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

by

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

Universitat Pompeu Fabra, Barcelona

November 21, 1999

Abstract

"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 the uncontrollability of these experiments. These results are then compared with lab experiments involving undergraduates and game theorists as subjects, in what must be one of the largest empirical corroborations of interactive behavior ever tried. We claim that all observed behavior, across a wide variety of treatments and subject pools, can be interpreted as iterative reasoning. Level-1 reasoning, Level-2 reasoning and Level-3 reasoning are commonly observed in all the samples, while the equilibrium choice (Level-Maximum reasoning) is only prominently chosen by newspaper readers and theorists. 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.

1. Introduction

In June 1997, Richard Thaler on one hand, and Antoni Bosch-Domènech and Rosemarie Nagel on the other, totally unaware of each other endeavor, designed and announced an experiment on the Beauty-contest game in two different business daily newspapers, in the UK and Spain, inviting the readers to participate. Five months later, Reinhard Selten and Rosemarie Nagel (1998) replicated the experiment in a German scientific magazine.

When designing an experiment, many elements are taken into account that can influence its results. Just to mention a few: 1) Physical environment, 2) Subject pool (gender, education and training, group identification and friendship, etc.), 3) Number of subjects, 4) Communication among subjects, 5) Information available to subjects, etc. Each represents a potential treatment or control. When experimenting with newspaper or magazine readers (i.e., announcing an experiment in a daily newspaper and inviting its readers to participate in it) the experimenter looses control simultaneously of all these elements. In addition, two other elements that play a crucial role in any experimental design can be changed dramatically: 6) Reward, 7) Duration of the experiment.

However, running experiments in a newspaper helps to answer questions like the following. 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 the test the critical assumption of "parallelism" between the lab and the field.

Also, experimenting in newspapers has advantages. They are cheap, since sponsors usually finance prizes. They do not require having a lab or easy access to students. They allow for a variety of subject pools in terms of interests, knowledge and nationalities, each pool corresponding to the particular readership of each newspaper. And potentially, they have a huge educational impact on the public at large, being advertised, described, and analyzed in the mass media.

The basic elements of a typical experiment are the following: It 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 equivalent to the minimum salary per hour. A number of experiments tried to go beyond the basic procedure of experimenting. The Iowa Electronic Markets may be the best known of them. The advent of Internet has allowed some experimenters to move out of the lab. Bossaerts and Plott (1999), for instance, have run several experiments using the Internet as a medium to collect experimental data, subjects being able to lock in any time they want within a range of several days. Lucking-Reiley (1999) and List and Lucking-Reiley (1999) test different auction mechanisms selling sports-cards in a real market or on the Internet. See also Isaac et al (1994) about the difficulties of large-scale experiments. For a pre-Internet experiment involving hundreds of subjects, Bohm (1972) is a classic example.

Economists have always questioned whether subjects other than students behave differently in a given setting. For example, Cooper et al. (1999) study behavior in ratchet effect games with students, experienced managers, and white-collar workers from China. The main difference observed is that managers behave differently in games with real world content and with abstract content. Students act the same in both contexts. Cross-cultural studies have become fashionable after the work of Roth et al (1991).

Yet, we are not aware of any experiment run in newspapers or magazines. These are experiments not involving students, with numbers of subjects (drawn possibly from different cultures) in the thousands, time availability extended to several weeks or months, and with rewards beyond the usual financial constraints. In a more fundamental sense, these are experiments with much less control, in a zone between a fully controlled experiment and a survey.

In this paper we first compare three newspaper experiments of the Beauty-contest game. 1) The experiment run in the *Financial Times* (FT from now on) by Richard Thaler (1997). 2)

The experiment run in *Expansión* (E from now on) -a Spanish daily business newspaper similar to *Financial Times*- by Antoni Bosch-Domènech and Rosemarie Nagel (1997a, 1997b, 1997c). And 3) the experiment run in *Spektrum der Wissenchaft* (S from now on) -a monthly German science magazine, the German edition of *Scientific American*- by Reinhard Selten and Rosemarie Nagel (1998). We also comment on some methodological points that arise when experiments are proposed to newspaper and magazine readers.

Second, we relate these experiments to similar ones run in labs, as reported in Nagel (1995), and to new experimental data involving students, economists and game theorists as subjects. Third, an essential feature of the present paper is the integration of the statistical analysis of these independent experiments -involving widely different subject pools, sample sizes, payoffs and settings- in a single statistical model. This model enables to capture aspects of the individual decisions that fit the predictions of a particular game-theoretical model, the model of iterated best reply, as discussed in Nagel (1995, 1998), Stahl (1996) and Ho et al. (1998).

The paper is organized as follows: in section 2 we explain the game and its theoretical predictions. Section 3 contains the design of the newspaper experiments. In section 4 we report the results of these experiments. In section 5 we compare the lab experiments with the newspaper experiments. In Section 6 we describe the statistical methods applied to further the comparison. Section 7 gives the results of the statistical analysis and Section 8 concludes.

2. The game

A basic Beauty-contest game is as follows. A certain number of players each chooses simultaneously a decimal number, let us say, from 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. 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.

The game is dominance solvable. The process of *iterated elimination of weakly dominated strategies* leads to the game's unique equilibrium in which everybody chooses 0[1]. Thus, a rational player does not simply choose a random number or his favorite number, nor does he choose a number above $100p$, since it is dominated by $100p$. Moreover, if he believes that the other participants are rational as well, he will not pick a number above $100p^2$; and if he believes that the others are rational and that they also believe that all are rational, he will not pick a number above $100p^3$ 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 subjects actually apply.[2]

Game theory has mostly developed using deductive and refinement concepts. However, their scarce predictive power in some experiments has turned the attention to alternative approaches based on bounded rationality and heterogeneity of beliefs. Obviously, once individuals are assumed not to be fully rational or to be diverse in their beliefs, selection of equilibria becomes an empirical matter (see, e.g., Schelling (1960), Stahl and Wilson (1995), Costa-Gomes et al. (1999)).

For the Beauty-contest experiments, Nagel (1995), Stahl (1996) and Ho et al. (1998), show that a model of *iterated best reply* describes subjects behavior better than the equilibrium obtained by iterated elimination of dominated strategies. They classify subjects according to the number of steps, or depth, of their reasoning. Accordingly, a Level-0 player chooses arbitrarily in the given interval, with the mean being 50. A Level-1 player gives best reply to Level-0 players and thus chooses $50 \cdot 2/3 = 33.333$. A Level-2 player chooses $50 \cdot (2/3)^2$ and so on. A player, who takes infinite steps and believes that all players take infinite steps, will choose the equilibrium 0.

---

[1] The number of steps is infinite. When subjects choose in [1,100] (as in E), a finite number of reasoning steps leads to the equilibrium. If only integers are allowed (as in F) there are several equilibria. In the case of $p = 2/3$, there is an additional equilibrium to "all choosing 0," which is "all choosing 1." This is a minor modification that should 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 López (1999)).

The hypothesis of iterated levels of reasoning predicts that choices will be clustered around the values 33.33, 22.22, 14.81, 9.88, … and 0.

3. Newspapers' Experiments

3.1 Design

Participants in the Newspapers' experiments are asked to choose a decimal number in [0,100] (in [1,100] in E, non-negative integers only in FT). The winner is the person who chooses the number closest to 2/3 of the average number submitted. Rewards offered and time available in the Newspapers' experiments were much larger than rewards offered and time available in the lab as can be seen in Table 1.

Thaler (1997) and Bosch-Domènech and Nagel (1997a) wrote the instructions without knowing about each other's plan of running a newspaper experiment. Selten and Nagel (1998) had both sets of E and FT instructions when writing their sets of rules for S. Table 1 summarizes common aspects and differences between the three games run.[3]

The newspapers' editors imposed 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 lawyer of FT. The lawyer thought that a game with decimal numbers becomes a game of pure luck, and gambles by private persons or institutions are 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."

---

[2] For a survey on the Beauty-contest experiments, see Nagel (1998).
[3] All data sets used in this paper are available upon request.

Items	Financial Times	Expansión	Spektrum
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 results by 2/3. The winning number is the closest to this 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
Announcement of the rules	Once	Pre-announcements of publication of the game; appearance of rules on 4 consecutive days	Once
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.
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, world wide 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

Expansión requested that the opening article included a reasoned justification for performing the experiment. This newspaper did also several pre-announcements of the game, days before the opening article appeared. Furthermore, without the authors' knowledge, *Expansión* published a shortened version of the opening article containing the rules of the game on the three consecutive days following its publication. The shorting resulted in the omission that comments were welcome and 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.

Finally, *Spektrum der Wissenschaften* asked Selten and Nagel to write a fairy tale, introducing the spirit of the game. The editor changed the story submitted of a nose-length Beauty-contest to a hair-length Beauty-contest.

These reported interventions on the part of the newspapers' editors are examples of how experimenters who use the mass media as a platform for their experiments may be constrained in ways that they do not face in the lab. Fortunately, we believe that none of the described interventions had a significant influence on the results. But other differences in the design of the experiments, imposed or not by the newspapers' editors, may have had an impact on the results and are worth reporting here.

1) Only entrants in the FT experiment were forced to explain their decision[4]. It is well known among experimentalists that requiring subjects to provide explanations may force them to think the decisions over, bringing about more thoughtful results. However, in E, the average choice of those entries, which submitted comments was 25.2, whereas the average choice of entries without comments was 25.5.

2) We know from lab experiments that giving examples can tip decisions. In FT, an example was provided with the number 20 as the winner. It is not farfetched to ponder whether using an example that focuses on low numbers did not push entrants towards small ones. Indeed, in

---

[4] The FT lawyer believed that if subjects had to make a comment showing their skills in choosing a number, this could be used in court to prove that the game was not a gamble.

FT, numbers above 50 were much less frequent than in the other two publications.[5] In addition, as teachers we know that examples are good didactic devices. Thus, giving an example may also have helped to educate entrants about the game.[6]

3) The number of participants was a parameter that was not directly controlled by the experimenters, but was certainly conditioned by the design. In E, in spite of having half the time to answer compared to FT and S, the number of participants was the largest of all. One reason was the fact that the game was prominently announced in E, where an advertisement on the first page was run for several days. Another reason might have been allowing fax and e-mail entries in E, unlike FT which only accepted mailed letters.

4) We all very well know that using e-mail requires little effort. This may have enticed the participation of people who did not want to spare one moment to think about the game ("noise players" if you want). Forbidding the use of e-mail and forcing entrants to sit down to write their decision on a piece of paper, put it inside an envelope, place a stamp on it, and post it, could also have had an "educational" effect. However, a comparison of the numbers submitted by mail or by e-mail does not show significant differences.

5) Communication among subjects could not be constrained. In the results section we will report on some of the consequences of this loss of control, for example, as observed collusion among some participants.

3.2. Results

Figures 1 to 3 show the relative frequencies of the numbers chosen (the rounding to an integer is from 0 to 0.5, 0.5 to 1.5, etc.) in the three Newspapers' experiments. The figures also indicate the number of participants, the average of all numbers and the number submitted by the winner. In total there were 7900 participants (3,696 in E, 1,476 in FT and 2,728 in S).

---

[5] We will show that when we exclude numbers above 50, the distributions of numbers in FT and S are strikingly similar.

[6] Which, incidentally, may help to explain the higher numbers of 0s and 1s in FT than in E.

The results seem to confirm the existence of a common pattern of decision-making among participants, previously identified in lab experiments of the Beauty-contest game and interpreted as steps of iterated dominance (see references above):

The most popular numbers in all three experiments are two-thirds of 50 (about 33), two thirds of this number (about 22) and the equilibria of the game (0 and 1 in FT, 1 in E and 0 in S). The steps of iterated dominance interpretation claims that in the Beauty-contest game people reason in steps. Step 0, which would be the preliminary step of any reasoning, translates into numbers that are arbitrarily distributed over the interval, resulting in an expected value of 50 (50.5 if numbers are from 1 to 100). Level-1 reasoning is $(2/3) \cdot 50 = 33.333$. Level-2 reasoning is $(2/3) \cdot 33.333 = 22.22$ and so on. Taking this reasoning to all its steps of iterated dominance would lead to choose the Nash equilibrium[7]. We report this finding as:

Fact 1: The numbers obtained by the process of reasoning in steps coincide with the peaks observed in the experiments. This is particular true of the first two steps and the Maximum step (and to some extent of the third as well).

From the submitted numbers it should be possible to identify those entrants that take one reasoning step, those taking two steps, perhaps even those taking three steps, and those taking all the steps down to the Nash equilibrium. This will be done in Sections 6 and 7, using statistical techniques.

Between the equilibrium choice and step 2 or 3 there are no other notable peaks and in the comments submitted we could not find anybody who stopped at just 4 steps of reasoning. Therefore we infer the following additional fact.

---

[7] As mentioned in our footnote 1, when subjects have to choose in [1,100], the number of reasoning steps necessary to reach the Nash equilibrium, 1, is finite. Not so when subjects have to choose in [0,100]. Ho et al. (1998) call the two types of games finite-threshold and infinite-threshold, respectively. In our one-shot games we do not observe any difference in the results of the two types of games that can be reasonably attributed to this particular characteristic.
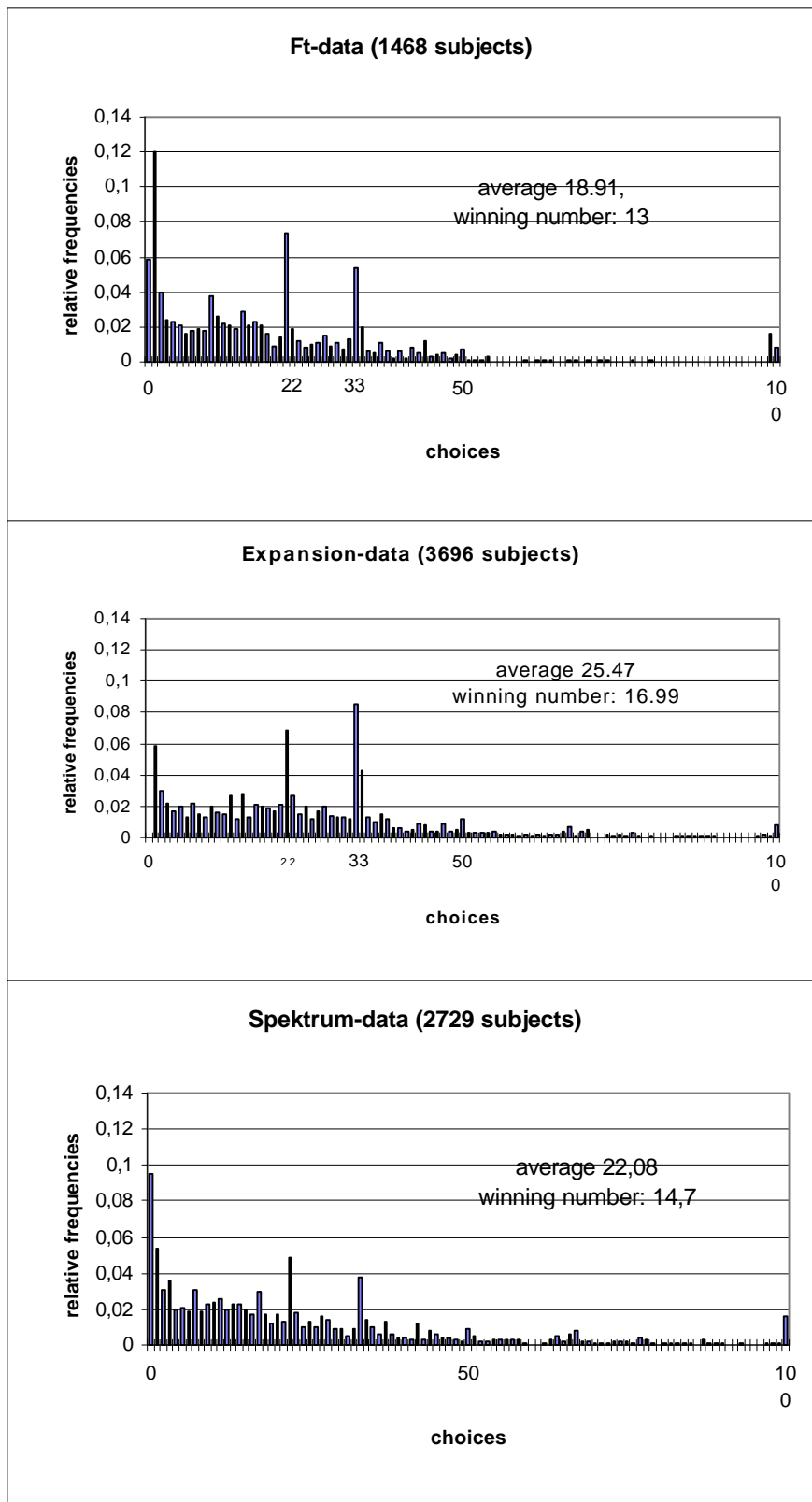
Figure 1 to 3. Frequencies of choices in the three Newspapers' experiments

Fact 2: Once subjects reach the second, or third reasoning level, they jump all the (infinite) steps towards Nash: One, two (three), infinity.

Analyzing the 50 comments of those entrants in the S experiment who describe the equilibrium process of infinite iteration[8], we find that the average number chosen by them is 6, with 80% of them choosing below 10. We state this as:

Fact 3: A significant proportion of the subjects (29/50 = 58% in our sample) who reach all the way to Nash, bounce back to choose a number larger than the equilibrium. However, most of them (19/29 = 65%), stay below 10 and thus clearly below the numbers corresponding to levels of reasoning from 1 to 3 (numbers around 33, 22 and 15).

Many 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 that confirms this observation: Many rational players do not choose the Nash equilibrium because they take into account the bounded rationality of others. But there is another side to this fact:

Fact 4: A significant proportion of subjects (21/50 = 42% in our sample) who comment on reaching the equilibrium, choose a number from 0 to 1.

Turning the previous statement upside down, we could say that a phenomenon that appears rational (choosing the Nash equilibrium) may be the result of players expecting, irrationally, that other players will behave rationally. In other words, that what is taken for rational behavior represents, in fact, a boundedly rational ignorance of other players bounded rationality. In psychology this is known as "false consensus" (see Dawes (1990)), a situation that appears when people assume that other players reason as themselves.

Notice that numbers above 50 are scarcer in FT than in the other two experiments. One reason for this fact, already suggested above, may be that in the FT description of the game an example that selected a low number was presented. Since this may have pointed the way towards lower numbers in the FT experiment, we may be justified in comparing the experiments using

---

[8] We randomly selected 50 comments out of all comments that identify the equilibrium, and classified their behavior.

only numbers lower than 50. In Figure 4 we compare the S results and the FT results in a quantile-quantile plot, when all the results above 50 have been removed form the two experiments. In the figure, it appears that the distributions of the numbers chosen in both experiments are almost identical, *despite the non-normality of both distributions, the different subject pools, and the large number of subjects involved.* Indeed, there is no significant difference between the two distributions when applying a conventional test, like the Mann-Whitney-U-test, despite the high power of the test induced by the large number of observations. Here we have, then, two experiments, run among two large populations presumed to be rather different (one made up mostly of UK businessmen and economists, the other made up of German scientists), whose results are undistinguishable.
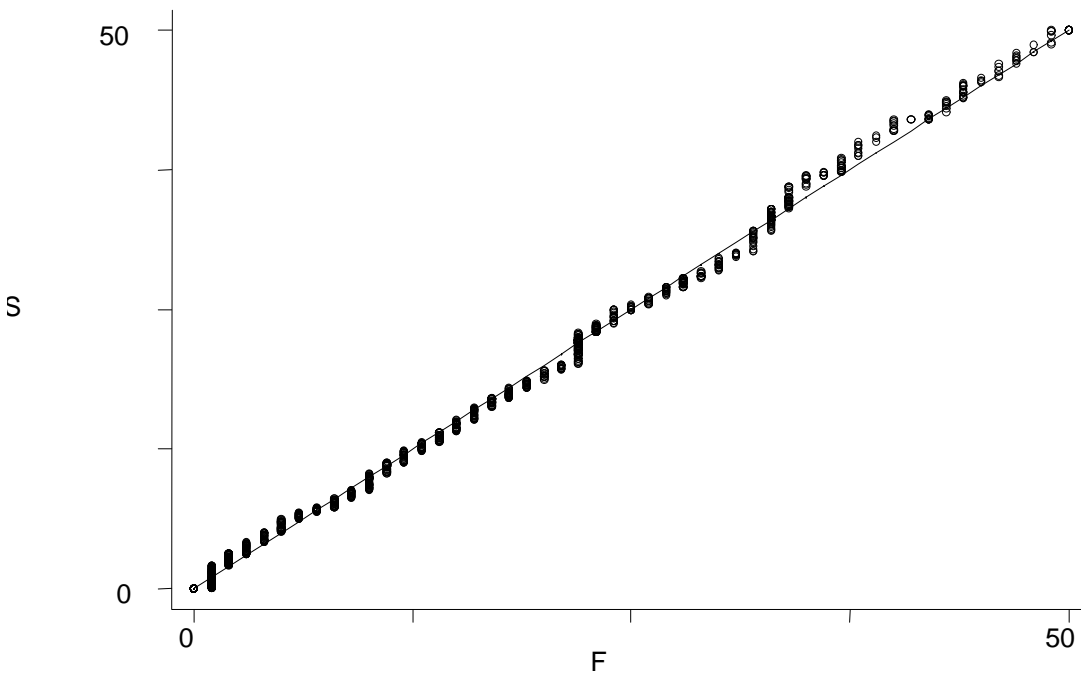
Figure 4. Quantile-quantile plot of choices 0 to 50 in FT and S

We can summarize this observation as:

Fact 5: Two large subject pools, drawn from different countries and from different professional backgrounds, answer the Beauty-contest game in ways that are statistically indistinguishable.

13

5. Comparisons with lab experiments

One purpose, perhaps the main 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 due, most of them, to the increased loss of control occurred in experiments carried out in newspapers. This loss of control should reshape the design of experiments' methodology when moving away from the full control provided by the lab. It involves:

1.   Sampling error

Experimentalists know that there is sampling error in their experimental results. They know that their subjects are not representative of the population at large, but are confident they have the ability to control the relevant characteristics of their samples. In a newspaper experiment, the experimenter looses some additional control of the sample.

2. Information seeking

Subjects of newspaper experiments may go to great lengths to submit informed answers. One interesting variety of information-seeking behavior on the part of some subjects consists, of all things, in running a parallel experiment. Some participants mentioned that they had run versions of the same experiment among friends, relatives or school classes, to help them decide what number to submit.

In one case that we comment below in footnote 10, a potential subject of the S experiment started in the Internet his own Beauty-contest. Another reader of S run the experiment

in her math-class and then submitted the joint bid of her class mates[9] (which was close to Nash equilibrium,).

Of the 21 subjects in the S game who reported that they had run their own experiment, 50% chose a number between 12 and 17 (the winning numbers in the three newspapers' contests were in this interval), whereas in the total population only 14% chose a number in this interval.

In addition, quite a few participants used an Excel spreadsheet in order to calculate the best response to a distribution of numbers selected by them, typically with the pivotal numbers 33 and 22 having a high number of frequencies. Whether this behavior may or may not have an undesirable effect on experiments run in newspapers is perhaps a moot issue, but it certainly helps to propagate out of academic circles the use of the experimental methodology as a road to enlightenment.

3.   Coalition forming

If an experimenter wants to avoid the forming of coalitions in the lab there are usually ways of ensuring that this is the case. Not so in a newspaper experiment. In fact, we know that in all three experiments there were attempts of coalition forming[10], although with little impact on the results (except for a larger than expected frequency of 100). Should newspapers experiment become more prevalent, and prizes larger, readers might find it worth organizing stable coalitions to participate in them, in the same way that it pays to organize rings of bidders in public auctions.

4.   Number of answers per subject

---

[9] The comment (see appendix) exemplifies a wide variety of comments we received: a) the idea of choosing a favorite number, b) reasoning according to the iteration model and c) the equilibrium concept.

[10] The attempt was blatant in E. 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. In fact, we were recipients of one such e-mails and what we saw was interesting. The proposition was not a profit-sharing colluding agreement. Instead, the ring leader asked friends to start a chain-letter of entrants playing 100, mentioning without further details that this way they were participating in an "economic experiment". Are we witnessing, thanks to Internet, the spawning of a new form of collusion, with a leader controlling both information and profits, and some instrumental followers who participate without gain or cost?

When large numbers of subjects are involved and the means of communicating between subjects and experimenters are diverse, it is quite difficult to enforce the rule of "one person one answer". On the other hand, this rule becomes less important as more subjects are involved. In newspaper experiments it may not even make any precise sense to talk about number of answers per person. One person can send many answers under different names or have her entire family, or her friends, sending answers on her behalf. We know, for instance, that a schoolteacher invited over 100 students to participate in the S experiment. The coalition forming problem and the problem of several answers on the part of a single individual may easily melt into each other. Nevertheless, it is conceivable that, in some newspaper experiments, a very determinate person might affect significantly the results by entering a large number of decisions.

Of these potential differences between lab and newspaper experiments, the first two have revealed themselves as the most interesting. In the remaining of this section we present and compare the main features of 17 different experiments collected from different sources (see Table 2). Experiments 1 to 5 involve undergraduate students (1-4 in Bonn, 5 in Caltech) with about 5 minutes to think and no communication among them. Experiments 6 and 7 are take-home experiment and experiments 8 and 9 are in-class sessions which all involve 2nd year undergraduate students from the Intermediate Micro classes in Universitat Pompeu Fabra with very limited knowlegde in game theory. Communication among the students was not constrained. Experiment 10 is a take-home exercise with game theory students of Bettina Rockenbach, University of Bonn. Experiments 11 and 12 involve economists (mainly game theorists and experimentalist). Experiment 13 was done by e-mail at Universitat Pompeu Fabra among colleagues and graduate students from the Department of Economics and Business. Experiment 14 is an e-mail game run by a participant of the S experiment[11], and Experiments 15 to 17 are the ones done with readers of S, E and FT, respectively, and reported above.

---

[11] To help making his decision, as mentioned above, one participant in the S experiment decided to run on Internet his own replication of the experiment. The answers that he received show the same common pattern of iterated reasoning. The winning number in his experiment was 14.81. He submitted a slightly lower number (14.2) and was very close to win the S prize, the winner being the number 14.7. A difference of 0.1 points in two experiments, one

Exp. Numbers (Month/ year)	Data collected By	Subject pool	No. of players per session (total)	Payoffs	Time to submit the number	Submission by type	Comments
1-5 (8/1991, 3/94)	Nagel (1995), Nagel (unpub.)	Undergraduates at Bonn and (various faculties) at Bonn and Caltech	15-18 (86)	20 DM to winners, 5 DM show up fee $ 20 and $5 show up fee	5 min.	Immediately	optional
6-9 (10/1997)	Various Instructors at UPF (unpub.)	2nd year economic undergraduates at UPF	30-50 (257)	3000 Pesetas ($24), split if tie	5 min. or 1 week	Immediately or hand-in personally	optional
10 (12/1997)	Rockenbach (unpub.)	3rd-4rth year undergraduates in game theory class, Bonn	54 (54)	30 DM (18 DM), split if tie	3 weeks	hand in personally	optional
11-13 (11/1995, 6/97, 10/97)	Nagel (unpub.)	game theorists/ economists in or before seminars	20-40 (92)		5 min. or 1 week	Immediately or by e-mail	optional
14 (10/1997)	Matthias, a Participant in S	Newsgroup in WWW	150	30 DM or book	1 week	e-mail	optional
15 (5/1997)	Thaler (1997) in *Financial Times*	Readers of F	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 2. Design and structure of 17 experiments

In order to assess the differences among these Beauty-contest experiments in labs and newspapers and to compare their results, we have constructed in Figure 5 the box plots associated to the responses in each of the experiments. Notice that box plots are drawn in such a way that 50% of the distribution is below the line dividing the box.

with 150 subjects, the other with 2,728!
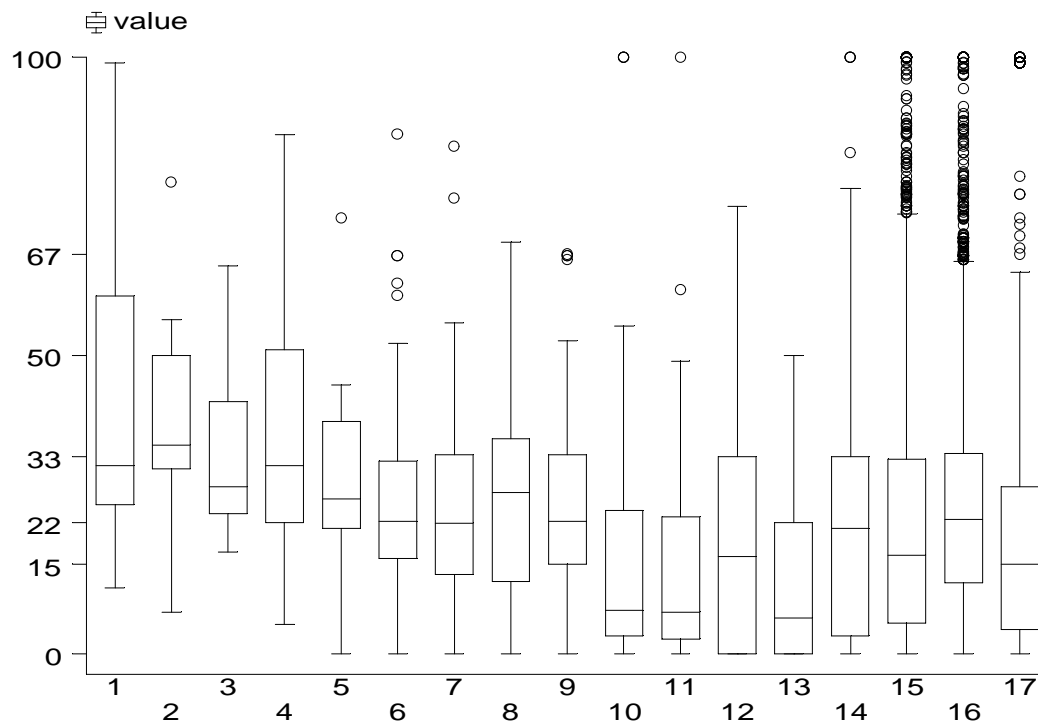
Figure 5. Distribution of data in single experiments.

The lower and upper side of the box indicates the 2nd and 3rd SÍMBOLOSÍMBOLOquartile of the distribution, Q_1 and Q_3. Therefore, the box spans the middle half of the data and the length of the box is the interquartile range (IQR). The "whiskers" at either end extend to the smallest and largest observations that are within $[Q_1 - 1.5 \, IQR, Q_3 + 1.5 \, IQR]$. Points that are outside this interval are shown with circles.

Note the high variation across experiments of the median of the distribution (i.e., the level of the line dividing the box). Among the 17 experiments, the first four are clearly distinguished from the rest by the fact that the Nash equilibrium of the game is *never* selected. This corresponds to four sessions run with undergraduates at the University of Bonn (see Nagel (1995)). This is the most significant difference in the choices observed in lab-experiments and in newspaper experiments. In experiments 10 and 11 the median of the distribution is close to zero (50% of the subjects were at the Nash equilibrium).

Other than subject pool, time availability seems to be a factor in the frequency differences observed in choosing the Nash equilibrium. To test it, we run, at Universitat Pompeu Fabra, a take-home experiment among undergraduate students with very limited knowledge of game theory, giving them one week to return their number choice (Experiments 6 and 7). We observed in it a clear increase in equilibrium choices (about 7%), still below the newspapers

18

average frequency of 13%. At the other extreme, when only game theorists are present in the experiment, the 0-choices are between 20% and 40% (experiments 10-13, with 10 and 13 as "take-home" experiments).

We can state these results as:

Fact 6: Time availability seems to have some impact on the frequency of equilibrium responses.

Fact 7: Subject-pool can have an important effect on the frequency of choices near the equilibrium.

Fact 7 represents a warning. Sampling error has to be carefully monitored in experiments.

In spite of these differences in frequencies, all experiments show a common and important pattern. We refer to the iterated best reply reasoning introduced in Nagel (1995). A thorough analysis of this pattern requires the use of statistical techniques. In the following section we comment on some of the statistical procedures used in previous work and describe the procedure that we use to analyze the data from the 17 experiments under consideration.

6. Statistical analysis of iterated reasoning

The seminal paper by Nagel (1995) uses the theoretical values of the iterated best replies starting from 50 (in experiments labeled here 1 to 4) as the center of step-k reasoning intervals. In the Beauty-contest game, this was the first attempt to characterize empirically the heterogeneous selection principles of agents assumed to iterate over best replies. From a statistical perspective, the fact that this method imposes the intervals' centers instead of estimating their locations, supposes a drawback.

Stahl (1996, 1998) describes the data of Nagel (1995) with a boundedly rational rule learning model using the iterated model $50*(2/3)^k$. An individual with type-k initial propensity (so called disposition) is more likely to use rule k than any other rule. In order to avoid identification problems, the set of dispositions is assumed to be equal to the set of behavioral rules. Players

evaluate the ex-post performance of these rules using step-*k* reasoning. The econometric specification is completed by assuming that the choices in each step of reasoning belong to the family of truncated normal distributions in the [0,100] interval. For the type called "–1", players that do not learn, Stahl (1996) includes an additional uniform distribution. From the econometric perspective the specification becomes a mixture of distributions where the parameters to be estimated are the means, variances and proportions of each type. However, from the underlying theoretical model and the particular specification assumed, the moments of these distributions are constrained across them. The estimation procedure is maximun-likelihood.

Ho, Weigelt and Camerer (1998) use also Beauty-contest results to investigate the proportion of players with different levels of reasoning. Their procedure assumes that players, at each level of reasoning, believe they are at one level of reasoning deeper than the rest. Ho et al. determine the mean and variance of each level-of-reasoning interval from the value of the mean and variance of Level-0 players who choose from a truncated normal distribution. The idea being to assign the outcomes in each interval to the different levels-of-reasoning types. Using maximum-likelihood, they estimate the mean and variance of Level-0 players and the proportion of subjects in each interval.

This approach, although interesting, suffers from several shortcomings. First, the results are very sensitive to the assumption about subjects' perceptions of each other. Even without changing the basic iterated dominance criterion, the analysis leads to very different results if players exhibit different levels of self-confidence. Second, the authors impose a restrictive specification that ties the mean and the variance of each level-of-reasoning to the mean and the variance of level-0 distribution.

Stahl and Haruvy (1998) use also maximum likelihood to uncover and test the selection principles used by different types of agents. The econometric specification is, as in the case of Ho, Weigel and Camerer (1998) and Stahl (1996, 1998), a mixture distribution model. The experimental results are obtained from twenty symmetric 3x3 games. The selection principles

tested include three deductive strategies (payoff dominance, security selection and risk dominance), three bounded rationality rules (Level-*n* rules), one optimistic rule, one pessimistic rule and a hybrid rule. Their methodology also allows testing the homogeneous population model versus the heterogeneous model. Haruvy and Stahl show that the predictive power of deductive equilibrium principles is very small and that even a Level-1 rule of bounded rationality has a hit rate much higher than deductive rules in the homogenous case. For this reason their base model includes only one deductive rule (uniform Nash equilibrium) instead of the original three. Their results show that the most likely set-up includes heterogeneous agents with a high weight of Level-1 bounded rationality rules. Deductive selection principles do not add any predictive power.

Haruvy (1998) proposes a non-parametric technique to find modes, or local maxima, in a kernel estimated probability density function. He applies this method to two data sets. The first is a set of fifteen symmetric 3x3 games. The second one is essentially the same as in Haruvy and Stahl (1998), with the players being asked to specify the distribution of the other participants' choices that led to their choices. Therefore, the data were in the form of the probabilities of the strategies of other agents as described by each agent[12]. The main idea of this non-parametric estimation is to change the bandwith parameter in order to modify the smoothing of the density surface and to find modes where probability mass is concentrated. The larger the bandwith the smoother is the surface and, therefore, less modes could be observed.

Our approach, as in Ho et al (1998) and Haruvy and Stahl (1998), is based on a mixture distribution model, but *without imposing any particular structure on the mean and variance of each distribution* in the mixture. The set of experiments that we are analyzing provides multiple group data $\{x_{ig}, i = 1, …, n_g\}_{g=1}^{G}$, where x_{ig} corresponds to the choice of the *i*th subject in the *g*th group. As can be seen from Table 3 below, we have aggregated the 17 experiments in 6 different

---

[12] An important question with this kind of data is the possibility that individuals are not able to quantify their beliefs using probabilities.

groups ($G = 6$), from Lab experiments to Newspapers' experiments. In the same table one can also find the values of the group sample sizes, n_g.

We assume that the data x_{ig} correspond to iid observations from a mixture of various normal distributions truncated at 0 and 100 and a uniform distribution in [0, 100]. The normal components are assumed, with one exception spelled below, to have *unconstrained* means m_k and variances s^2_k, $k = 1,2,\ldots, K\text{-}1$. Here K denotes the number of normal components of the mixture. The normal distributions correspond to choices under levels of reasoning 1 to Maximum, while the uniform distribution collects the remaining choices, the level 0 reasoning.[13]

We distinguish two types of parameters: 1) those that characterize the *mixing proportions*, i.e. the weights of each level of reasoning in the population and 2) the *means* and *variances* of the normal components of the mixture.

The mixture model that we consider can be written as

$$f_x (x, q) = p_1 f_1(x, z_1) + \ldots + p_K f_K(x, z_K) + p_u f_u,$$

where $f_x (x, q)$ denotes the density function of the mixture distribution at the value x. $q = (p', z')'$, where π is a vector that collects the mixing proportions p_k's, and z is a vector collecting the component parameters $z_k = (m_k, s^2_k)'$. The f_k's correspond to normal distributions of mean m_k and variance s^2_k, for $k = 1,\ldots, K$, and f_u corresponds to a uniform distribution in [0, 100].

As mentioned above, we keep the component parameters unconstrained[14], i.e., they are determined by the data themselves, with one exception. For the sake of the stability of the model when applied to small samples, we chose one of the normal distributions, the one that corresponds to the Maximum reasoning level, f_K, to have mean 1 and variance 5. The mean of 1

---

[13] We also run an estimation assuming a mixture of log-normal distributions plus one uniform distribution. The fitted distributions were very symmetrical except for the distribution of choices at the Maximum level of reasoning. The present specification with truncated normals reflects both the asymmetry at the Maximum level and the symmetry of the remaining distributions, while keeping the description of the procedure simpler.

[14] Identifiability of the mixture model requires the means of the components of the mixture to be sufficiently apart from each other, and within the permissible parameter space. That the estimation procedure converges is an indication that the condition is fulfilled.

seemed a natural choice[15], while the variance of 25 was selected for consistency with the observed variance when the model with unconstrained component parameters was fitted in the largest sample size experiment. We observed that the results of the statistical analysis were not sensitive to moderate changes in the value of this variance.

The key aspect of the empirical investigation is whether the pattern of choices postulated by the iterated reasoning hypothesis shows up across the variety of groups (newspapers, students theorists). Specifically, we want to see if choices cluster around the theoretical choices of 33.33, 22.22, 14.81, …. Another aspect of the statistical analysis is to evaluate the weight across groups of each level of reasoning.

We consider two alternative model specifications, which differ according to the degree of invariance across groups of the model's parameters. The *unconstrained* model, in which the different component parameters, means and variances, are unconstrained across groups, and the *constrained* model, in which the component parameters are constrained to be equal across groups. In both analyses, the mixing proportions are allowed to vary across groups.

The technique we use is ML estimation of the mixture model using the expectation maximization (EM) algorithm (e.g., Dempster et al., 1977). The EM algorithm is a procedure that iterates estimation between z and π, i.e., given z, π is computed, and the other way round, till convergence is achieved. Estimation of the mixing proportions for each group is obtained from the mean, within groups, of the conditional probabilities of belonging to component *k*.

Note that the different analytic role of π and z in the likelihood function makes an EM algorithm the natural approach to our problem. In fact, the EM algorithm is found to be quite fast and stable in the problem at hand. Once convergence is attained, standard errors of the parameter estimates are obtained using the approximation to the information matrix as proposed in Louis (1982).

---

[15] Recall that in the E data set, choices were in [1,100]. And in FT many participants thought the range to be from [1,99]. In the remaining data sets choices were from [0,100].

7. Results of the statistical analysis

In Table 3 we show the results of the statistical method described above for $K = 4$. The first six rows describe the results of the unconstrained model, while the estimated parameters in the seventh row correspond to the constrained model.

	Number of Observations		Level-1 (33.33)	Level-2 (22.22)	Level-3 (14.81)	Level-Max (0/1)	Level-0
1. Lab (Experiments 1-5)	86	μ	35.31	23.60	14.53	1.00	
		σ^2	41.86	2.66	31.98	25.00	
		Proportion	37.88	18.01	9.99	0.00	33.99
2. Classroom (Experiments 8, 9)	138	μ	33.41	22.69	12.80	1.00	
		σ^2	0.88	12.31	6.07	25.00	
		Proportion	19.63	26.88	8.78	14.19	30.52
3. Take home (Experiments 6, 7)	119	μ	31.74	22.45	15.43	1.00	
		σ^2	16.94	0.94	12.68	25.00	
		Proportion	22.31	11.73	36.01	7.13	22.81
4. Theorists (Experiments 10-13)	146	μ	37.83	21.78	14.82	1.00	
		σ^2	89.54	3.19	3.50	25.00	
		Proportion	17.00	9.53	7.74	55.09	10.64
5. E-mail (Experiment 14)	150	μ	32.88	24.10	15.33	1.00	
		σ^2	0.69	10.57	7.90	25.00	
		Proportion	12.30	19.00	13.30	30.35	25.06
6. Newspapers (Experiment 15-17)	7889	μ	33.51	22.90	13.48	1.00	
		σ^2	0.88	16.19	8.32	25.00	
		Proportion	10.60	22.95	12.05	27.29	27.12
7. Constrained Model	8516	μ	33.45	22.56	13.48	1.00	
		σ^2	10.38	7.27	13.88	25.00	
		Proportion	18.23	17.50	18.12	24.82	21.32

Table 3. Results of the statistical analysis.

Sets of experiments with similar subject pools are aggregated in the groups described in the first column of the table. The second column shows the group sample sizes, n_g. We report means m_k and variances s_k^2, and the proportion variable which indicates the proportion, in each group, of the population with different levels of reasoning.

As an illustration, in Figure 6, we represent the four different normal distributions - obtained in the unconstrained analysis of Group 6 (Newspapers)- associated to the levels of reasoning 1, 2, 3 and Maximum.

Figure 6: The Components of the mixture for the Newspapers'group (exp.15-17).
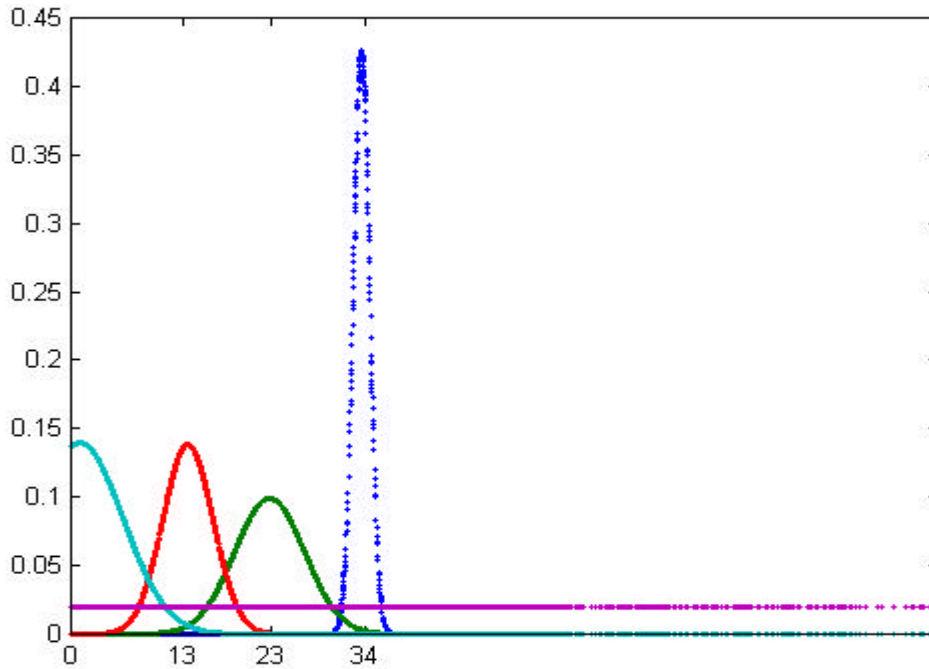
In addition, in Table 4, we present the log-likelihoods of the different estimations. It is interesting to note that the mean log-likelihood is similar for groups of experiments, an indication of similar goodness of fit among groups.

	Number of Observations	- log L	mean - log L
1. Lab (experiment1-5)	86	361.37	4.20
2. Class (experiments 8,9)	138	557.88	4.04
3. Take home (experiments 6,7)	119	478.82	4.02
4. Theorists (experiments 10-13)	146	544.19	3.73
5. E-mail (experiment 14)	150	593.88	3.96
6. Newspapers (experiments 15-17)	7889	32045.17	4.06
Constrained model	8528	30305.37	3.55

Table 4. Values of minus-log-likelihood of the estimated models

The first result we want to highlight is that all groups have means close to the theoretical predictions. The constrained model shows means of 33.45, 22.56, and 13.48, when the theoretical prediction puts them at 33.33, 22.22, and 14.81. If we turn to the unconstrained model and look in particular at the group of experiments with the largest sample size (Newspapers, row 6), the means estimated in the decomposition of the data, 33.51, 22.90, and 13.48, appear again quite

close to the theoretical predictions. Equally important, a comparison of the μ-values in the different groups shows little differences.

On the basis of the statistical analysis above, we can state the following facts:

Fact 8: Stability across very disparate experiments of the mean parameters of the component distributions.

Fact 9: Estimated mean values are close to the theoretical values predicted by the game-theoretical model of iterated reasoning.

One could be tempted to carry out a statistical test for the equality of the means of the composing distributions to the theoretical values predicted by the iterated reasoning hypothesis, i.e., values of 33.33, 22.22, 14.81, etc. To carry out such test, however, we need to face the issue of high heterogeneity of the sample sizes across groups. Such heterogeneity induces a high variation across groups of the power of a test for the same hypothesis. Take for example Group 6 (Newspapers' experiments) with a sample size of 7,889. If we test for the equality of the first-level mean to its theoretical value of 33.33, we obtain a t-value of 3.76 that clearly rejects the null hypothesis (p-value is smaller than 1%). In contrast, if we take a group like Classroom with sample size 138, the t-value turns out to be 0.30, which clearly leads to acceptance of the null hypothesis (p-value greater than any reasonable significance value). Therefore, in the large sample-size group we reject the hypothesis of the first-level mean being *equal* to the theoretical value, while we accept it in the small sample size group. But note that we reject the hypothesis of equality, the estimated mean differs from the theoretical prediction by .12, while we accept the equality hypothesis when the difference is 1.98. It is important to realize that the variation of the sample size that conditions the test, is an observation design artifact, nothing to do with the truth or falsity of the null hypothesis. Clearly, as in many settings, an adequate interpretation of the test statistics requires to perform a power sensitivity analysis, in order to distinguish between

26

statistical and substantive significance. This is an issue that we are currently investigating, but that, we feel, exceeds the scope of the present paper.[16]

With regard to the estimated *proportions* of the different reasoning levels, and especially the Maximum-level reasoning, we observe that they vary considerably across experiments, a confirmation of Fact 7 above[17]. This can be summarized as:

Fact 10: The proportions of subjects in each level of reasoning vary across experiments, i.e. it is highly determined by subject pools.

In particular, the frequency of equilibrium choices in the Newspapers experiments and in experiments with subjects trained in game theory is much higher than with un-trained undergraduates. This reflects the wide variation in the backgrounds of subject pools across experiments.[18]

The hypothesis that subjects have reasoning Levels 0, 1, 2, 3 or Maximum seems to account, as has been shown, for the choice patterns of a wide variety of very heterogeneous experiments. One may wonder whether the statistical analysis of the data would give a better fit if we assumed that no subjects had Level 0 reasoning, i.e., if our mixtures model did not include a uniform distribution. To assess the need for the uniform distribution in the mixture, we have performed the analysis without such non-normal component. The results show a significant deterioration of the fit of the model, as measured by the Schwarz information criterion. This

---

[16] Extending the simple t-test to the multivariate setting of the equality of the three- (four-) level means to their corresponding theoretical values would yield, again, results dependent on the sample size. Indeed, the corresponding Wald test for the equality of the three means to their theoretical values yields for the Newspapers and the Classroom groups the values 117.75 and 3.0084 respectively which, compared with a χ^2-distribution with 3 degrees of freedom, induce a clear rejection of the null hypothesis in the case of the Newspapers group and acceptance in the case of the classroom group (p-values are 0 and . 39 respectively).

[17] This is also true for the conditional proportions in the constrained model, that we did not print here.

[18] In fact, a formal test of the null hypothesis that the proportions are equal across groups, conditional that the component parameters (means and variances of each component) are equal, could be carried out. We take 2 times the difference of log-likelihoods for the constrained model and the model that assumes identical components and proportions across groups (this analysis would correspond to pooling all the cases in a single group). The test statistic equals 8734, which in relation to a χ^2-distribution with 25 degrees of freedom (the difference on the number of parameters estimated) clearly rejects the null hypothesis that all the proportions are equal across groups. Thus, the data clearly supports the Fact 10.

confirms that reasoning Level-0 should be included as a uniform distribution in the mixture model.

Another issue of interest is the testing of the *number* of normal components of the mixture model. A test statistic that takes care of the non-standard issue that the null value of the parameter is at the boundary of the parameter space can be constructed using bootstrap methods. But the power of such a test would be probably low given the large difference in sample sizes among the groups of experiments.[19]

6. Conclusions

Experimental results are influenced by what Marshak (1968) called the different costs of thinking, calculating, deciding and acting. Grand-scale experiments of the sort that can be run through a newspaper can test whether the results of the lab experiments change under variations in sample sizes and rewards but also when these different costs are modified. A population more heterogeneous than undergraduate subjects -as one is likely to encounter in a newspaper experiment- will include subjects with wide different costs of thinking and calculating (due to different education, training or information), different cost of deciding (at leisure vs. the time-constraint imposed in the lab), and different costs of acting (ready access to e-mail and fax or not). A richer world with less control.

That three experiments, run in different countries, for different newspapers, catering at different populations, yield results that are similar and, in two cases, indistinguishable from each other, is a clear indication that we are observing a pattern of behavior that must be quite common. That, in addition, the iterated reasoning observed with these large and diverse populations is also a pattern observed in lab experiments with subject pools of undergraduate, graduate students and economists, seem to clearly indicate that the "parallelism" assumption between lab and field has

---

[19] We repeated the previous estimation procedure for K = 5. The results indicate that the theoretical regularities are sustained (the mean of the fourth composing distribution is on average at 10.14, when the theoretical prediction would put it at 9.88). But the convergence problems for one group (Theorist) when we use the k=5 specification

been vindicated for the iterated reasoning hypothesis in Beauty-contest experiments. In fact, we are not aware that such a wide empirical corroboration of individual behavior has been previously observed, or even tried.

In addition, the paper shows that some basic patterns of behavior, specifically the individual depth of reasoning, are subject specific. This observation should warn experimentalists to apply with caution the results of their experiments to the wider world.

To sum up, the two main conclusions of the paper are:

Conclusion 1: Across subject pools, different sample sizes, and different methods of selecting the data, iterated reasoning patterns are stable and remain similar to the theoretical values predicted by the bounded rational model of iterated reasoning.

Conclusion 2: The proportions of subjects employing different levels of reasoning are subject-pool specific.

---

could also be taken as indication of overparameterization, at least in the case of small sample size groups.

References

Bosch-Domènech, Antoni and Rosemarie Nagel (1997a). "Cómo se le da la bolsa." *Expansión,* June 4, 1997, p. 40.

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

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

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

Bossaerts, Peter and Charles R. Plott (1999). "Basic Principles of Asset Pricing Theory: Evidence From Large Scale Experiments." California Institute of Technology, July.

Cooper, David, John Kagel, Wei Lo, and Qing Liang Gu (1999). "An Experimental Study of the Ratchet Effect: The Impact of Incentives, Context and Subject Sophistication on Behavior." *American Economic Review* (forthcoming).

Costa-Gomes, Miguel, Vincent Crawford, and Bruno Broseta (1999). "Cognition and Behavior in Normal-Form Games." working paper.

Dawes, Robin M. (1990). "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.

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

Haruvy, Ernan (1998). "Models in Beliefs." Mimeo University of Texas.

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

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

Lopéz, Rafael (1999). "The Beauty-Contest Integer-Game: A Theoretical Analysis." Universitat Pompeu Fabra. Mimeo.

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

Lucking-Reiley, David H. (1999). "Using Field Experiments to Test Equivalence between Auction Formats: Magic on the Internet." *American Economic Review*, forthcoming.

List, John A. and Lucking-Reiley, David H. (1999). "Demand Reduction in Multi-Unit Auctions: Evidence from a Sportscard Field Experiment." *American Economic Review*, forthcoming.

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

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

Nagel, Rosemarie (1998)."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), p.105-142.

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

Plott, Charles R. (1993). "Rational Individual Behavior in Markets and Social Choice Processes." Caltech Social Science Working Paper: 862.

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir (1991)."Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study." *American Economic Review*, 81, December, 1068-1095.

Schelling, Thomas. C. (1960). *The Strategy of Conflict.* Cambridge: Harvard University Press.

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

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

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

Stahl, Dahl O. and Ernan Haruvy (1998). "Empirical Tests of Equilibrium Selection Principles in Symmetric Normal-Form Games." working paper University of Texas.

Stahl, Dahl O and Paul W. Wilson (1994)."Experimental Evidence on Players' Models of Other Players." *Journal of Economic Behavior and Organization*, 25(3), 309-27.

Thaler, Richard (1997)."Giving Markets a Human Dimension." *Financial Times*, section Mastering Finance 6, June 16, 1997.

Appendix

Comment by participants in the S-experiment (translation from German into English):

I would like to submit the proposal of students of my math class Grade 8e of the Felix-Klein-Gymnasium Goettingen for your game: 0.0228623. How did this value come 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. To send one number for all, maybe one could take the average of all these numbers.

A first concern came from Ulfert, who stated that numbers greater than 66 2/3 had no chance to win. Sonja suggested to take 2/3 of the average. 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. 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 has to submit a very small number.

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.