"One, Two, (Three), Infinity, …: Newspaper and Lab Beauty-Contest Experiments"[1]

by

Antoni Bosch-Domènech, Jose García-Montalvo, Rosemarie Nagel, and Albert Satorra

Universitat Pompeu Fabra, Barcelona

January 31, 2002

Acknowledgments: We wish to thank Gary Charness, Brit Grosskopf, Ernan Haruvy and Joaquim Silvestre for their remarks on a preliminary version. Special thanks also for two referees who provided extremely helpful comments. We acknowledge financial support from the Spanish Ministry of Education through grants SEC98-1853-CE, SEC00-1034, BEC2000-0983 and PB98-1076, and the EU-TMR Research Network ENDEAR (FMRX-CT98-0238). We thank the Spanish newspaper *Expansión* and the German magazine *Spektrum der Wissenschaft* for letting us use their platforms to run our experiments, Richard Thaler for giving us his data from the *Financial Times* experiment, and Gary Charness, Sjaak Hurkens, Angel López, and Bettina Rockenbach for running an experiment in their classes.

[1] George Gamow (1988) wrote the popular book *One, Two, Three,…, Infinity: Facts and Speculations in Science*. Note that, in our title, Three is in brackets and the dots are not between Three and Infinity. The name of the game has been adapted from John Maynard Keynes (1936, p.156) well-known metaphor of Beauty-contest games played in the nineteen-twenties in some UK newspapers. See also at the end of this paper an extended quotation.

Abstract

A “Beauty-contest” is a game in which participants have to choose, typically, a number in [0,100], the winner being the person whose number is closest to a proportion of the average of all chosen numbers. We describe and analyze Beauty-contest experiments run in newspapers in UK, Spain, and Germany and find stable patterns of behavior across them, despite some uncontrollability of these experiments. These results are then compared with lab experiments involving undergraduates and game theorists as subjects. We claim that a large proportion of the observed behavior, across a wide variety of treatments and subject pools, can be interpreted as iterative best reply. Level-1 reasoning, Level-2 reasoning, and occasionally Level-3 reasoning, are commonly observed in all the samples, while the equilibrium choice (Level-∞ reasoning) is only prominently chosen when subjects have training or time to think. Analyzing the comments sent by the participants, we also identify further reasoning patterns. The results show the empirical power of experiments run with large subject-pools, and open the door for more experimental work performed on the rich platform offered by newspapers and magazines.

J.E.L. classification codes: C7, C9

Keywords: experiments, bounded rationality, parallelism, games, Keynes

1. Introduction

In June 1997, Richard Thaler and Bosch-Domènech and Nagel independently designed and announced an experiment on the Beauty-contest game in two different daily business newspapers (the *Financial Times* (FT) in the UK, and *Expansión* (E) in Spain), inviting the readers to participate. Five months later, Reinhard Selten and Nagel replicated the experiment in the monthly *Spektrum der Wissenschaft* (S), the German equivalent of *Scientific American*.

Experimenting with newspaper or magazine readers means losing control over some important elements. However, it opens up the possibility to experiment with 1) larger numbers of subjects, 2) larger rewards, 3) longer timescales, and 4) a more diverse subject pool than would be possible in the lab. Also, experiments in newspapers can be inexpensive, since sponsors may be induced to finance prizes. And potentially, they have a huge educational impact on the public at large, being advertised, described, and analyzed in the mass media.

Most important, running experiments in a newspaper helps to answer the following question. Are the results of lab experiments different from those obtained with large numbers of subjects, who are not the usual students, have plenty of time to ponder their decisions, and can obtain large prizes? To say it differently, by running experiments in newspapers we put to test the critical assumption of "parallelism" between the lab and the field.

A laboratory experiment usually consists of a relatively small group of persons (up to 20 subjects), who arrive at the lab at the same time, participate in an experiment for 1 or 2 hours, and are paid slightly above the minimum wage. A number of experiments tried to go beyond this basic procedure. Peter Bohm (1972) pioneered public good experiments carried out by the Swedish Radio TV Broadcasting Company, with hundreds of subjects. The Iowa Presidential Stock Market (Robert Forsythe, Forrest Nelson, George R. Neumann and Jack Wright, 1992) engaged the University of Iowa community to test how well markets work as aggregators of information. Mark Isaac, James Walker and Arlington Williams (1994) ran repeated public good games with 40 or 100 subjects over several days. More recently, the advent of Internet has allowed some experimenters to move out of the lab. Peter Bossaerts and Charles Plott (1999), for instance, have run market experiments using the Internet as a medium to collect experimental data, subjects being able to log in any time they want within a range of several days. Other examples are David Lucking-Reiley (1999) and John List and Lucking-Reiley (2000), who tested different auction mechanisms selling sports-cards on the Internet or in a real market. On a different track, a large number of field studies of social programs involving thousands of participants, the so-called social experiments, have been performed in the last decades.[2] See also Randall Bennett and Kent Hickman (1993) and Jonathan Berk, Eric Hughson and Kirk Vandezande (1996), who used data from the show "The Price is Right" in order to study rational behavior.

Some researchers even ran experiments in magazines. In 1993, inspired by Robert Axelrod (1984) and the column "Metamagical Games" by Douglas Hofstadter (1983a, 1983b)) in *Scientific American*, Jean-Paul Delahaye and Philippe Mathieu (1993, 1996) asked the readers of *Pour la science* to send in programs for an iterated prisoners dilemma experiment with the possibility to exit the PD-game. Ninety-five readers responded with interesting comments and programs.

The present paper analyzes a rich data set on the Beauty-contest game. We first describe and compare the three Newspaper experiments. We then relate these experiments to similar ones run in labs (as reported in Nagel (1995)), and to new experimental data collected in classrooms, conferences, by e-mail or through newsgroups. These non-laboratory sessions may allow more time to students, or use economists, game theorists, or the general public as subjects. Finally, we integrate the analysis of these independent experiments -involving different subject pools, sample sizes, payoffs, and settings- in a single statistical model. This analysis enables us to capture aspects of the individual decisions that fit the predictions of the iterated best reply model, as discussed by Teck Ho, Colin Camerer, and Keith Weigelt (1998), Nagel (1995, 1998), and Dale Stahl (1996). We also identify other reasoning patterns that were absent in the analyses of these papers and quantify their frequencies with the help of the many comments received from the participants of the Newspaper experiments.

[2] David Greenberg and Mark Schroder (1997) report 143 social experiments completed by the end of 1996 and 74 additional

2. The game and reasoning processes

In a basic Beauty-contest game, each player simultaneously chooses a decimal number in the interval [0,100]. The winner is the person whose number is closest to *p times the mean of all chosen numbers*, where $p < 1$ is a predetermined and known number.[3] The winner gains a fixed prize. If there is a tie, the prize is split amongst those who tie or a random draw decides the winner.[4] In this game there exists only one Nash equilibrium in which all choose zero, or the lowest possible number.[5]

We analyze the data and comment sets according to the following five types of reasoning processes. The first two are related to the game theoretic analysis; types three and four have been introduced and discussed in the previous literature on the Beauty-contest game.

1. The lowest number of the interval is the unique equilibrium. Anybody who deviates unilaterally from it will deviate from the winning number, i. e., from p times the mean. This is the typical *fixed-point argument*.

2. The game is dominance solvable. The process of *iterated elimination of weakly dominated strategies* (which will be called ID) leads to the game's unique equilibrium in which everybody chooses 0.[6] Thus, a rational player does not choose numbers above $100p$, which are weakly dominated by $100p$. Moreover, if he believes that the other participants are also rational, he will not choose above $100p^2$ and so on, until all numbers are eliminated but zero. The concept of iterated dominance is an important concept in game theory. The Beauty-contest game is an ideal tool to study whether individuals reason in steps and how many iterated levels they actually apply.

3. For the Beauty-contest experiments, Ho et al. (1998), Nagel (1995), and Stahl (1996) show that a model of iterated best reply (IBR), starting at a uniform prior over other players' choices,

---

experiments not yet completed by that time. See also Greenberg, Schroder and Matthew Onstott (1999).

[3] Here we will only discuss the cases $0 < p < 1$. For the other cases see e.g. Nagel (1995).

[4] See Nagel (1998) for a survey on the Beauty-contest experiments.

[5] If only integers are allowed (as in F) there are several equilibria; in the case of $p = 2/3$, in addition to the equilibrium "all choosing 0," there is an equilibrium "all choosing 1." This is a minor modification that does not change the game in an important way. However, if p had been equal to 0.9, the equilibria would have been "all choosing either 0, 1, 2, 3, or 4," instead of just a unique equilibrium as in the case of real number choices (see Rafael López (2000)).

describes subjects' behavior better than the model of iterated elimination of dominated strategies. These authors classify a subject according to the number of levels of his reasoning and assume that, at each level, every player has the (degenerate[7]) belief that he is one level of reasoning deeper than the rest.[8] Therefore, a Level-0 player chooses randomly in the given interval [0, 100], with the mean being 50. A Level-1 player gives best reply to the belief that everybody is Level-0 player and thus chooses 50p. A Level-2 player chooses $50p^2$, a Level-k player chooses $50p^k$, and so on. A player, who takes infinite levels and believes that all players take infinite levels, chooses zero, the equilibrium. This hypothesis of iterated best reply together with p=2/3 predicts that choices will be on the values 33.33, 22.22, 14.81, 9.88, … and, in the limit, 0. This kind of process will be called IBRd where "d" stands for degenerate beliefs[9]. Note that the main difference between the iterated best reply model and the iterated dominance reasoning lies in the different starting point (50 vs.100).

4. Stahl (1998) tests whether a model of non-degenerate beliefs explains the data. We denote by IBRnd ("nd" stands for non-degenerate), the *iterated best reply to the non-degenerate beliefs* that other players are at more than one level of reasoning.

5. Lastly, we add a type of procedure that has not been mentioned in the previous literature. Players might realize that through "armchair" reasoning the "right" number could not be found. From comments submitted by participants in the E and S experiments we learnt that some of them avoided this problem by running their own experiment with a sample of people. We will call these subjects *experimenters*.

3. Newspaper Experiments

3.1 Design

---

[6] The number of iterations is infinite. When subjects choose in [1,100] (as in E), a finite number of reasoning steps leads to the equilibrium.

[7] In general, by degenerate we mean that the player assigns probability one to all the other players being at one specific level of reasoning. We say that a player has non-degenerate beliefs if he gives positive probabilities to the other players being at more than one level of reasoning.

[8] Ho, Camerer and Weigelt (1998) state that "while this is logically impossible, it is consistent with a large body of psychological evidence showing widespread overconfidence about relative ability (see e.g., Camerer and Dan Lovallo, 1996)."

Participants in the three Newspaper experiments (and in all other experiments discussed in the paper) are asked to choose a decimal number in [0,100],[10] the winner being the person whose number is closest to 2/3 of the average number submitted. Rewards offered to the winners and time available in the Newspaper experiments were much larger than those in the lab[11]. Table 1 summarizes common aspects and differences between the three Newspaper experiments.[12]

Thaler (1997a) and Bosch-Domènech and Nagel (1997a) wrote the instructions independently of each other. Selten and Nagel (1997) had both sets of instructions when writing for S. The newspapers' editors induced some of the differences in the instructions. Thaler had to limit the choices to integers instead of decimal numbers. The reason was a legal restriction imposed by the FT attorney, who felt that a game with decimal numbers becomes a game of pure chance. Gambling by private persons or institutions is not allowed in the UK. This restriction causes a higher number of ties. In order to decide the winner in FT's contest, "the judges consider the best answer to be the tie breaker."

Only in the FT experiment were entrants obliged to explain their decisions. Many experimentalists are concerned that requiring explanations from subjects may force them to think their decisions over, bringing about more thoughtful results.[13] Indeed, in S, the average choice of entrants submitting comments (24% of all entrants) was 14.4, while the average of those without comments was 26.8. In FT all entrants were supposed to submit comments and their average was 18.9. However, the average choice of those in E with comments (4,5% of all entrants) was 25.2, whereas without comments it was 25.5.

Similarly, providing examples in the instructions may affect decisions. In FT, Thaler used an example (with number 20 as a winner) in order to prevent choices above 50. Indeed, in FT, numbers above 50 were less frequent than in the other two publications: 4% in FT, 9% in E and 10% in S.

---

[9] We use "beliefs" as synonym of "beliefs about the choices of others".
[10] As mentioned, in E the choice was in [1,100], and in FT it was restricted to non-negative integers.
[11] All data sets used in this paper are available upon request from the authors.
[12] Many of the methodology aspects mentioned here also hold for Internet experiments or experiments done for third parties such as a government or firms.
[13] About the effect on decisions of prompting subjects to think more carefully, see for instance Rachel Croson (1999).

E requested that the opening article include a reasoned justification for performing the experiment. This newspaper did also several pre-announcements of the game, days before the opening article appeared. This probably caused a higher number of participants than in the other newspaper experiments. Furthermore, without the authors' knowledge, E published a shortened version of the opening article containing the rules of the game on the three consecutive days following its publication. The shortening resulted in the omission that comments were welcome and, consequently, we received fewer comments from E than from the other newspapers. It also omitted mentioning that only one number per person would be accepted. In fact, several participants submitted multiple numbers. However, they only amounted to about 1% of the entries.

Insert Table 1 about here

3.2. Results

3.2.A. choices

In part A of this section we analyze and compare the data sets of choices from the three Newspaper experiments. In part B we make use of the large number of comment received for these experiments.

Figures 1a-c show the relative frequencies of the chosen numbers (in intervals [0, 0.5); [0.5, 1.0); etc.), the average choice, the winning number and the number of participants in the three Newspaper experiments. The figures indicate the similarity of choices despite the differences in subject pools and notwithstanding the uncontrollability of such experiments. In addition, the results confirm the existence of a common pattern of decision-making, previously identified in the lab experiments of the Beauty-contest game as levels of iterated best reply (IBRd, see Section 2). We report these findings as:

---

[15] The spike at 33.5 in Figure 1 results from the choice in E being constrained to the interval [1, 100], so that 2/3 of the average is 33.66. The rounding up of this and other numbers from 33.5 to 34.00 in the figure yields 33.5. The interval constraint in E and the restriction to integers in FT also causes the spike at 1.

Fact 1: The three Newspaper experiments result in similar frequency distributions. In particular, they all show spikes at number choices 33.33, 22.22, and 0.[15]

In line with previous work, we take spikes 33.33 and 22.22 as an indication that a number of participants follow Level-1 and Level-2 based on the IBRd-model.[16] The process of infinite iterated dominance or the fixed point argument can also explain the spike at 0. Models that incorporate non-degenerate beliefs do not offer plausible explanations of these spikes. Indeed, we find that none of the 72 participants in E and S[17] whose comments indicated a reasoning process according to IBRnd chose 33.33, 22.22 or 0.

Insert Figures 1a-c about here.

3.2.B. Comments

Here we analyze the set of comments[18] received from the participants of the Newspapers experiments in order to gain insight into the reasoning process behind their choices.[19] A detailed classification of the comments[20] according to the five types of reasoning processes mentioned in Section 2 results in the following observations (see also Table 2). From the 786 comments in E and S[21], 55% used iterated best reply degenerate (IBRd), of which 12 percentage points correspond to Level-0 (random choice); 14% used iterated dominance (ID); 9% iterated best reply non-degenerate (IBRnd); 5% ran their own experiment; 3% used a fixed point argument; and 15% described the equilibrium without explicitly

[16] Level 3 is less compelling.

[17] We do not have the comments submitted to FT.

[18] All comments used in this paper are available upon request from the authors.

[19] I do not understand In experimental economics exExaminations of reasoning processes or verbal comments thatwhich are not merely choices or outcomes, has not been done very often in experimental economics. A short overview of some exceptions can be found in Nagel (1998).

[20] To interpret comments presents significant difficulties, which might result in different classifications by different examiners. Therefore, two of us independently classified the set of comments from E readers according to the types of reasoning mentioned in Section 2. We then compared both classifications and settled any differences. After this, we classified the remaining comments.

[21] In E we received 166 comments. In S it was made clear that comments were welcome, and we received 645. Of these, we exclude 29 comments, which did not fit in any of the types mentioned in Section 2.

detailing their reasoning.[22] This last group may include fixed point reasoning, as well as IBRd Level-∞ and ID.

Insert Table 2 about here

If we disregard the 15% of equilibrium comments that cannot be classified, we can state the following fact:

Fact 2: A majority (64%) of comments show subjects using an IBRd argument, of which 15 percentage points correspond to Level-0 (random choice).

It is interesting to note that almost all subjects who provided comments describing IBRd only mentioned Levels 0, 1, 2, 3 and Level-∞. Even comments based on non-degenerate beliefs assign positive probabilities mainly to those levels.

Insert Figure 2a-c about here

In order to visualize the connection between types of comments and choices, Figures 2a-c plot the relative frequencies of choices (in intervals [0, 0.5); [0.5, 1.0); etc.) made by the subjects who submitted comments to E and S. Figure 2a represents the distributions of choices of those subjects who identify the equilibrium in their comments. We separate these subjects in three groups according to whether they describe their reasoning processes as ID Level-∞, IBRdLevel-∞, or fixed-point. The choices of those subjects who do not explicitly state their reasoning are pooled together with those in the last group. Figure 2b plots the distributions of choices of the subjects who *do not* reason all the way to the equilibrium. These subjects are again separated into three groups, according to whether their reasoning fits ID, IBRd (without Level-0) and IBRd Level-0. Figure 2c represents the choice distributions of the experimenters and of those subjects who apply IBRnd.

Comments describing IBRd Level-0 are associated with the highest dispersion of choices (Figure 2b). Comments describing IBRd Level-1, Level-2 or Level-3 are associated with large spikes at 33.33, 22.22 and near 14.81 (Figure 2b). More precisely, of all subjects describing these three

[22] See Appendix A for an example for each of these reasoning processes.

levels, 42% choose exactly 33.33, 22.22, 14.81. Choices with ID-comments (excluding Level-∞) show some concentration near the theoretical values 66.6, 44.4, 29.6, 19.75 (Figure 2b). In contrast to this, the choice distributions of experimenters and of those following IBRnd show no systematic features (see Figure 2c).

Finally, the three choice distributions of the group of subjects who identify the equilibrium are very similar and all have a large spike near 0/1 (see Figure 2a). Analyzing these 422 choices in this group, we find the following:

Fact 3: The large majority (81%) of subjects describing the Nash equilibrium choose a larger number than the equilibrium.

Some economists (see Plott, 1993) have argued that phenomena that appear irrational could be the result of rational players expecting others to behave irrationally. Fact 3 is an example of this phenomenon. This explains the three dots in the title after "infinity".

Turning the previous statement upside down, those who choose the equilibrium (19%), and thus appear rational, incorrectly expect that the other players will behave rationally. In psychology this is known as "false consensus" (see Ross, Greene, and House, 1977), meaning that a player assumes that other players reason as himself.[24]

4. Comparisons with lab experiments

[24] But Robin Dawes (1990) argues that expecting others to behave like oneself may not be that irrational after all.

As mentioned, one purpose of running experiments out of the lab is to help critically assess the assumption of "parallelism". Do we see, then, similarities or differences between Beauty-contest experiments run in labs and in newspapers?

Before entering into a detailed comparison, it is worth mentioning some of the basic differences between the two types of experiments, often due to the increased loss of control in newspaper experiments:

a) *Subjects' socio-demographic profiles:* Experimentalists know that their lab subjects are not representative of the population at large. They are aware, however, of some of their basic socio-demographic characteristics (age, sex, training…). In a newspaper experiment, we obtain a larger, but also uncertain, range of socio-demographic profiles.

b) Information seeking: Subjects of newspaper experiments may go to great lengths to submit informed answers. One interesting variety of observed information-seeking behavior consists in running a parallel experiment. Thirty-nine participants in S, and one in E, reported that they had run an experiment among students, friends, and relatives, to help them decide what number to submit. Of those, 31% chose a number between 12 and 17 (see also Figure 2c), the smallest integer interval containing all 2/3 of the averages in the three Newspaper experiments.[25] By contrast, among the entire population of all Newspaper experiments, only 11% chose in this interval (see Figure 3.6).

In one case, a participant in the S experiment decided to run his own replication of the experiment on an Internet newsgroup, with responses sent via e-mail (for the distribution of choices, see Figure 3.5). The winning number in his experiment was 14.81. He submitted 14.2 and was very close to winning the S prize, the winning number being 14.7. This is a difference of 0.1 points between one experiment with 150 subjects and another with 2,728![26] We state these results as follows:

---

[25] A group of German experimental physicists reported (see Selten and Nagel, 1998, p. 17): "We have concluded that we do not have any reasonable reference point. Therefore we decide to indulge the Deities of empiricism by running the game quickly with about 50 friends." Their choice was 15.768361, very close to the winning number.

[26] As noted by a referee, it is striking how close the "experimenter" came to the correct answer. Take, for instance, the 95% confidence interval (CI) for the winning number derived from the "experimenter" data (assuming the same sample size of S), which is found to be [14.24, 15.30]. This interval contains the winning number in S. Moreover, it is very similar to the 95% CI for the winning number in the S experiment, which is [14.15, 15.27].

Fact 4. Those subjects who conducted their own experiments in order to decide which number to choose were, on average, closer to the winning answer than theorists and the general public.

Another reader of S discussed the experiment in her math class and then submitted the joint bid of her classmates. Her account appears in Appendix A, and exemplifies the wide variety of comments received ranging from choosing a favorite number, to a finite IBR process, or choosing according to an experiment run in class. Her account is also a description of group decision-making reaching the equilibrium by infinite IBRd, and finally choosing close to equilibrium. But, as reported in the survey by Norbert Kerr, Robert MacCoun, and Geoffrey Kramer (1996), there is no clear evidence that groups make fewer judgmental errors than individuals.[27]

c) *Coalition formation:* In the lab the experimenter can easily avoid the formation of coalitions, but this is not possible in a newspaper experiment. In fact, we know that in all of these there were attempts of coalition formation,[28] although with little impact on the results (except for a larger than expected frequency at 100).

In the remainder of this section we present and compare the main features of 17 different experiments, collected from different sources. These experiments are pooled in eight groups described in Table 3.

Insert Table 3 about here. Insert Figure 3 about here.

To compare the results of the Beauty-contest experiments we plotted in Figure 3 the relative frequencies of choices of the six groups of experiments, separately. The first group, Lab-experiments with undergraduates, is clearly distinguished from the rest, because the Nash equilibrium was only

[27] Yet, Gary Bornstein and Ilan Yaniv (1997) report more rational behavior in group decision-making in ultimatum games. Similarly, Alan S. Blinder and John Morgan (2000) show that group decisions are on average superior to individual decisions. However, in a recent paper on group decisions in the Beauty-contest game, Martin Kocher and Matthias Sutter (2000) found that, in the first period, 15 groups (with 3 members in each and 5 minutes discussion time) do not choose differently than 15 single players. Differences between group decision-making groups and individuals will probably depend on the particular decision tasks and on the decision rules applied in the groups.

[28] The attempt was blatant in E. By allowing for the use of e-mail to submit numbers, we made it easy for a ring leader to spread the word among his e-friends to enter the number 100, so that he could increase his chance of winning by choosing a large number. Thaler (1997) reports that a "group from a College in Oxford all gave the answer 99…" Removing all 99 and 100 entries "the winning number would have been 12 instead of 13." In S the authors report that "the grandparents and parents Kennel […] send 100 [...] in order to irritate seemingly rational players who choose near 0 […] and in order to increase the winning chances of [their] daughter", who chose 5.5.

once (1%) selected. As soon as subjects have some training in game theory, the proportion of subjects choosing the equilibrium increases. The highest frequencies are attained when experimenting with theorists (Group 4, 15%), in which case, the greater confidence that others will reach similar conclusions may be reinforcing the effect of training. In Newspapers, the frequency of equilibrium choices (6%) falls somewhere in between, as should be expected from the heterogeneous level of training of their readers. We can, consequently, state the following observation:

Fact 5: Training, and playing with other trained subjects, seems to increase the frequency of choices near the equilibrium.[29]

Other than training, time availability may be a factor in the frequency differences observed in choosing the Nash equilibrium. To test this hypothesis, we ran two Classroom experiments (Group 2) and two Take-home experiments (Group 3) at the Universitat Pompeu Fabra, among undergraduate students with very limited knowledge of game theory, giving them about 5 minutes and one week, respectively, to return their number. These experiments show a small increase in equilibrium choices (these being 3% and 4%, respectively) with respect to Lab-experiments (1%), but almost no difference between them. However, analyzing the comments we find that only 9% indicate the equilibrium in Group 2 vs. 20% in Group 3. A similar comparison can be done with E and S equilibrium choices (3% in E choose 1, and 4% in S choose 0) and equilibrium comments (33% in E vs. 60% in S) with one and two week deadlines, respectively.[30] Time, therefore, helps subjects in identifying the equilibrium. But our particular game also allows subjects to find reasons not to stick to it.

We can state this result as:

[29] In Bosch-Domenech, García-Montalvo, Nagel and Satorra (2001), we construct a mixture distribution model and estimate the proportion of the different composing distributions. A t-test for equality of the proportions at Level-∞ in Classroom and Theorist experiments gives the value of t = -5.40 which is significant at any typical significance level; the same t-test for Classroom and Newspaper experiments, gives a value of t = -3.19 which is also highly significant. Even more significant would be the differences between Lab experiments and Theorists or Newspapers.

[30] Weber (2000) ran 10 period Beauty-contest games, in which no information was reported to the players until the end of the experiment. In spite of this, choices converged (albeit slowly) to equilibrium. This result is interpreted as implying that the choice in a game is affected by repeatedly thinking about it. More time allows more repeated thinking.

Fact 6: Time availability seems to increase the frequency of equilibrium comments, but not of equilibrium choices.

Time is also associated with the appearance of comments indicating that subjects follow IBRnd. This thinking process is absent in comments from experiments in Groups 1 and 2, and in the experiments of Group 4 with little time to think. However, in Group 3 and in the Bonn-experiment (Group 4) it is 10% and 5%, respectively.

Most important, all experiments show, in spite of these differences, a common pattern of choices already described as Fact 1 in relation to the Newspaper experiments. Fact 2 is also confirmed by the comments submitted. These comments show similar percentages of IBRd.[31] Excluding Level-0 reasoning (random choice) from it, we observe 49% for Group 1, 44% for Group 2, 46% for Group 3, and 46% for Group 4, just above the 43% observed in the Newspaper experiments. We can restate these facts as follows:

Fact 1b. All experiments analyzed result in frequency spikes at number choices 33.33 and 22.22 and also, in all but the Lab experiments, at equilibrium. Furthermore, in all experiments the modal reasoning process described in the comments is IBRd.[32]

7. Conclusions

Experimental results are influenced by what Jacob Marshak (1968) called the different costs of thinking, calculating, deciding and acting. Large-scale experiments of the sort that can be run through a newspaper can test whether the lab experiment results are robust to variations in sample sizes, rewards, and in the different costs mentioned by Marshak. In a newspaper experiment, one is likely to encounter a population more heterogeneous than undergraduate subjects. There may be subjects with widely different costs of thought and calculation (due to different education, training or information),

[31] For a complete classification of the comments from the lab experiments see Nagel (1993).

[32] In Bosch-Domenech et al (2001) we estimate from the experimental data the means and variances (except one) of a mixture distribution model. We show that across all very disparate experiments, the estimated means of the component

different decision costs (at leisure vs. the time-constraint imposed in the lab), and different costs of taking action (ready access to e-mail and fax or not). This is a richer world with less experimental control.

The fact that three experiments involving thousands of subjects, run in different countries, for different newspapers, catering to different populations yield very similar results is a clear indication that we are observing a pattern of behavior that must be quite common. In addition, this pattern is replicated in lab experiments with subject pools of undergraduate, graduate students and economists. This clearly indicates that the "parallelism" assumption between lab and field has been upheld.

The paper also shows that:

First, *iterated best reply (degenerate) is prevalent* across subject pools, different sample sizes, and diverse elicitation methods. In particular, Level-1, Level-2, Level-∞, and perhaps Level-3, are steps of IBRd reasoning frequently observed. In much lesser proportions, IBRnd and ID are also present, especially among trained subjects and with ample time available.

Nevertheless, the proportions of subjects employing different levels of reasoning may vary across experiments. They may depend on the following factors, among others: 1) Subjects' training, as in Lab experiments vs. Theorists; 2) Time availability, as in Classroom experiments vs. Take-home experiments; and 3) Information gathering efforts, as in Newspaper experiments. Also, for a number of participants following infinite IBR or ID, their choice ultimately depended on their confidence in others' ability to reach similar results.

Second, *newspaper experiments can be done and are fruitful*. Some economists may be skeptical about the future of newspaper experiments. We are not.[33] As readers become familiar with the Web pages of newspapers and magazines, experimenters can run Internet-like experiments from these Web pages.[34] This will provide experimenters with access to large and heterogeneous

---

distributions remain similar and close to the theoretical values predicted by IBRD.

[33] A recent example of a newspaper experiment is Werner Gueth, Carsten Schmidt, and Matthias Sutter (2001). They ran an ultimatum game using the platform of the *Berliner Zeitung*.

[34] Ernst Fehr and Suzann-Viola Renninger (2000) discuss the results of a Beauty-contest experiment announced in *DIE ZEIT* in which the participants were explicitly invited to debate about the game in the web site of the newspaper. About 100

populations, to sponsorship, and to a unique platform for publicizing the experimental methodology and divulging economics principles.

Before ending this paper, let us read again the famous quote by Keynes:

> […] professional investment may be likened to those newspaper competitions in which the competitors have to pick out the six prettiest faces from a hundred photographs, the prize being awarded to the competitor whose choice most nearly corresponds to the average preferences of the competitors as a whole; so that each competitor has to pick not those faces which he himself finds prettiest, but those which he thinks likeliest to catch the fancy of the other competitors, all of whom are looking at the problem from the same point of view. It is not a case of choosing those which, to the best of one's judgment, are really the prettiest, nor even those which average opinion genuinely thinks the prettiest. We have reached the third degree where we devote our intelligences to anticipating what average opinion expects the average opinion to be. And there are some, I believe, who practise the fourth, fifth and higher degrees." [Keynes (1936), p.156)].

Professional investors use their experience to ascertain what average opinion expects the average opinion to be… What should investors new to the trade do? Run experiments. This lesson we learnt from smart contestants in the Newspaper experiments.

participants used this forum. The authors report no difference with our results, except for a larger number of 100s "possibly encouraged by one participant via the Internet".

References

Adams, Douglas. "The Hitchhiker's Guide to the Galaxy." Ballantine Books, New York, 1995.

Axelrod, Robert. *The Evolution of Cooperation*. Basic Books, New York, 1984.

Berk, Jonathan B; Hughson, Eric; Vandezande, Kirk. "The Price is Right, but are the Bids? An Investigation of Rational Decision Theory." *American Economic Review*, 1996, 86 (4), pp. 954-70.

Bennett, Randall W; Hickman, Kent A. "Rationality and the 'Price is Right'." *Journal of Economic Behavior & Organization*, 1993, 21 (1), pp. 99-105.

Blinder, Alan S. and John Morgan. “Are Two Heads Better than One? An Experimental Analysis of Group vs. Individual Decision Making." *NBER WP*, No. 7909, September 2000.

Bohm, Peter. “Estimating Demand for Public Goods: An Experiment." *European Economic Review*, 1972, 3, pp. 111-30.

Bornstein, Gary, and Ilan Yaniv. "Individual and Group Behavior in the Ultimatum Game: Are Groups more “Rational” Players?" *Experimental Economics*, 1998, 1, pp. 101-108.

Bosch-Domènech, Antoni, José García-Montalvo, Rosemarie Nagel and Albert Satorra. “One, Two, Three, Infinity, …: Newspaper and Lab Beauty-Contest Experiments.” Mimeo, Universitat Pompeu Fabra, November 2001.

Bosch-Domènech, Antoni and Rosemarie Nagel. “Cómo se le da la Bolsa.” *Expansión,* June 4, 1997a, p. 40.

Bosch-Domènech, Antoni and Rosemarie Nagel. “El juego de adivinar el número X: una explicación y la proclamación del vencedor." *Expansión*, June 16, 1997b, pp. 42-43.

Bosch-Domènech, Antoni and Rosemarie Nagel. “Guess the Number: Comparing the Financial Times and Expansion’s Results." *Financial Times*, section Mastering Finance 8, June 30 1997c, p. 14.

Bossaerts, Peter and Charles R. Plott. “Basic Principles of Asset Pricing Theory: Evidence from Large Scale Experiments.” California Institute of Technology, July 1999.

Camerer, Colin and Lovallo, Dan. "Overconfidence and Excess Entry: An Experimental Approach." *American EconomicReview*, March 1999, 89 (1), pp. 306-18.

Cooper, David, John Kagel, Wei Lo, and Qing Liang Gu. “Gaming against Managers in Incentive Systems: Experimental Results with Chinese Students and Chinese Managers.” *American Economic*

Review, September 1999, Vol. 89 (4). Pp. 781-804..

Croson, Rachel. “The Disjunction Effect and Reason-Based Choices in Games." *Organizational Behavior and Human Decisions*, 1999, 80(2). pp. 118-133.

Dawes, Robin M. “The Potential Nonfalsity of the False Consensus Effect,” in *Insights in Decision-Making: A Tribute to Hillel J. Einhorn*, Robin M., editor. Chicago: University of Chicago Press, 1990.

Delahaye, Jean-Paul and Philippe Mathieu. “L' altruisme perfectionné." *Pour la science*, 1993, 187, pp.102-107.

Delahaye, Jean-Paul and Philippe Mathieu. “Le monde agité de la coopération." *Pour la Science*, September 1996, 227.

Dempster, Arthur P., Nan M. Laird and Donald B. Rubin. “Maximum Likelihood from Incomplete Data via de EM algorithm." *Journal of the Royal Statistical Society B*, 1977, 39, pp. 1-38.

Fehr Ernst and Suzann-Viola Renninger. “Gefangen in der Gedankenspirale.” Die Zeit, November 23, 2000, Wirschaft, p. 31. DIE ZEIT Nr.48.

Gamow, George. One, Two, Three,...Infinity: Facts and Speculations of Science. Dover Publishing, New York, 1988.

Greenberg, David and Mark Schroder. *The Digest of Social Experiments*. The Urban Institute Press, Washington DC, 1997.

Greenberg, David, Mark Schroder and Matthew Onstott. “The Social Experiment Market." *Journal of Economic Perspectives*, 1999, 13-3, pp. 157-172.

Gueth, Werner, Carsten Schmidt, and Matthias Sutter. "Fairness in the Mail and Opportunism in the Internet -A Newspaper Experiment on Ultimatum Bargaining." Mimeo (2001).

Ho, Teck, Colin Camerer, and Keith Weigelt. “Iterated Dominance and Iterated Best-Response in Experimental ‘P-Beauty-contests’." *American Economic Review,* 1998, 88, 4, pp. 947-969.

Isaac, R. Mark, James Walker, and Arlington Williams. “Group Size and the Voluntary Provision of Public Goods: Experimental Evidence Utilizing Large Groups.*" Journal of Public Economics,* 1994, 54(1), pp. 1-36.

Hofstadter, Douglas, R. "Metamagical Games: Computer tournaments of the Prisoner's Dilemma Suggest how Cooperation Evolves." *Scientific American*, May 1983a, pp. 16-26.

Hofstadter, Douglas, R. "Metamagical Games: The Calculus of Cooperation is Testes through a Lottery." *Scientific American*, June 1983b, pp. 14-28.

Kerr, Norbert L., Robert.J. MacCoun and Geoffrey.P Kramer. "Bias in judgment: Comparing Individuals and Groups." *Psychological Review*, 1996, 103, No. 4, pp. 687-719.

Keynes, John, M. *The General Theory of Interest, Employment and Money.* London: Macmillan, 1936.

Kocher, Martin, G. and Matthias Sutter. "When the 'Decision Maker' Matters: Individual Versus Team Behavior in Experimental 'Beauty-Contest' Games." Institute of Public Economics, discussion paper 2000/4.

López, Rafael. "The Beauty-Contest Integer-Game: A Theoretical Analysis." Universitat Pompeu Fabra. UPF-working paper, 2001.

Louis, Thomas A. "Finding the Observed Information Matrix when Using the EM Algorithm." *Journal of the Royal Statistical Society B*, 1982, 44, pp. 226-233.

Lucking-Reiley, David H. "Using Field Experiments to Test Equivalence between Auction Formats: Magic on the Internet." *American Economic Review*, 1999, 89-5, pp. 1063-1080.

List, John A. and Lucking-Reiley, David H. "Demand Reduction in Multi-Unit Auctions: Evidence from a Sportscard Field Experiment." *American Economic Review*, September 2000, vol. 90, no. 4, pp.961-972.

Maddala, G.S. *Limited Dependent and Quantitative Variables in Econometrics*. Cambridge University Press, 1983.

Marshak, Jacob. "Economics of Inquiring, Communications, Deciding." *American Economic Review Proceedings*, 1968, 58, pp. 1-18.

Nagel, Rosemarie. "Experimental Results on Interactive Competitive Guessing." Discussion Paper No. B-236, University of Bonn, 1993.

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

Nagel, Rosemarie. “A Survey on Experimental “Beauty-Contest Games”: Bounded Rationality and Learning," in *Games and Human Behavior, Essays in Honor of Amnon Rapoport.* Eds. D. Budescu, I. Erev, and R.Zwick. Publisher: Lawrence Erlbaum Associates, Inc., New Jersey, 1998, pp.105-142.

Netlab Workshop Report. http://www.uiowa.edu/~grpproc/netlab.htm. 1998.

Plott, Charles R. “Rational Individual Behavior in Markets and Social Choice Processes: The Discovered Preference Hypothesis,” in Arrow, Kenneth, Colombatto, E. Perlaman, M. and Schmidt, C., editors, *The Rational Foundations of Economic Behavior.* Macmillan: London, 1996.

Ross L., D. Greene and P. House. “The ¨False Consensus Effect¨: An Egocentric Bias in Social Perception and Attribution Process." *Journal of Experimental Social Psychology*, 1977, 13, pp. 279-301.

Selten, Reinhard and Rosemarie Nagel. "1000DM zu gewinnen." *Spektrum der Wissenschaft*, November 1997, p. 10.

Selten, Reinhard and Rosemarie Nagel. “Das Zahlenwahlspiel-Hintergruende und Ergebnisse." *Spektrum der Wissenschaft*, February 1998, pp. 16-22.

Stahl, Dahl O. “Rule Learning in a Guessing Game." *Games and Economic Behavior*, 1996, 16(2), pp. 303-330.

Stahl, Dahl O. “Is Step-j Thinking an Arbitrary Modelling Restriction or a Fact of Human Nature?" *Journal of Economic Behavior and Organization,* 1998, 37(1), pp. 33-51.

Thaler, Richard. "Competition." *Financial Times*, section Mastering Finance 1, May 9, 1997, p.29

Thaler, Richard. “Giving Markets a Human Dimension." *Financial Times*, section Mastering Finance 6, June 16, 1997, pp. 2-5.

Weber, Roberto. “’Learning’ with no Feedback in a Competitive Guessing Game." Carnegie Mellon University, mimeo, 2000.

Appendix A

Examples of the 5 different types of reasoning processes and of group decision-making by participants in the E and S experiments (translated from Spanish or German):

1. Fixed point
E#986: "I choose 1. This is what is nearest to x = 0, which is the only number equal to 2/3 of itself. Logical answer."

2. ID plus rounding, trembling and other rules of thumb
E#3237: "If everybody would choose 100, the maximum number that could be chosen is 66.6. Therefore, theoretically nobody will send a number over 66.6 and, if you multiply this by 2/3 we get 44.4. Therefore, in theory, nobody should be sending either a number over 44.4. Following this reasoning process the only number that should be sent is 1. However, I understand that many different people participate in this game and not everybody will apply the reasoning process explained above. Therefore, and taking into account that the majority of people would go all the way up to 1, I choose 6.8."

3a. IBRd Level-∞ plus rounding, trembling and other rules of thumb
S#1206: "In case that all numbers are equally distributed, the average will be 50. 2/3 of that is about 33. Since the readers of Spektrum are certainly not the dumbest, they will all get to 33 at the first step. However, 2/3 of that is 22. Since certainly all will calculate this, one has to take 2/3 of that... The series continues at infinitum and at the end you get 0! However, I choose, despite that logic, 2.32323."

3b. IBRd Level-1 plus rounding, trembling and other rules of thumb
E#663: "If all the numbers had the same probability of being chosen, the mean would be 50 and the choice should be 2/3 50=33.33. However, I have estimated a percentage of deviation around 33.33 of 10% and, therefore, I choose the number 30."

3c. IBRd Level-0
S#1591 [chooses 42 with the following explanation]: "Even though I know I won't win, I take the answer from the question of life, universe and the rest (see Douglas Adams, "The Hitchhiker's Guide to the Galaxy") and use it for everything. Maybe I will also use it for this quiz."

4. *IBRnd*
E#1811: "I choose the number 15.93. The reasoning is the following: I assume
10% do not have a clue and pick the mean 50
20% give a naive answer: 33=50*2/3
50% go a second round: 22=33*2/3
5% go a third round: 14=22*2/3
5% are really devious and choose 10=14*2/3
10% are crazy mathematicians who choose 1."

5. *Experimenter*
E#1984: "I decided to run an experiment with a group of friends. Since I believed that the sample was representative of the participants in the general experiment, I assumed the result of the experiment would be a good indicator of the solution. People used the following reasoning. One said simply the mean, 50 (!!!). Some others multiplied 2/3 by 50 and said 33.33. One said 25 because "today is the

25th". In some other cases people said 1, or a number close to 1 even though in one case the reason was "to pick a number at random". The mean was around 13 and, therefore, my answer is 8.66666."

6. *Group decision-making [italics added]*
S#1172: I would like to submit the proposal of students of my math class Grade 8e [*about 14 years old*] of the Felix-Klein-Gymnasium, Goettingen, for your game: 0.0228623. How did this value came up? Johanna …asked in the math-class whether we should not participate in this contest. The idea was accepted with great enthusiasm and lots of suggestions were made immediately. About half of the class wanted to submit their favorite numbers [*IBRd Level-0*]). To send one number for all, maybe one could take the average of all these numbers [*experimenter*].

A first concern came from Ulfert, who stated that numbers greater than 66 2/3 had no chance to win (*ID*). Sonja suggested taking 2/3 of the average [*IBRd*]. At that point it got too complicated for some students and the decision was postponed. In the next class Helena proposed to multiply 33 1/3 by 2/3 and again by 2/3 [*IBRd*]. However, Ulfert disagreed, because starting like that one could multiply it again by 2/3. Others agreed with him that this process could be continued. They tried and realized that the numbers became smaller and smaller. A lot of students gave up at that point, thinking that this way a solution could not be found. Others believed to have found the path of the solution: one just had to submit a very small number [*IBRd*].

However, they could not agree about how many of the people participating would become aware of this process. Johanna supposed that the people who read this newspaper were quite sophisticated. At the end of the class, 7 to 8 students heatedly continued to discuss the problem. The next day I received the following message: […] we think it is best to submit the number 0.0228623 [*we classify this comment as Level-∞ IBRd plus trembling*].

	Financial Times	Expansión	Spektrum der Wissenschaft
Number of participants	1,476 participants	3,696 participants	2,728 participants
Numbers/ Interval to choose from	Integer number in [0,100]	Number in [1,100]	Number in [0,100]
Explanation of "2/3 of the mean"	With an example: 5 people choose 10, 20, 30, 40, 50. The average is 30, 2/3 of which is 20. The person who chooses 20 wins.	With a definition: suppose 1000 persons participate. Sum the chosen numbers and divide them by 1000. Multiply the result by 2/3. The winning number is the closest to the last result	No explanation of mean or 2/3 of mean is given. 2/3 of mean is called "target number"
Comments asked	"Please describe in no more than 25 words the thought processes you went through in arriving at your number"	"If you want to add some comment about how you decided to choose your number, we are interested in it"	"We will be glad when you also tell us how you got to your number"
Prize	2 return Club Class tickets to New York or Chicago donated by British Airways	100.000 Pesetas (about $800), paid by Expansión	1000 DM (about $600) paid by Spektrum
Announce-ment of the rules	Once	Pre-announcements of the game; appearance of rules on 4 consecutive days	Once in print and in their webpage
Time to submit	13 days	1 week	2 weeks
Submission form	Postcards	Letters, fax, or e-mail	Letters or e-mail
Other restrictions	One entry per household, minimum age 18, resident of UK; excluded: employees of FT or close relatives, any agency or person associated with the competition	One entry per person. Personnel of Universitat Pompeu Fabra and direct family excluded	One entry per participant. Employees of Spektrum excluded
Cover story, context of experiment	Competition as "appetizer for the FT Mastering Finance series"... "Contest will be discussed … in an article on behavioral finance…. The series will offer a mix of theory and practical wisdom on … corporate finance, financial markets and investment management topics"	"This is an exercise, an experiment ... related to economics and human behavior. John Maynard Keynes could say that playing at the stock market is similar to participating in a Beauty-contest game…"	"Who is the fairest of them all? The average... according to psychological tests. However, sometimes it helps being different from the average by the right amount." Tale about a country Hairia where the most beautiful person is the one who has 2/3 of the hair-length of all contestants
Language	English	Spanish	German
Description of newspaper/ magazine	Daily business paper, worldwide distribution, printed in England, with 391,000 copies per day.	Daily business paper, distributed in Spain with 40,000 copies per day.	Monthly magazine, German edition of Scientific American, distributed in Germany, with about 120,000 copies per month.
Authors	Thaler	Bosch, Nagel	Selten, Nagel

Table 1. Main features of the Newspapers experiments

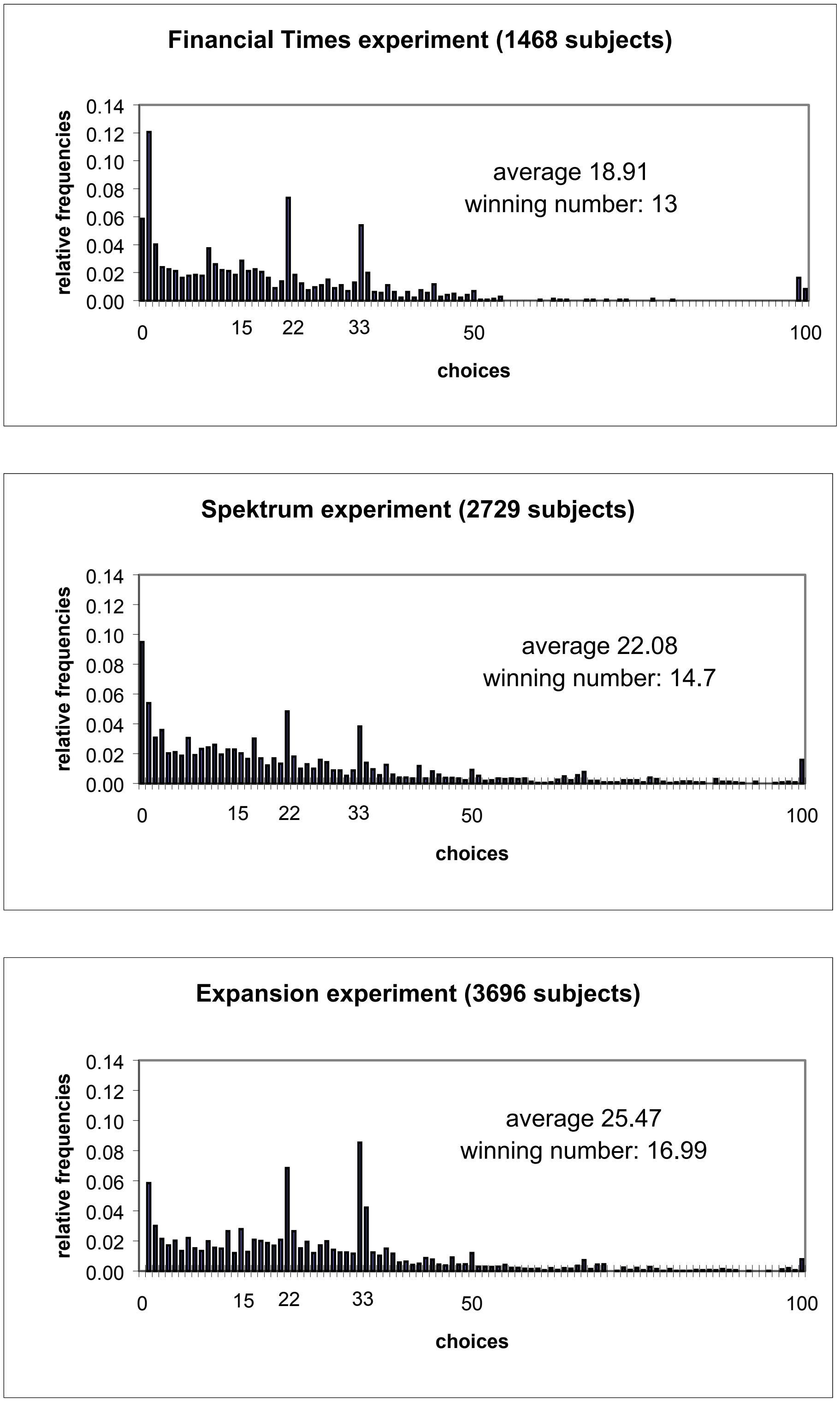
Types of reasoning processes	Relative frequencies
Fixed point	2.56%
Equilibrium, without further explanation	14.61%
Iterated dominance (ID)	13.77% of which 11.10 percentage points are Level-∞
Iterated best reply degenerate (IBRd)	54.71 % of which 25.45 percentage points are Level-∞ 12.47 percentage points are Level-0
Iterated best reply non degenerate (IBRnd)	9.28%
Experimenters	5.09%

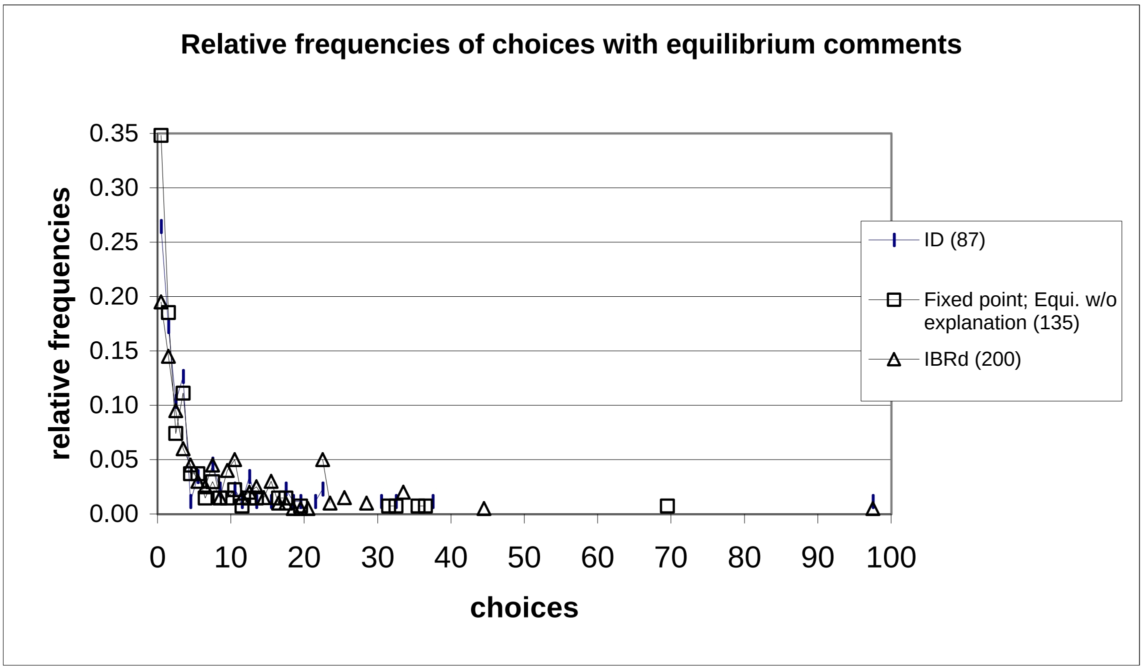
Table 2: Relative frequencies of the different types of reasoning from the comments of E and S experiments

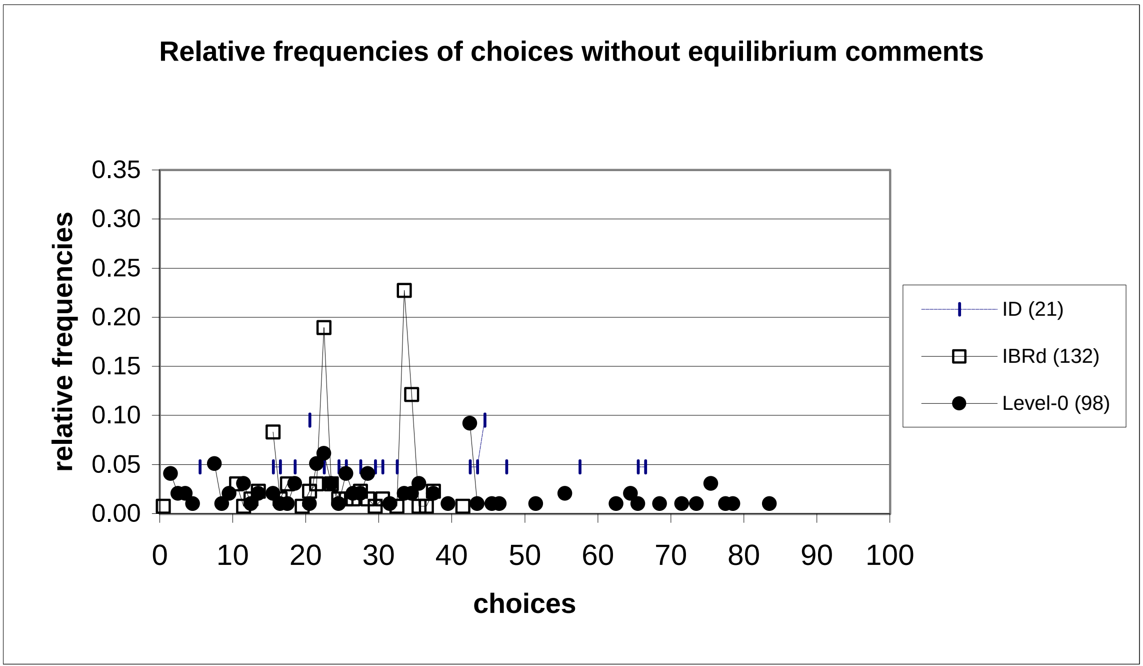
Experiment **(Month/ year)**	**Data from**	**Subject pool**	**No. of players per session (total)**	**Payoffs**	**Time to submit the Number**	**Submission by type**	**Comments**
1. Lab # 1-5 (8/1991, 3/1994)	Nagel (1995), Nagel (1998)	Undergraduates from various Dpts. at Bonn and Caltech (# 5)	15-18 (86)	20 DM to winners, 5 DM show up fee, \$20 and \$5 show up fee, split if tie	5 min.	Immediately	Optional
2. Classroom # 6,7 (10/1997)	Collected by Teachers at UPF: Charness , Hurkens, Lopez, Nagel	2nd year economic undergrads UPF, in Economic Theory class. Limited knowledge in game theory	30-50 (138)	3000 Pesetas (\$24), split if tie	5 min.	Immediately	Optional
3. Take-home # 8,9 (10/1997)			30-50 (119)		1 week	Hand-in personally	Optional
4. Theorists # 10 (12/1997)	Collected by Rockenbach	3rd-4rth year undergraduates in Game Theory class, Bonn	54	30 DM (\$18), split if tie	3 weeks	Hand-in personally	Optional
# 11, 12 (6, 10/1997)	Collected by Nagel	Game theorists/ Economists in Conference	20-40 (92)	\$20 split if tie	5 min.	Immediately or e-mail	Optional
# 13 (11/1995) by e-mail		Profs/doctorates of Department Business/Economics in UPF		Handbook of Exper. Economics. Random draw if tie	1 week		
5. Internet newsgroup # 14 (10/1997)	Collected by Participant in S. See Selten and Nagel (1998)	Newsgroup in Internet (responses Via e-mail)	150	30 DM (\$18) or book	1 week	e-mail	Optional
6. Newspaper # 15 (5/1997)	Thaler (1997) in *Financial Times*	Readers of FT	1476	2 tickets London-NY or London-Chicago	2 weeks	Letters	Required to become a winner
# 16 (5/1997)	Bosch, Nagel (1997) in *Expansión*	Readers of E	3696	100.000 Pesetas (\$800)	1 week	Letter, e-mail fax	Optional
# 17 (10/1997)	Selten, Nagel (1998) in *Spektrum der Wissenschaft*	Readers of S	2728	1000 DM (\$600) random draw if tie	2 weeks	Letter, e-mail	Optional

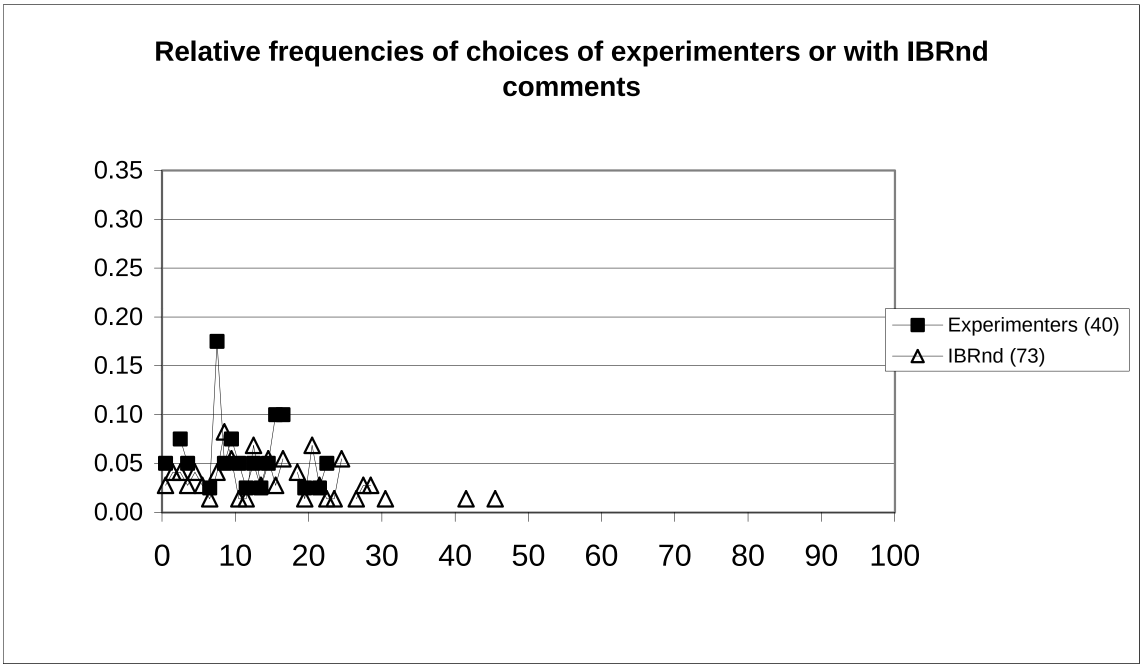
Table 3. Design and structure of 17 experiments, classified into six groups

Figure 1a-c: Relative frequencies of choices in the three Newspaper experiments









Figures 2a-c: Relative frequencies of choices of those subjects who made comments (in parenthesis of the labels are the numbers of observations of each type)

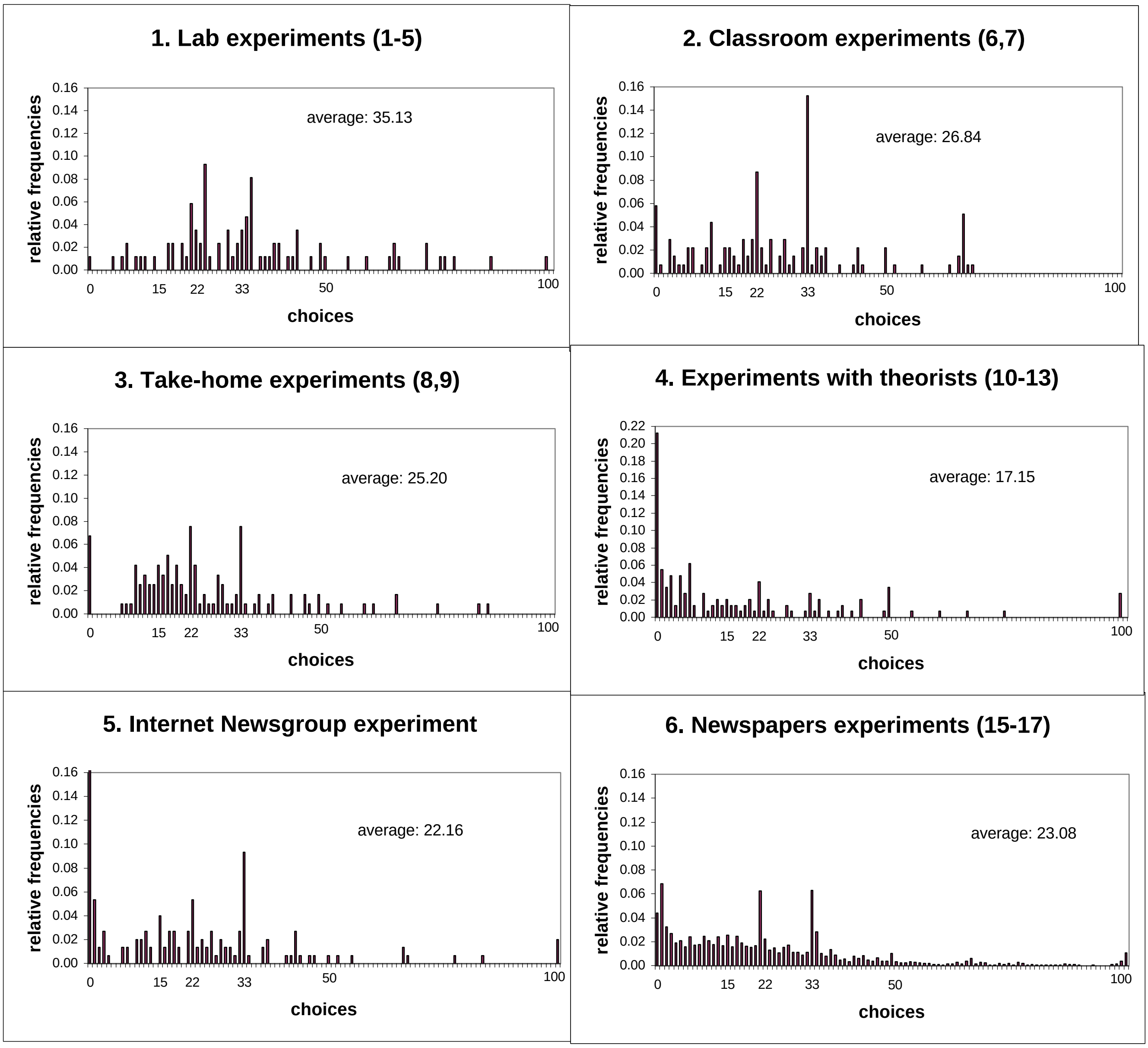


Figure 3: Relative frequencies of choices in the six groups of experiments