10 games. The predicted choice distributions under the NI model are depicted in the lower panel of Figure 3 for treatment 11-20 and Figure 4 for treatment 1-10.

We will measure performance by the likelihood of the observed data given a model's prediction. Given a game $g \in G'$ and the estimated values, θ^*_M, shown in Table 2, each model generates a probability distribution over the possible actions $p_M(a \,|\, \theta^*_M, g)$. We use this to calculate each subject's likelihood for making the particular choices in all games in G'. We then take logarithms and sum up for all subjects to obtain the log-likelihood of the observed data

$$\mathcal{L}_M = \sum_{i=1}^{144} \log\left(L^i_M(\theta^*_M \,|\, x^i, G')\right)$$

Notice that we treat all choices made by a particular subject across all games she played as a single observation. This **strong consistency** requirement does not make a difference for homogeneous models but sets a higher bar for models with heterogeneity. It implies that a subject maintains his type across games. We believe this is the correct way of evaluating models with heterogeneity, unless one has a model of how individuals' types change across game froms.[12] However, our results are robust to imposing only **weak consistency**, i.e. when subjects' types are allowed to vary across games (see Appendix A).

To derive a score that lies between 0% and 100%, we compare this log-likelihood with two benchmarks. One is the upper-bound on the log-likelihood set by the model that exactly reproduces the choice frequencies observed in the experiment. The other is a "lower-bound" set by completely random choice. Let $n_g(a) = \sum_i \sum_s \mathbf{1}(x^i_{g,s} = a)$ denote the total number of a choices in game g then the upper-bound on the log-likelihood is given by

$$\overline{\mathcal{L}} = \sum_{g \in G'} \sum_{a \in A} n_g(a) \log\Big(\frac{n_g(a)}{3 \cdot 72}\Big)$$

The lower-bound based on uniform randomization is simply $\underline{\mathcal{L}} = (72 \cdot 6 + 72 \cdot 9) \cdot \log(1/10)$, since subjects made 3 choices from a set with 10 possible actions in each game, and we consider 2 games (we exclude the $B_{11\text{-}20}$ game) for the 72 subjects that played the 11-20 games and and all 3 games for the 72 subjects that played the 1-10 games. We can now define a model's likelihood score as

$$S^{\mathcal{L}}_M = \frac{\mathcal{L}_M - \underline{\mathcal{L}}}{\overline{\mathcal{L}} - \underline{\mathcal{L}}} \times 100\%$$

The likelihood scores for the different models are listed in the final column of Table 2 and are also shown in the left panel of Figure 5.

Result 1. *Based on likelihood scores, models that employ cost-sensitive payoff trembles predict choices significantly better than those based on action trembles.*

Support. The models are non-nested but a Vuong closeness test shows that the five models that use cost sensitive errors (five right bars in the left panel of Figure 5) perform better at the 0.001% level than the three models based on action trembles (three left bars). □

Result 2. *Based on likelihood scores, models that assume common knowledge of noise predict choices significantly better than those that don't.*

[12] There is an active literature focusing on the issue of persistence of types across games. See for example Georganas et al. (2015) and Cooper et al (2015). Alaoui and Penta (2015) develop a theoretical a model in which levels of thinking are determined endogenously.

Support. A Vuong closeness test shows that the QRE and NI models that assume common knowledge of payoff trembles (two right bars in the left panel of Figure 5) perform better at the 0.001% level than the three payoff-tremble models that don't (three middle bars). □

Interestingly, the homogeneous QRE model performs as well as the heterogeneous NI model, at least in terms of likelihood. The reason for good performance differs for the models, however. The QRE model has a higher error rate and, hence, results in "flatter" choice distributions than NI. When likelihood is the scoring criterion this helps in the sense that even though QRE is less likely to be "right," when it is "wrong" the penalty is not that high.

4.4. Out-of-sample performance: beliefs

The top panel of Figure 6 and Figure 7 presents the observed belief distribution by treatment and game. The bottom panel shows the predictions under the NI model. To measure how well each model predicts beliefs, we follow a similar procedure as described in the choice section.

Given a game, $g \in G$, and the estimated parameters, θ^*_M, each of the models predicts a belief distribution, $b_M(a \,|\, \theta^*_M, g)$, over the opponent's actions. We use $b^i_{g,a}$ to denote i's guess about how many others choose action a in game g. Like for choices, we require **strong consistency**. Thus, an individual's likelihood function for beliefs evaluated at θ_M, given the reported beliefs, x^i_g, in the set of games G for the homogeneous models is given by:[13]

$$B^i_M(\theta^*_M \,|\, b^i, G) \;=\; \prod_{\substack{g\,\in\,G \\ s\,=\,1\ldots3}} \prod_{a\,\in\,A} b_M(a \,|\, \theta^*_M, g)^{b^i_{g,a}}$$

and for models with heterogeneity by

$$B^i_M(\theta^*_M \,|\, b^i, G) \;=\; \sum_{k\,=\,0}^{9} f(k;\tau) \prod_{\substack{g\,\in\,G \\ s\,=\,1\ldots3}} \prod_{a\,\in\,A} b_M(a \,|\, \theta^*_M, g)^{b^i_{g,a}}$$

We then can define the log-likelihood for beliefs as

$$\mathcal{B}_M \;=\; \sum_{i\,=\,1}^{144} \log\left(B^i_M(\theta^*_M \,|\, b^i, G)\right)$$

[13] In both cases we ignore a multinomial coefficient, $\frac{(b^i_{g,1}+\ldots+b^i_{g,10})!}{b^1_{g,1}!\cdots b^i_{g,10}!}$, as it would also appear in the upper and lower bounds and therefore cancels when defining the likelihood score for beliefs.

Note that we now consider all six games, including the B_{11-20} game as reported beliefs were not used in model parameter estimation. The upper bound is given by

$$\overline{\mathcal{B}} = \sum_{i=1}^{72} \sum_{g \in G} \sum_{a \in A} b^i_{g,a} \log \left(\frac{\sum_i b^i_{g,a}}{23 \cdot 72} \right)$$

while the lower bound is $\underline{\mathcal{B}} = 144 \cdot 96 \cdot \log(1/10)$. The likelihood score is then

$$S^{\mathcal{B}}_M = \frac{\mathcal{B}_M - \underline{\mathcal{B}}}{\overline{\mathcal{B}} - \underline{\mathcal{B}}} \times 100\%$$

The calculated values for all models are presented in the right panel of Figure 5.

Result 3. *Based on likelihood scores, level-k models predict beliefs significantly worst than uniformly random beliefs.*

Support. A Vuong closeness test shows that all four level-k model specifications perform worst at the 0.001% level than a model in which beliefs are draws from a uniform distribution over all possible choices. This random model defines the 0% limit for the likelihood score. □

Result 4. *Based on likelihood scores, models that assume common knowledge of noise predict beliefs significantly better than those that don't.*

Support. A Vuong closeness test shows that QRE and NI perform significantly better than Nash and the level-k models at the 0.001% level. The difference between QRE and NI is not statistically significant (p-value = 0.35). □

4.5. Choice Consistency

Up to this point we find QRE and NI are the two winning models in predicting subjects' aggregate behavior and there is no significant difference between these two. To further compare the performance of these two models, we turn our attention to individual choices. There is significant heterogeneity in choices not only across subjects, but within subjects as well. Subjects often switched to different choices when playing the same game again in different stages of the experiment. To evaluate how well either model predicts individual switching patterns, we calculate the expected number of times a particular subject will make the same choice in a particular game across all three stages (every time, only twice, never) based on the estimated models and compare it to the actual data. The results are shown in Figure 8.

The QRE model captures some heterogeneity but it tends to overestimate the number of times a subject never repeats the same choice and underestimate the times a subject consistently

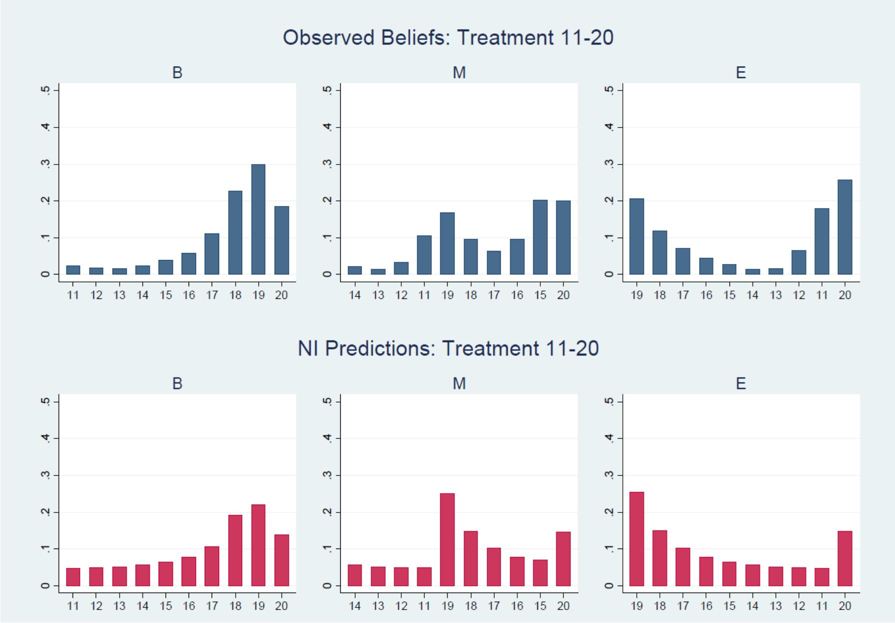


Figure 6: Observed (top) and predicted (bottom) belief distributions by game in the 11-20 treatment.

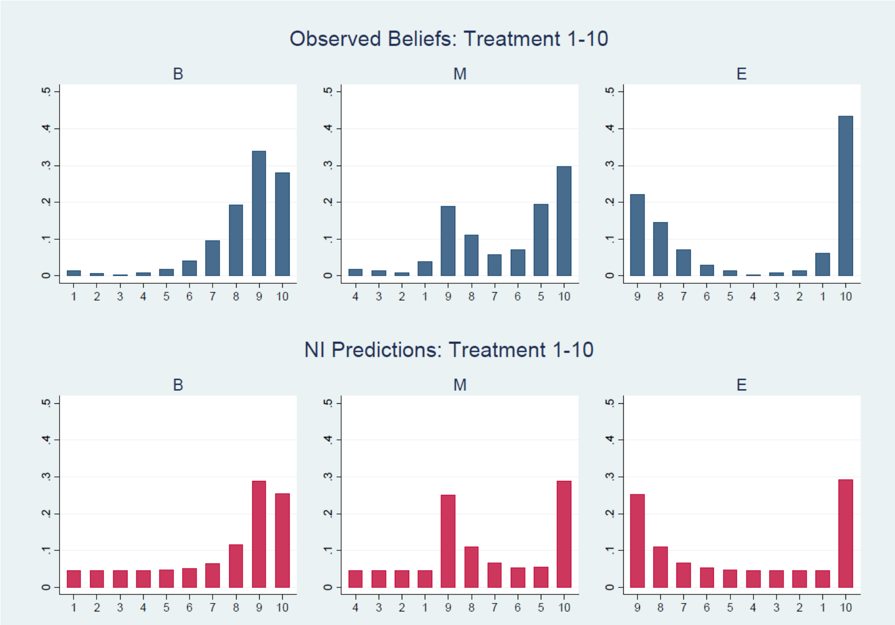


Figure 7: Observed (top) and predicted (bottom) belief distributions by game in the 1-10 treatment.

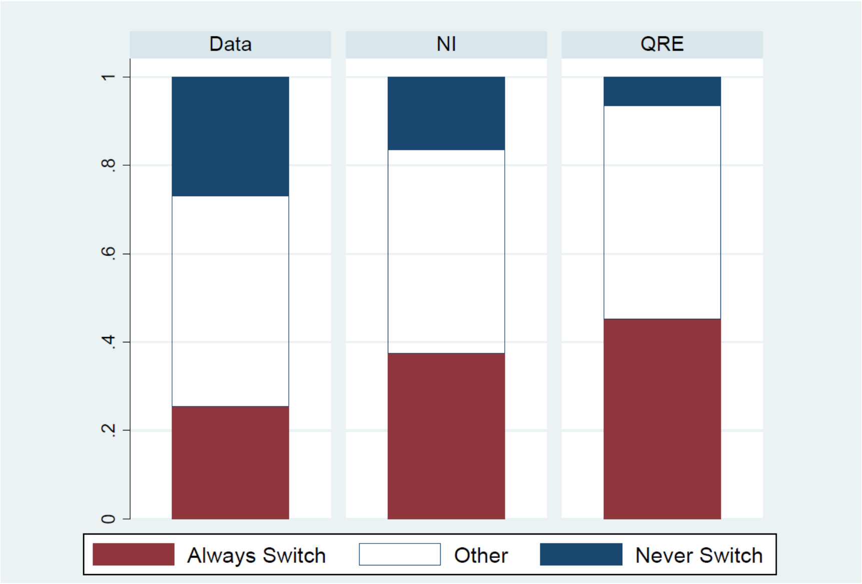


Figure 8: Observed distribution of choices.

repeats the same choice in all three stages.[14]

Result 5. *The Noisy Introspection model predicts switching behavior significantly better than the QRE model.*

Support. The Fisher's exact test reports significant difference in the overall distribution between the QRE and NI predictions, $p < 0.001$. The predicted percentage of always switching behavior falls from 45.2% under the QRE to 37.4% under the NI model, getting closer to the observed 25.5%, and this difference is significant according to a proportion test, $p = 0.02$. The predicted percentage of never switching behavior increases from 6.4% under the QRE to 16.5% under the NI model, moving towards the observed 26.9%, and also this difference is significant according to a proportion test, $p < 0.001$. □

Noisy introspection outperforms QRE because it predicts heterogeneity across subjects and choice consistency within subjects. It should be pointed out though that the NI predictions are significantly different from the data at the 1% level according to a Fisher exact test.

[14] More specifically, a two-sided proportion test shows significant difference at 1% level between QRE prediction and actual data in the percentage of always switching behavior and the percentage of never switching behavior. Using the Fisher's exact test to compare the distributions across three categories reveals significant difference at 1% as well.

5. Conclusions

Arad and Rubinstein (2012) suggest the 11-20 game as a tool to study level-k reasoning. We concur but propose to take their suggestion a step further: the 11-20 game plus some variations form an ideal tool to study a variety of models of strategic thinking, not just level-k. After all, choice behavior in the basic 11-20 game is well explained by several models and it is natural to explore game variations that can discriminate between them. More generally, small variations, such as the ones proposed here, allow experimenters to probe a series of questions related to depth of reasoning, belief formation, and learning in games.

We find that the level-k models that assumes best-response behaviour does not perform better than Nash. Relaxing this assumption and allowing for payoff dependent noise seems to be a step in the right direction in terms of performance with respect to our data, although predictions for specific games may be significantly different than the ones of the simple best-response model. Even so, an important ingredient seems to be missing. Our data on beliefs indicate that players are aware of the noise in others' behaviour. In fact we find that the best performance across all variations of the 11-20 game we used is achieved by the models that incorporate such "common knowledge of noise": QRE and noisy introspection.

There is one ingredient of the level-k model that while not decisive, does seem to reflect an important feature of the data: heterogeneity. The noisy introspection model we estimate extends the homogeneous model of Goeree and Holt (2004) with a hierarchy of types. This element moves predictions closer to the data compared to the QRE model, although still not close enough. In light of these results we encourage further investigation to understand the heterogeneity in strategic thinking, but we strongly encourage this to be done in a framework where payoff dependent noise and "common knowledge of noise" are explicitly accounted for. The noisy introspection model provides such a framework.

References

Alaoui, L. and Penta, A. (2015) "Endogenous Depth of Reasoning," *Review of Economic Studies*, forthcoming.

Arad, A. and Rubinstein, A. (2012) "The 11-20 Money Request Game: A Level-*k* Reasoning Study," *American Economic Review*, 102(7), 3561–3573.

Cooper, D. J, Fatas, E., Morales, A. and Qi, S. (2015) "The Types They Are a-Changin': An Experimental Study of Persistence of Types in Level-k Models," Paper presented at: *Behavioral Game Theory Workshop*, June 23-24, 2015, University of East Anglia, Norwich, UK.

Georganas, S., Healy, P. J. and Weber, R. (2015) "On the Persistence of Strategic Sophistication," *Journal of Economic Theory*, 159(A), 369–400.

Fischbacher, U. (2007) "z-Tree: Zurich Toolbox for Ready-made Economic Experiments," *Experimental Economics*, 10, 171–178.

Goeree, J.K. and Holt, C.A. (2001) "Ten Little Treasures of Game Theory and Ten Intuitive Contradictions," *American Economic Review*, 91(5), 1402–1422.

Goeree, J.K. and Holt, C.A. (2004) "A Model of Noisy Introspection," *Games and Economic Behavior*, 46(2), 365–382.

Greiner, B. (2015) "Subject Pool ecruitment Procedures: Organizing Expriments with ORSEE," *Journal of the Economic Science Association*, 1, 114–125.

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

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

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

Stahl, D.O. and Wilson, P.W. (1995) "On Players' Models of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior*, 10(1), 218–254.

A. Appendix: Estimation Results Under Weak Consistency

The predictive performance of the estimated models is calculated imposing *strong consistency.* While this requirement seems to us to be appropriate given the models' assumptions, one might worry that our results are driven by it. Here we show that similar results are obtained if we only apply *weak consistency*, which treats a subject's choices or reported beliefs in one game as independent from those in another game. In fact, results are even more favourable to the noisy introspection model than under *strong consistency.* Notice that this affects only the level-k and noisy introspection models. Since QRE and Nash are homogeneous models it does not make a difference whether choices are considered independently for a game or even for each subject: log-likelihood is calculated by taking the logarithm of a product of products. On the other hand, for models with heterogeneity we estimate a finite mixture model and the log-likelihood is the logarithm of a product of a weighted sum of products. Formally, relaxing consistency from strong to weak results in the following individual's likelihood function for choices in models with heterogeneity:

$$L^i_M(\theta_M \,|\, x^i, G) \;=\; \prod_{g\,\in\,G} \left(\sum_{k\,=\,0}^{9} f(k;\tau) \prod_{s\,=\,1}^{3} p_M(x^i_{g,s} \,|\, \theta_M, g) \right)$$

and for beliefs:

$$B^i_M(\theta_M \,|\, b^i, G) \;=\; \prod_{g\,\in\,G} \left(\sum_{k\,=\,0}^{9} f(k;\tau) \prod_{s\,=\,1}^{3} \prod_{a\,\in\,A} b_M(a \,|\, \theta_M, g)^{b^i_{g,a}} \right)$$

The likelihood scores obtained given *weak consistency* are reported in Table 3.

Baseline model	level-0	heterogeneity	payoff trembles	common knowledge of noise	error distribution	choice prediction performance	belief prediction performance
Nash					uniform	27.8%	21.0%
			√		logit	62.7%	55.5%
QRE			√	√	logit	79.4%	65.8%
Level-k	20	√			uniform	35.7%	-0.8%
	20	√	√		logit	60.5%	1.5%
	uniform	√			uniform	54.0%	9.8%
	uniform	√	√		logit	69.3%	12.9%
NI		√	√	√	logit	82.8%	89.6%

Table 3: Predictive performance for choices and beliefs under *weak consistency*. Prediction with weak consistency is based on the same estimated parameter values used for prediction with strong consistency and reported in Table 2.

B. Appendix: Experimental Instructions

Instructions – Treatment “11-20”

Thank you for participating in this session. The experiment will involve a series of decisions. Each of you may earn different amounts. The amount you earn depends on your decisions and those of others. The exchange rate used in the experiment is **4 points** for **1 CHF.** You also receive a **10 CHF participation fee**. Upon completion of the experiment, you will be paid individually and privately. **24** participants take part in today’s experiment.

Please remain quiet!
You will be using the computer terminal for the entire experiment, and your decisions will be made via your computer terminals. Please DO NOT talk or make any other audible noises during the experiment. If you have any questions, raise your hand and your question will be answered so that everyone can hear.

The experiment consists of three parts. You will now receive instructions on the first part of the experiment.

Part 1

The game: You and another participant in the experiment are randomly matched to play the following game:

13	16	20	18	19	12	15	11	17	14

On your screen you will see 10 boxes in line, containing different amounts (like the ones above). Each participant requests an amount of points. The amount is chosen 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 20 points if the selected amount is exactly **'one to the left'** of the amount that the other participant chooses.

Profit: You will play this game three times. One of these three rounds will be chosen randomly and your earnings in that round will be your profits for this part of the experiment. You will be informed about your profits in this and the other parts at the end of the experiment. **Please take your time and think through each of your decisions as you will be paid for only one of them.**

Example:

Suppose the boxes are arranged as shown above, and you choose 12:

- If your opponent chooses 17, you earn 12 points and he earns 17.
- If your opponent chooses 15, you earn 12+20= 32 points and he earns 15 points.
- If your opponent also chooses 12, you both earn 12 points.
- If your opponent chooses 19, you earn 12 points and he earns 19+20=39 points.

Part 2

The game: In this part of the experiment you will play the same game as in part 1, except that now your choice will be matched against each of the choices made by all the other participants in the experiment. Your earnings in points will be equal to the average earnings you get from playing against all other participants.

Example

Suppose there are only 4 participants in the experiment: you, Participants 1, 2 and 3. The amounts are as follows:

13	16	20	18	19	12	15	11	17	14

You choose 15 and the other's choices are:

- Participant 1 chooses 16
- Participant 2 chooses 19
- Participant 3 chooses 11

Your earnings against each other participant are:

- Against Participant 1: 15
- Against Participant 2: 15
- Against Participant 3: 35

Your average earnings in this round are:

$$\frac{Sum\ of\ earnings\ against\ each\ participant}{\#\ of\ other\ participants} = \frac{15+15+35}{3} = 21.7$$

Profit: You will play this game three times. One of these three rounds will be chosen randomly and your average earnings in that round will be your profits for this part of the experiment.

Part 3

The game

In this part of the experiment you will play the same game as in part 2, except that this time before you make a choice you will have to guess what the other participants will do. Below each of the 10 amounts you will see a box where you can enter your guess about the number of other participants you think will choose that amount. The sum of these numbers must be equal to 23, the number of other participants in today's experiment.

You will earn additional points from accurately guessing others' choices. The closer your guess ends up being to the actual distribution of choices, the more points you earn. In particular, the following formula will be used to determine how many points you earn from your guess:

$$Points = 20 - \frac{10}{23^2} \times \sum_{over\ all\ amounts} (Guess_{amount} - Realized_{amount})^2$$

Example

There are 23 other participants and the amounts are arranged as follows:

13	16	20	18	19	12	15	11	17	14

Suppose that you guess that:

- 5 participants will choose 16
- 3 participants will choose 18
- 9 participants will choose 12
- 4 participant will choose 15
- 2 participant will choose 11

Then you should fill out the boxes as follows:

13	16	20	18	19	12	15	11	17	14
0	**5**	**0**	**3**	**0**	**9**	**4**	**2**	**0**	**0**

(continue reading on the back)

Now suppose it turns out that:

- 8 participants chose 16
- 3 participants chose 18
- 4 participants chose 12
- 8 participants chose 17

The following table shows your guess and the realized choices, as well as the number that corresponds to the square of the difference between your guess and the realized number, for each amount

Amount	**13**	**16**	**20**	**18**	**19**	**12**	**15**	**17**	**11**	**14**
Guess	0	5	0	3	0	9	4	0	2	0
Realized	0	8	0	3	0	4	0	8	0	0
$(\boldsymbol{Guess_{amount}} - \boldsymbol{Realized_{amount}})^2$	0	9	0	0	0	25	16	64	4	0

The formula then gives:

$$\boldsymbol{Points} = 20 - \frac{10}{23^2} \times (0 + 9 + 0 + 0 + 0 + 25 + 16 + 64 + 4 + 0) = \mathbf{17.95}$$

Given this formula, you maximize the expected number of points you earn by giving a guess that is as close as possible to your true estimate of what other participants will do.

Profit

You will play this game (guess + choice) three times. One of these three rounds will be chosen randomly and your average earnings in that round (guess + choice) will be your profits for this part of the experiment.

After finishing this part you will be presented with a table indicating your profit for the experiment. Please wait patiently at your seat until you are called to collect your profit.