Estimating Bounded Rationality And Pricing Performance Uncertainty

C. Nicholas McKinney Jr. and John Van Huyck

August 2004

Abstract: This paper estimates bounded rationality in two outcome strictly competitive extensive form games of perfect information without chance moves. In the baseline sessions, the average subject can reason effectively to about rank 6. Twelve percent of these subjects have a perfect bound on their ability to achieve the substantively rational outcome: they win all simpler games and they lose all games more complicated than the estimated rationality bound. All subjects in the baseline sessions are within 5 classification errors of a perfect bound. Most subjects don't have an accurate expectation about their performance against a procedurally rational opponent. Usually, they are overconfident and value the games at more than they are likely to earn if they actually play them. The paper also reports robustness checks with respect to subject pool and game samples.

Key Words: Bounded rationality, performance uncertainty, perfect information, Nim, human behavior, experiment.

JEL Classification: c72, c78, c92, d83

Acknowledgments: Nick Rupp and Chris Ball provided able research assistance. Ted Bergstrom inspired our investigation of subject pool differences. Tom Palfrey, Sean Gailmard, Walter Yuan, and Julie Malmquist assisted us in various ways collecting the Caltech subject data. Ray Battalio provided helpful advice on software development. Jana Joswiak helped with recruiting and accounting. Andreas Ortmann brought the literature on capacity limits for the focus of attention to our attention. The Private Enterprise Research Center and the National Science Foundation provided financial support. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Private Enterprise Research Center or the National Science Foundation.

I. Introduction

Game theory provides a compelling solution to strictly competitive extensive form games of perfect information without chance moves: PI-games for short. Backwards induction on the extensive form even provides a solution algorithm. This solution was first discussed by Zermelo (1913). The value of the game when played by people capable of computing Zermelo's algorithm can in theory be determined. Reasonable people ought to agree that there exists a substantively rational way to play PI-games.[1]

However, practitioners of PI-games, like checkers or chess, almost never refer to the algorithm, much less use it, even in high stakes games with millions of dollars on the table. Chess books written for practitioners never mention the idea that chess has a value. Instead, they focus on developing the ability to recognize and create winning positions. The reason, of course, is that Zermelo's algorithm when applied to chess is not computable by either man or machine. It is not a procedurally rational way to play, see Simon and Schaeffer (1992).

There are some PI-games for which clever people have found ways to construct algorithms or procedures that achieve the outcome prescribed by substantive rationality: examples include Tic-tac-toe, Connect Four, and Nim. Nim is a two player game. Players alternate turns. During a turn the player must remove one or more stones from any one of the *m* rows. The player who takes the last stone wins. The rank, the number of moves in the longest play path, of a Nim game can vary from 1 to an arbitrarily large number. Algorithms to solve Nim are elegant because they apply to the whole class of Nim games and, unlike backwards induction, they do not require the construction of the games extensive form. Since Nim is a game of varying complexity with both a substantively and procedurally rational way to play, it is ideally suited to measure the bounds of human rationality.

[1] Which is not the same thing as saying it is not controversial. The definition slips in a mutual consistency requirement that has been criticized by Selten (1975), Binmore (1987) and many others. If the value of the game is not a player's most preferred outcome and one knows they are playing against a player who can not find a winning move when confronted with certain positions, then effective play requires exploiting this known weakness in an opponent's behavior, see Luce and Raiffa (1957, section 4.11) for an early discussion and a warning about sandbaggers.

Performance uncertainty is the uncertainty a player has about his ability to implement a substantively rational strategy and, hence, earn the value of a PI-game. One possibility is that there is always a small chance of suboptimal performance at every information set and this chance doesn't depend on the overall complexity of the continuation game. Selten (1975) has argued that performance uncertainty is an essential component of a model of substantive rationality. Including a tremble at every information set forces rational players to consider what happens off the equilibrium path. Here we investigate not only a trembling hand bound on rationality, but also the possibility that performance deteriorates rapidly as the complexity of a situation exceeds some threshold level.

It is not obvious that a boundedly rational player of this sort will possess self-knowledge, since they cannot satisfy basic consistency requirements of rational choice, like the reduction of compound lotteries. To investigate self-knowledge we use a Becker, DeGroot, Marshak (1964) like protocol to elicit a player's beliefs about the profitability of playing a particular game of Nim against a procedurally rational opponent.

The main results are that the average subject can reason effectively about Nim up to about rank 6. People differ in their ability to reason about Nim, we have observed players able to solve all games up to rank 17. Most people don't have an accurate idea about their ability to play Nim. Usually, they are overconfident and value the games at more than they are likely to earn if they actually play them. Experience appears to reduce overconfidence.

II. Analytical Framework

A game of Nim can be summarized by a $1\times m$ vector of natural numbers, g, with elements g_i denoting the number of stones in row i. Notice that this requires a 'row' to have at least one stone. Let G denote the set of all Nim games, where $G = \{g \in \mathbb{N}^m \mid m \in \mathbb{N}\}$ and $\mathbb{N}$ denotes the natural numbers. The rank of a game is the length of the longest play path through its extensive form. The rank of Nim game g, $r(g)$, is equal to the number of stones used in the game: $r(g) = \sum_{i=1}^{m} g_i$.

Consider using rank as a complexity measure. Intuitively, the longer the play paths the more difficult it is to think through. By the rank measure, (5,3,4,5) is more complex than (1,1,1), because 17 is more than 3. While the rank provides a useful and general measure of complexity, it may miss essential elements limiting human rationality. A game with a large rank may be easy to solve if it has a short winning play path.

An alternative measure is the complexity introduced by the shortest play path through the extensive form. The shortest play path would predict that (1,1,1) with a shortest play path of 3 was more complex than (3) with a shortest play path of 1, which is counter intuitive. It is not possible to lose game (1,1,1) as the first mover, because the game contains no decision nodes in which it is possible to make a choice that changes the value of the game. It is a trivial game.

Denote decision nodes that can't change the value of the game trivial decision nodes. It is easy to see that all games with $r = m$ are trivial games and it is only slightly more difficult to show that all games with $r > m$ are non-trivial games. An alternative measure of complexity, denoted NT-complexity, is obtained by counting the number of non-trivial decision nodes in the extensive form of the game. The choices made at non-trivial decision nodes matter because they can potentially change the value of the game.

For example, the following games all have the same rank of 7: (1,1,1,1,1,1,1), (1,1,1,4), (1,1,5), and (7). The NT-complexity measures for the four rank 7 games are 0, 220, 124, and 32 respectively, which corresponds to our intuition about their relative complexity.

The NT-complexity measure captures a general aspect of PI-games. Nim is special in that it can be solved using rules-of-thumb or tricks like strategy copying or take all stones in games with a shortest play path equal to 1. So some nodes that are measured as non-trivial by the general NT-complexity measure are trivial for players who are aware of some applicable Nim solving trick, like strategy copying. However, further modification of the measure would require assumptions about players' ability to recognize Nim tricks and implement the applicable strategies.

A forth measure of complexity is the probability a blunderer wins the game when making uniformly random legal moves and playing against a procedurally rational opponent. A blunderer wins (1,1,1,4), (1,1,5), or (7) with probability 0.143, but wins the trivial game (1,1,1,1,1,1,1) for sure.

By Zermelo's theorem we know that every Nim game $g \in G$ has a value. The value is either a win for the first mover, W, or a loss, L. Let the value function be denoted $v(g) : G \rightarrow \{W,L\}$.

A general computable algorithm to determine the value of g must not require construction of the extensive form representation of g. There exist several algorithms for computing $v(g)$ that are procedurally rational, see Bouton (1902) and Conway (1976). Here, we describe the Bouton algorithm based on Binmore (1992).

Convert the decimal representation of the natural numbers in g into the equivalent binary representation. Let $b(g)$ denote the binary representation. For example $b((1,1,5)) = (001, 001, 101)$. Let $d_j(g_i)$ denote the digit in the 2^{j-1} position of $b(g_i)$, where $d_j(g_i) \in \{0,1\}$. For example, consider the 5 stones in the last row. The binary representation is 101, thus $d_2(5) = 1$, $d_1(5) = 0$ and $d_0(5) = 1$.

The set of balanced Nim games, B, is $B = \{g \in G | \sum_{i=1}^{m} d_j(g_i) \text{ is even } \forall j\}$ and the set of unbalanced Nim games, U, is $U = \{g \in G | g \notin B\}$.

To understand why the algorithm works we review a few propositions. The first mover in a balanced game cannot win on his first move, because a balanced game has stones in at least two rows and a player may only remove stones from one row during a turn.

Proposition 1: Any balanced game g must be left unbalanced.

Proof: Since g is balanced, $\sum_{i=1}^{m} d_j(g_i)$ is even for all j. A player must remove at least one stone, which will change the sum for at least one value of j by 1, which leaves it odd.

Proposition 2: Any unbalanced game g can be left balanced.

Proof: Since g is unbalanced, $\sum_{i=1}^{m} d_j(g_i)$ is odd for at least one j. Let J denote the set of binary positions j with odd sums. Find a row i with the largest value of j such that $\sum_{i=1}^{m} d_j(g_i)$ is odd and remove stones from that row in such a way as to change each of the $\sum_{i=1}^{m} d_j(g_i)$ for $j \in J$ by 1, which leaves all of the sums even and a balanced game.

Proposition 3: The first mover in an unbalanced game of Nim wins and in a balanced game of Nim loses. Proof: Consider the first mover in an unbalanced game first. By proposition 2 it is always possible to find a way to convert an unbalanced game into a balanced game. A player who does so always leaves his opponent with an unwinnable position. An opponent who can never win on the next move can never win at all. Since the game is finite and has only two outcomes, the first mover must win.

Consider the first mover in a balanced game. By Proposition 1, the first mover must convert it to an unbalanced game in which case the second mover can force a win by the argument in the first part of the proof.

The value function can be expressed in our notation as follows:

$$v(g) = \begin{cases} W \text{ if } \ g \in U \\ L \text{ if } \ g \in B \end{cases}.$$

Substantively rational players win all unbalanced games and when playing against other substantively rational players lose all balanced games.

III. Pricing Performance Uncertainty

Suppose that a player earns \$*z* for a win and \$*y* for a loss, where \$*z* > \$*y*. A substantively rational player ought to be willing to pay as much as \$*z* for the first movers role of an unbalanced game no matter what his risk attitude, because there is no uncertainty about the outcome of the game. Becker, DeGroot, Marshak (1964) describe a (BDM) procedure for eliciting the value of a lottery, given people who conform to the axioms of subjective expected utility. It is straight forward to show that the procedure also works for two outcome PI-games.[2]

---

[2] The quadratic scoring rule can not be generalized to PI-games. Unlike a lottery, a player can throw a winning position and thus be sure the outcome will be a loss.

Our adaptation of the BDM procedure endows a player with a two outcome PI-game *x* and then the player names an asking price at which they would be willing to sell the game. A bid price is generated by the uniform distribution over the interval [*a*,*b*], where $a \le y < z \le b$. If the bid price is greater than or equal to the asking price, the player sells *g* for a price equal to the randomly generated bid price. Otherwise, the player's earnings are determined by playing *g* against a procedurally rational opponent.

Proposition 4: A substantively rational player uses the following ask function in the adapted BDM procedure:

$$a(g) = \begin{cases} \$z \text{ if } v(g) = W \\ \$y \text{ if } v(g) = L \end{cases}.$$

Proof: If sincere play in the game is incentive compatible with the elicitation procedure, then standard arguments give the ask function. The only non-standard difficulty here is establishing that a substantively rational player would want to win winnable games. Suppose the player asks \$*a* < \$*z* for an unbalanced (winnable) game, then with positive probability the player sells the game for a price \$*s* between \$*a* and \$*z*. Since \$*z* > \$*s*, the player would have been better off to hold the game and play it to win. Given that the player did not sell the game it is always in his interest to win unbalanced games that he holds.

However, a boundedly rational player may not be willing or able to implement an optimal strategy in an unbalanced game. This performance uncertainty may cause a boundedly rational player to discount the earnings from a win.

IV. Experimental Design

The experiment consists of two treatments. In treatment one, the subjects played Nim games against a procedurally rational computer program, see table one. Subjects were always in the role of the first mover. The games range in complexity from rank 3 to rank 17 and have shortest play paths ranging from 1 to 5. When a subject won a game they earned \$0.60 and when a subject lost a game they earned \$0.10, that is, *z* = \$0.60 and *y* = \$0.10 (In the summer sessions, *z* = \$0.70 and *y* = \$0.00).

Table 1: Experimental Design

	Session	Subject Pool	Number of subjects	Game Sample	Number of games in treatment 1	Number of games in treatment 2	$z	$y
Summer Sessions	1	TAMU	12	1	10	50	0.70	0.00
	2	TAMU	13	1	10	50	0.70	0.00
Baseline Sessions	3	TAMU	20	2	27	54	0.60	0.10
	4	TAMU	20	2	27	54	0.60	0.10
Caltech Sessions	5	Caltech	7	2	27	54	0.60	0.10
	6	Caltech	11	2	27	54	0.60	0.10
Spring Sessions	7	TAMU	20	3	55	55	0.60	0.10
	8	TAMU	20	3	55	55	0.60	0.10

Figure 1 is a screen grab of the graphical user interface used in treatment one. Subjects play the game by clicking on the black circles to remove them and everything to the right. Instructions are found to the right of the game. Records are kept at the bottom of the screen. A dialog reminding the subjects of what just happened appears between the game grid and the record grid. Treatment two adds the BDM interface in the upper right corner of the screen. The record grid becomes more complicated in treatment two.

Between treatment one and two subjects answer a questionnaire. It insures that subjects understand how to use our adaptation of the BDM procedure, before they proceed into treatment two. Subjects progress at their own pace.

In treatment two, the subjects were endowed with Nim games that they could sell in the BDM procedure. Subjects then played any unsold games. In treatment 2, the games range in complexity from rank 3 to rank 17 and have shortest play paths ranging from 1 to 5[3].

A surprising discovery changed the design of the experiment. Treatment 1 was originally designed to give subjects experience with Nim before they were asked to price Nim games. However, when we examined

[3] See McKinney (2001) for a screen grab of treatment two as well as a complete list of the Nim games used in both treatments.

the subjects performance in unbalanced games we discovered that 8 of 25 subjects had perfect bounds on their performance in the summer session (1 and 2). They won all unbalanced games with rank up to $\hat{r}$ and lost all other games.

The *baseline sessions* (3 and 4) expand the number of games included in treatment 1 to allow us to use treatment 1 to measure the bound on their performance before subjects are asked to price the games. This avoids the selection bias from measuring the bound on games they may have selected through their pricing behavior.

At the suggestion of a discussant we investigated the robustness of our findings to subject pool differences. The baseline design was used with subjects recruited from the Caltech student body. We will refer to these as the *Caltech sessions* (5 and 6).

In the baseline and Caltech sessions, the games were selected through a block randomization process from the set of Nim games. The blocking variables were: balanced games, unbalanced games with even rank and unbalanced games with odd rank. (No odd ranked balanced games exist.) Within each block, the games were chosen so as to cover multiple combinations of r and m. The selected games were then ordered by rank, divided into groups of three and distributed between treatments one and two.

The randomization procedure used for the baseline and Caltech sessions resulted in only one unbalanced game with a shortest play path equal to 4 in treatment 1. So this class is under-represented in the baseline sample and may account for the differences observed between the summer sessions and the fall sessions. Half the games in the summer sessions had m = 4. Sessions 7 and 8 will be called the *spring sessions*. These sessions sampled games in such a way as to restrict the measurement of the influence of changing the length of the shortest play path to the crucial range of rank 6 to 10. Also, the same sample of games was used in treatment 1 and treatment 2. Hence, in the spring sessions subjects were always pricing games they had played before. The rows in the game were scrambled in order to make it difficult to recognize this similarity between treatment 1 and treatment 2 in the spring sessions.

The subjects for summer, fall, and spring sessions were recruited from the Texas A&M University (TAMU) student body and subjects for the Caltech sessions were recruited from the California Institute of Technology (Caltech) student body. A total of 123 subjects participated in the experiment. The experiment was conducted in the Economic Research Laboratory and the Hacker Social Science Experimental Laboratory respectively. A substantively rational subject's expected earnings are $40.11 in the baseline and Caltech sessions, $30.80 in the summer sessions, and $52.76 in the spring sessions. The length of the experiments varied across subjects. Some subjects finished in 30 minutes and others took up to 93 minutes.

V. Experimental Results

A. Estimating Bounded Rationality

Figures 2*a*, *b*, and *c* report the frequency subjects won the unbalanced games in treatment 1. The games in the figures are sorted first by blunderer win probability and then by NT-complexity. For comparison, the figures include the blunderer and substantively rational subject's winning probability.

Figure 2*a* reports the data for the summer subjects. Subject performance lies between blunderer and substantively rational play. As games increase in complexity performance moves away from substantively rational play and towards blunderer play. Notice that both the rank and blunderer measures treat (0,0,0,7), (1,0,1,5), and (1,1,1,4) as having equivalent complexity, but the win frequency is declining as predicted by NT-complexity.

Figure 2*b* reports the data for the baseline subjects and the Caltech subjects. Baseline subject performance is always statistically distinguishable from substantive rationality and is statistically indistinguishable from blunderer play by game (2,0,1,2). The baseline subject performance deteriorates more quickly than the Caltech performance. The Caltech performance is statistically indistinguishable from substantive rationality for the first three non-trivial games and does not become indistinguishable from blunderer play until game (1,4,2,1,2).

Figure 2*b* also reveals evidence of the use of Nim tricks. Both the baseline and Caltech subject performance spikes up at games (6,0,5,0,0) and (6,0,2,6,0). These games can be won by reducing them to strategy copying games. The Caltech performance spikes up more dramatically suggesting a wider recognition of the strategy copying trick amongst Caltech subjects.

Figure 2*c* reports the data from the spring subjects. It uses a different sample of games in treatment 1. The spring subject performance now has two non-trivial games that are statistically indistinguishable from substantive rationality: (2,0,1,0,0) and (1,1,2,0,0). In this sample of games, it is not until (2,0,2,3,0) that the spring subject performance becomes indistinguishable from the blunderer performance. Again, there is some evidence that subjects recognize Nim tricks. Game (5,5,0,4) can be won by reducing it to a game in which one can use the strategy copying trick and subject performance spikes up in a statistically distinguishable way from blunderer play for this game.

The blunderer measure orders subject performance more accurately than predictions based on the assumption of substantively rational play. However, almost all subjects out perform the blunderer model for sufficiently simple games.

In treatment one of the baseline sessions, 5 of the 40 subjects won all games up to rank $\hat{r}_i$ and lost all games more complex than $\hat{r}_i$. The bound on their performance was like a wall. Depending on how strict one is about the criteria most of the other baseline subjects appear to have a bound on their rationality in the interval between rank 4 and rank 10.

In order to measure these less than perfect bounds, we estimate a logit regression of the probability a subject wins against the rank of the game. The estimated rationality bound for subject *i*, $\hat{r}_i$, is either the perfect bound on *i*'s performance or if less than perfect then the point at which a logit regression for *i* estimates the subject as being equally likely to win or lose games of that rank.

Sixty percent of the baseline subjects were within 2 games of a perfect bound at $\hat{r}_i$ and all of the baseline subjects were within 5 classification errors of $\hat{r}_i$. Table 2*a* reports the number of classification errors, that is, the number of games the subjects won above their estimated bound plus the number of games they lost below their estimated bound.

The Caltech data has no subjects with a perfect bound. The Caltech sample also has more classification errors suggesting that the wall model does not fit their behavior as well as it did for the baseline subjects, see table 2*b*. One obvious reason is that they recognized more Nim tricks. A good Nim tricks player is not well modeled as having a wall based on rank.

The spring data has slightly more than twice as many games as in the baseline sample, and the spring subjects have about twice as many classification errors as the baseline subjects, see table 2*c*.

Table 2*a*: Classification Errors Given $\hat{r}_i$ Baseline Sessions

Classification Errors	Number of Subjects	Cumulative Count	Percentage	Cumulative Percentage
0	5	5	12.5%	12.5%
1	4	9	10.0%	22.5%
2	15	24	37.5%	60.0%
3	12	36	30.0%	90.0%
4	1	37	2.5%	92.5%
5	3	40	7.5%	100.0%

Table 2*b*: Classification Errors Given $\hat{r}_i$ Caltech Sessions

Classification Errors	Number of Subjects	Cumulative Count	Percentage	Cumulative Percentage
0	0	0	0.0%	0.0%
1	1	1	5.6%	5.6%
2	8	9	44.4%	50.0%
3	2	11	11.1%	61.1%
4	4	15	22.2%	83.3%
5	2	17	11.1%	94.4%
6	1	18	5.6%	100.0%

Table 2*c*: Classification Errors Given $\hat{r}_i$ Spring Sessions

Classification Errors	Number of Subjects	Cumulative Count	Percentage	Cumulative Percentage
0	0	0	0.0%	0.0%
1	0	0	0.0%	0.0%
2	1	1	2.5%	2.5%
3	5	6	12.5%	15.0%
4	5	11	12.5%	27.5%
5	4	15	10.0%	37.5%
6	13	28	32.5%	70.0%
7	3	31	7.5%	77.5%
8	2	33	5.0%	82.5%
9	5	38	12.5%	95.0%
10	1	39	2.5%	97.5%
13	1	40	2.5%	100.0%

Figures 3*a*, *b*, and *c* report histograms of $\hat{r}_i$. The median bound is 6 for the baseline and spring subjects and 8 for the Caltech subjects. All the subjects are contained in the interval [3,14].

We estimate a random effects logit model to characterize the unconditional probability that a subject of type h_i drawn from a normal distribution produces a vector of wins and losses, w_i, as a function of the rank[4], r_i, of the n_i games played by subject *i*. Let q_i index the unbalanced games played by subject *i*, then the model is as follows:

$$\Pr(w_i|r_i) = \int_{-\infty}^{\infty} \frac{e^{-h_i^2/2\sigma_h^2}}{\sqrt{2\pi\sigma_h}} \left[\prod_{q_i=1}^{n_i} F\left(r(q_i)\beta + h_i\right) \right] dh_i$$

where n_i is the number of games played by *i*, h_i is the random effect for *i*, and *F* is given as follows:

$$F\left(r(q_i)\right)\beta + h_i = \begin{cases} \dfrac{1}{1+e^{r(q_i)\beta+h_i}} & if \quad w(q_i) = W \\ 1-\dfrac{1}{1+e^{r(q_i)\beta+h_i}} & otherwise \end{cases}$$

[4] We use rank because it is the most intuitive of our measures of difficulty. Models using NT-Complexity or the Blunderer measure yield similar results.

Table 3 reports the estimated parameters for the rank model by subject pool. A Wald test fails to reject the hypothesis that the baseline and spring data are derived from the same population. The pooled estimate is reported in the row labeled TAMU. A Wald test does reject pooling the TAMU and Caltech data. (The Wald statistic is 41.06, which is significant at all conventional levels.) The main difference between the TAMU and Caltech samples appears in the constant term rather than the precision parameter. The representative TAMU subject has an estimated rationality bound of rank 6.27. The representative Caltech subject has an estimated rationality bound of rank 8.97.

Table 3: Logit Model of Winning Probability by Subject Pool

Subject Pool	Rank	Constant	σ_h	LogLikelihood
Overall	-0.5769 (0.03)	3.8764 (0.23)	1.0927 (0.11)	-1009
Baseline	-0.5816 (0.05)	3.5262 (0.38)	0.8233 (0.18)	-253
Spring	-0.5908 (0.03)	3.8070 (0.31)	1.0157 (0.15)	-601
TAMU	-0.5893 (0.03)	3.6949 (0.23)	0.9505 (0.11)	-855
Caltech	-0.5376 (0.06)	4.8218 (0.61)	0.8473 (0.24)	-141

Standard errors are reported in parentheses
σ_h represents the panel-level variance

B. Pricing Performance Uncertainty

A substantively rational player would set an asking price of \$0.60 for the unbalanced games and an asking price of \$0.10 for the balanced games in sessions 3 to 8. A straight test of this proposition reveals little of the predicted difference in asking price. In order to give the theory its best shot we distinguish between simple and complex games.

In what follows, we use the individual estimates, $\hat{r}_i$, reported in figures 3*a*, *b*, and *c* to sort games into *simple games* for subject *i*, which have a rank below the subject's estimated bound, and *complex games* for subject *i*, which have a rank above the subject's estimated bound. Notice that this sort is by subject performance in treatment 1. Because the summer session has so few treatment 1 games we are not able to estimate $\hat{r}_i$ accurately and, hence, we will not use this data when making a distinction between simple and complex games for a particular subject.

At least for simple games, subjects value balanced and unbalanced games differently. In the baseline sessions, the average price for the simple unbalanced games is $0.47 and for simple balanced games is $0.19, see table 4. The difference in asking price for simple games is $0.28, which is more than four times the difference in asking price for complex games of $0.06. The spring sessions differences are both within $0.02 of these averages essentially replicating the baseline findings.

The Caltech sessions come closer to the theoretical ask function, see table 4. The difference in asking price for simple games is $0.36, which is twice the difference in asking price for complex games of $0.15. The Caltech subjects are able to discriminate between balanced and unbalanced games more effectively.

Table 4: Average Asking Price

		Simple Nim Games			Complex Nim Games		
Sessions	Overall	Unbalanced	Balanced	Difference	Unbalanced	Balanced	Difference
Baseline	$0.28	$0.47	$0.19	$0.28	$0.27	$0.21	$0.06
Caltech	$0.36	$0.52	$0.16	$0.36	$0.35	$0.20	$0.15
Spring	$0.24	$0.40	$0.14	$0.26	$0.23	$0.18	$0.04
a(g)	$0.44	$0.60	$0.10	$0.50	$0.60	$0.10	$0.50

We estimate a random effects tobit model to characterize the asking price, *a(g)*, as a function of NT-Complexity and the games value *v(g)*. Table 5 reports the estimated parameters for a tobit model. LogNT represents the natural log of NT-complexity. BAL = 1 if *v(g)* = *L* and BAL = 0 when *v(g)* = *W*. NTBal is the product of LogNT and Bal, the interaction variable. σ_ε and σ_h represent the overall and panel-level variances.

Table 5: Tobit Asking Price Model by Subject Pool Samples

	LogNT	NTBal	Bal	Constant	σ_h	σ_ε	LogLikelihood
Overall	-1.4199 (0.06)	1.2938 (0.11)	-24.2699 (1.23)	41.0805 (0.69)	11.2546 (0.35)	19.4256 (0.19)	-25414
Summer	-1.2898 (0.16)	1.1236 (0.26)	-22.7462 (3.37)	50.1216 (2.06)	12.0159 (0.80)	22.5188 (0.55)	-4623
Baseline	-1.7072 (0.09)	1.4642 (0.19)	-23.1186 (2.13)	43.2371 (1.04)	11.2467 (0.47)	17.8105 (0.30)	-8353
Spring	-1.1603 (0.11)	1.1141 (0.19)	-20.2342 (1.98)	37.1535 (1.10)	11.1743 (0.54)	19.2981 (0.33)	-8449
Caltech	-1.5071 (0.13)	1.5358 (0.28)	-38.6305 (3.22)	56.6033 (2.32)	6.8266 (0.82)	18.1455 (0.44)	-3873

Standard errors are reported in parentheses
σ_h represents the panel-level variance
σ_ε represents the overall variance

All of the variables are statistically significant at conventional levels. The statistically significant negative coefficient on LogNT imply that for the unbalanced games the predicted asking price, $\hat{a}(g_i)$, declines as NT-complexity rises in all four samples. In the balanced games, the change in asking price is far less sensitive. For the balanced games, $\hat{a}(g_i)$ is downward sloping for all TAMU subject pools but has a slight positive slope for the Caltech subjects. The estimated values for the tobit models imply that as the games become more complex the subjects' ability to recognize unbalanced games diminishes, thus they set lower asking prices.

We reject all pooling hypotheses. Wald tests comparing the four subject pools against the overall model always yield statistics of at least 17.24, which is significant at all conventional levels of statistical significance. The difference in summer, baseline, and spring samples is surprising. We attribute it to the increasing amount of experience subjects had in treatment one. Summer sessions had 10 games, baseline had

27, and spring had 55. Spring subjects were much better calibrated and we conjecture it is because going into treatment 2 subjects had much more performance experience. Also, this performance experience was directly relevant as they were pricing the same games they had played in treatment one.

How accurate are subjects at valuing their own performance? Call a subject overconfident when the difference between a subject's average asking price and actual earnings is positive. All of the subjects in the summer sessions had a positive difference. All but 3 of 40 in the baseline sessions, all but 4 of 40 in the spring sessions, and all but 4 of 18 in the Caltech sessions had a positive difference. All four treatments reveal overconfidence: subjects price the games at more than they are likely to earn if called on to play the game by the BDM procedure.[5]

The mean overconfidence is significantly positive for all four treatments. Specifically, the mean overconfidence for the summer sessions was 26.7¢, for the baseline sessions was 12.7¢, and for the spring sessions was 9.8¢. Recall that the number of games in treatment 1 was 10, 27, and 55 for the summer, baseline, and spring sessions respectively. The mean overconfidence is decreasing in the amount of experience gained in treatment 1. The extra experience appears to have resulted in better calibrated pricing behavior and less overconfidence.

VI. Discussion

The subject pool differences detected in this experiment should not be over emphasized. While it was possible to find a subject pool that priced the games closer to their value under substantively rational play, the differences in average performance are small. The Caltech performance does not stochastically dominate the baseline or spring cohorts. For example, the two subjects with the highest estimated bound are both Aggies. The variations in estimated bounds distracts from the main lesson, which is that people's ability to

[5] See Camerer (1997) for a review of the overconfidence literature.

reason effectively in parlor games is bounded by a small number when compared to the complexity of ordinary parlor games.

Bounded rationality models might be compared to models of iterated reasoning. Nagel (1995) and Ho, Camerer and Weigelt (1995) found that subjects typically do 1 to 3 iterations of reasoning in p-beauty contest games. Stahl and Wilson (1995) define "level-*k*" players as players that best respond to players that perform *k-1* iterations of reasoning. They find mostly level-1 and level-2 players, implying 1 to 2 levels of iteration on the rationality of others. Our bounds are typically larger than their reported levels of iterative reasoning, because they are measuring a different thing. Our bounds measure the rank of a winnable game in which a player has an even chance of winning against a procedurally rational opponent.

Gabaix and Laibson (2000) also attempt to measure bounded rationality. They find some evidence for a wall and report an estimated value of 8, which is in the range of values we have observed across subject pools. Ultimately they reject the wall model in favor of a heuristic that ignores contingencies that have relatively small impact on expected payoffs. Their experimental design presented subjects with decision paths and it was relatively easy to follow specific paths through to a terminal outcome. In our experiment, subjects would have to construct the play paths in their head, which we don't observe, if they are using the Gabaix and Laibson heuristic. A single mistake in this "pruning" results in a loss against a procedurally rational opponent in Nim. Given the complexity of the extensive form representation of even simple Nim games, we are skeptical of subjects ability to imagine the extensive form.

Our estimated bounds are in most cases larger than the capacity limits for the focus of attention reported in Cowan (2001). Earlier estimates, such as Miller (1956), found that people can remember about 7 chunks in short-term memory, but more recent work has lowered this earlier estimate to 4. While our work does not measure the limits of attention and we made no effort to prevent "chunking", the bounds we discovered may ultimately be related to deeper limits on brain function.

REFERENCE

Gordon M. Becker, Morris H. DeGroot, and Jacob Marschak, "Measuring Utility by a Single Response Sequential Method." *Behavioral Science* 9, 226-32, 1964.

Ken Binmore, "Modeling Rational Players," *Economics and Philosophy* 3 1987.

Ken Binmore, *Fun and Games: a text on game theory*, D.C. Heath and Company, Lexington, 1992.

Colin F. Camerer, "Progress in Behavioral Game Theory," *The Journal of Economic Perspectives* 11,167-188, 1997.

Charles L. Bouton, "Nim A Game with a Complete Mathematical Theory," *The Annals of Mathematics* 1901–1902,3(1), 35-39.

J. H. Conway, *On Numbers and Games*, Academic Press, New York, 1976.

Nelson Cowan, "The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity," *Behavioral and Brain Sciences* 24(1) 2001.

Xavier Gabaix and David Laibson, "A Boundedly Rational Decision Algorithm," *American Economic Review: Papers and Proceedings*, 90(2) May 2000.

Teck-Hua Ho, Colin Camerer and Keith Weigelt, "Iterated Dominance and Iterated Best Response in Experimental p-Beauty Contests," *American Economic Review* 88(4) September 1995.

R. Duncan Luce and Howard Raiffa, *Games and Decisions: Introduction and Critical Survey*, John Wiley and Sons, 1957.

G.A. Miller, "The Magical Number Seven, plus or minus two: Some limits on Our Capacity for Processing Information," *Psychological Review* 104, 1956, 81-97.

C. Nicholas McKinney, "Essays on Bounded Rationality." Ph.D. dissertation, Texas A&M University, 2001.

Rosemarie Nagel, "Unraveling in Guessing Games: An Experimental Study," *American Economic Review* 85(5) December 1995.

Herbert A. Simon and Jonathan Schaeffer, "The Game of Chess," *Handbook of Game Theory, vol. 1*, edited by R.J. Aumann and S. Hart, Elsevier Science Publishers B.V., 1992, p. 1-16.

Reinhard Selten, "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory* 1975, 4(1), 25-55.

Dale Stahl and Paul Wilson, "On Players' Models of Other Players: Theory and Experimental Evidence." *Games and Economic Behavior* 10(1) July, 1995

E. Zermelo, "Uber eine Anwendungen der Mengenlehre auf die Theorie der Schachspiels", *Proceedings of the International Fifth Congress of Mathematicians, Cambridge 1912, vol. II.* Cambridge: Cambridge University Press, 1913, 501-504.

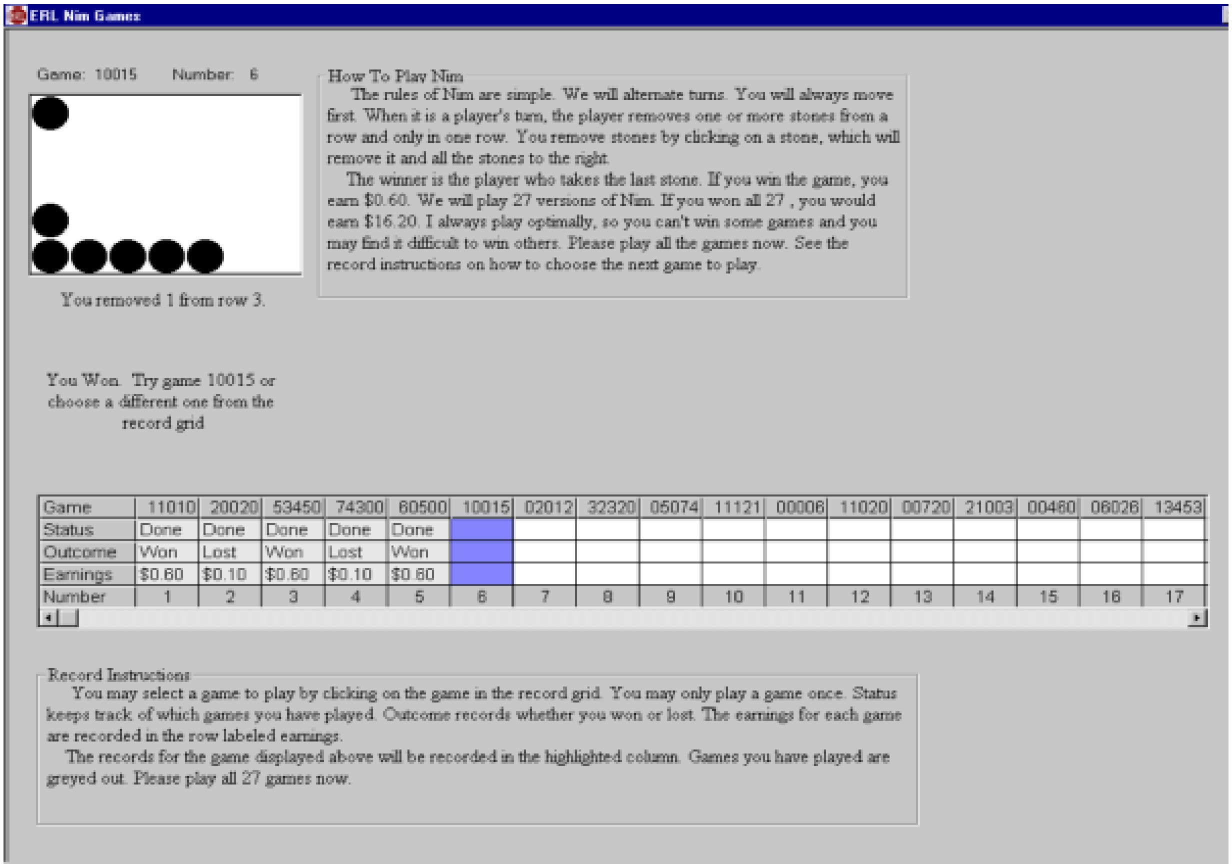

Figure 1: Treatment 1 user interface

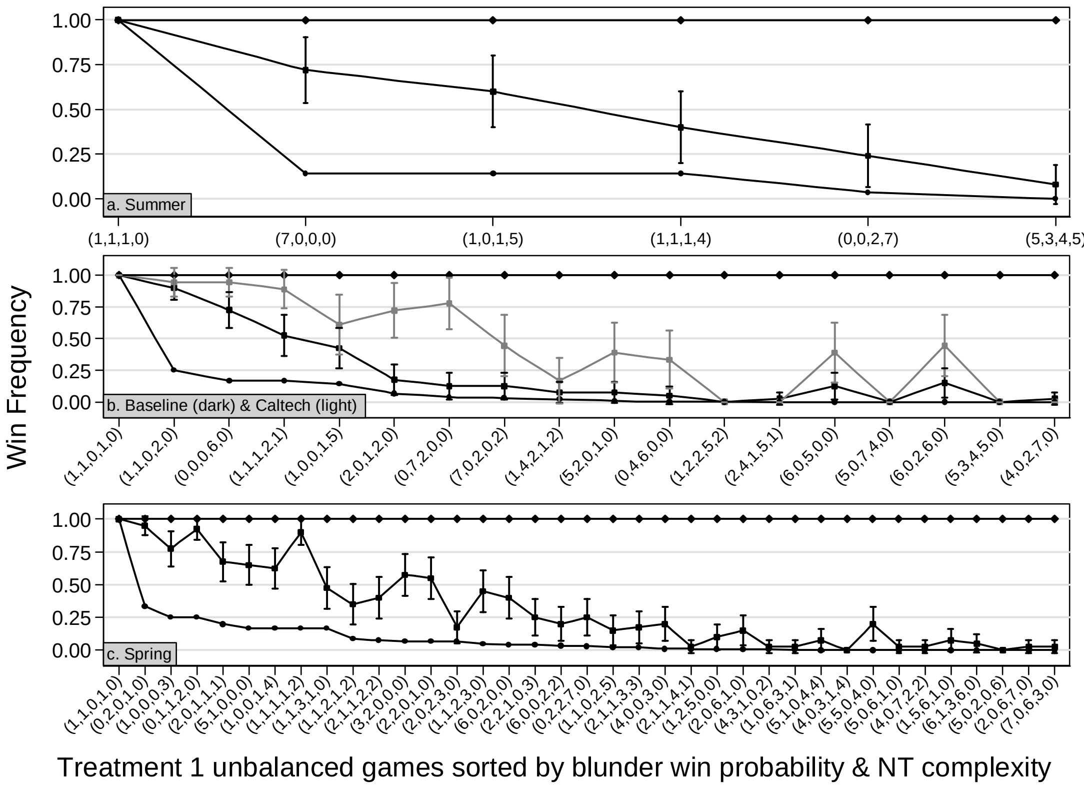

Figure 2: Subject win frequency (■) for unbalanced games. The games are sorted by blunderer win probability and NT complexity. The upper line (♦) represents the procedurally rational model and the lower (●) line represents the blunderer model.

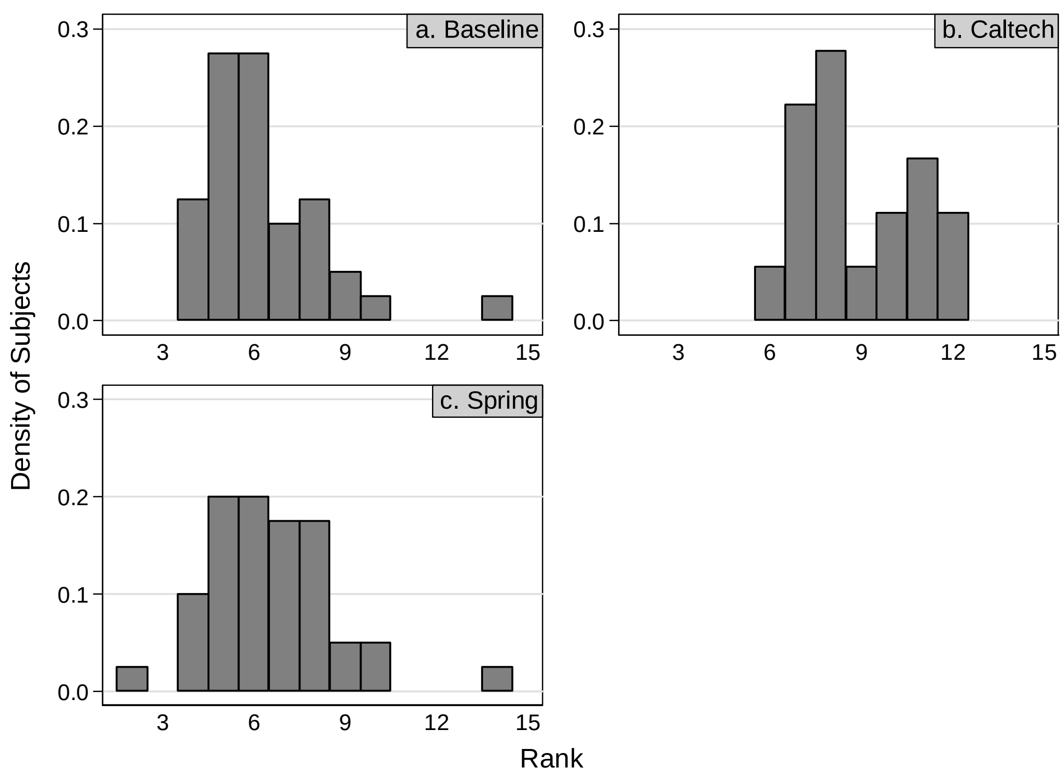

Figure 3: Histogram of $\widehat{r}_i$