Levels of reasoning between students: An experiment on level-k

Alejandro Medina Sandín

Co-advisor: Alessandro De Chiara
Co-advisor: Ester Manna

June 2022

Abstract

Bosch-Domenech et al. (2002) designed a very powerful model to study the levels of reasoning of subjects, and up to which point they can iterate their thinking. I adapted this framework to study the differences in reasoning between business students and economics students. I hypothesized about the degree of reasoning of each type of student and ran an experiment in my faculty. Subjects were asked to solve an adapted version of the Keynesian beauty contest game. I compared the responses of business students against the responses of economics students. No significant difference was found in their pattern of responses. However, there is a large number of observations of business students in the last level of reasoning of my experiment. This supports the hypothesis that business students solved the game by iterated dominance. I also provide a brief discussion about other factors that could drive different responses in the game such as gender, language of the class, game theory knowledge of the subject and nationality.

Keywords: Beliefs, level-k reasoning, behavioral economics, deepness of thinking, game theory, rational bounds, beauty contest.

1 Introduction

Level-k reasoning has been broadly studied in Behavioral Economics. Several games are trying to apply this framework in the literature, where the main objective is to understand the cognitive and reasoning process behind the choices of players. For instance, a K-level player (or infinite-level player) would be the one that has infinite intelligence and is perfectly rational, meaning that he would be able to anticipate the moves of all other players. However, this is not possible in human thinking (at least yet) since we have bounded rationality. Several papers try to model this type of behavior such as Alaoui and Penta (2016). This framework is contrasted with the concept of Nash Equilibrium. In fact, it proves that playing the Nash Equilibrium is not always the best strategy; seeking for it in these type of games will lead to the player losing. Nash Equilibrium is a fixed-point reasoning, such as a computer would proceed to solve games. Instead, most of the players deviate from the equilibrium, by anticipating that other players will play differently. Level-k reasoning tries to capture the degree of this "anticipation" and quantify it. More precisely, it captures the depth of thinking of each player. For instance, a level-0 subject would be an irrational player (usually it is assumed that he plays randomly), while a level-1 subject would assume that the rest of the players in the game are level 0, and he would act according to it. However, a level-2 player, would assume that the rest of the players in the game are level-1, and he would act according to it; and so on. More generally, a level-k player would assume that the rest of the players are level k-1. This theory also illustrates iterated elimination of dominated strategies. It consists in eliminating, step by step, dominated strategies that players would not play (assuming rationality) until no other dominated strategies are found. Level-k reasoning allows checking how many steps of elimination do players perform, and up to which point they stop. One of the most interesting things about this framework, is how it contradicts the assumption of rationality. A perfect rational player, would always be level-k, and would always play the Nash Equilibrium. However, it is proven that playing the Nash Equilibrium is not the best in this game. Bosch-Domenech et al. (2002) show that results with a wide sample of observations are far away from Nash Equilibrium (which is playing 0). In their original game, subjects have to choose a number between [0,100], where the winner is the one whose number is closer to 2/3 times the average of all numbers chosen. For example, let's take the case of 5 subjects. They each respond 10, 20, 40, 40 and 40. The sum is 150, and the average is 30; 2/3 of which is 20. The second subject would be the winner. Winning numbers range between [10,30] in the existing literature.

This assumption of common rationality is very important nowadays, not only in game theory, but also in other areas of economics. For example, core models of Industrial Organization are assuming common rationality. Let's take for instance the competition *A la Cournot*. In the simplest case, where two firms compete, they both know they are rational, moreover, they know their opponent is also rational. That's why they end up solving for the game with the best response function, as a function of their own quantity but also of their opponent's quantity. Another clear example of common rationality is the Hotelling model, applied to Political Eco-

nomics: The median voter theorem. Again, under the assumption of common rationality of all politicians, they compete to be closer to the median voter.

I explore the differences in the reasoning of different subjects in an academic environment. I contribute to the existing literature by testing the difference in reasoning between two types of students: business students and economics students. I expect them to reason differently because of their interests, especially those who have taken game theory. According to level-k reasoning, I expect to find significant differences in the responses of the two groups. I test 3 hypotheses regarding their way of responding to the game.

I designed an adapted version of the Keynesian Beauty contest (see Bosch-Domenech et al. (2002)) for undergraduate students. I randomly took students from *Universitat de Barcelona* from different years and lessons. I showed up in some classes, and they were asked to voluntarily participate in the experiment. They faced the following game:

- In this game you must choose a number with up to three decimals between 0 and 100. Your objective is to get as close as possible to 1/2 times the average of the numbers chosen by all the students in your classroom.

I also collected demographic data, such as gender, age and if they have Spanish nationality. There is no "correct" answer in this game, since it depends on the subjects playing. Results show that there is no statistically significant difference between the responses of two groups of students, their level of reasoning is not statistically different. I also found that a notable percentage of business students arrive at the last level of reasoning of the game (close to 0). I explored differences in other characteristics such as gender, language, nationality and if the subject had previous knowledge of game theory. None of these characteristics seemed to be a key factor in the responses of the subjects.

2 Motivation

Usually, students are taught how to mathematically solve a problem, how to find the equilibrium of a game, or to arrive at the steady-state of a macroeconomics model. However, sometimes teaching does not involve this degree of reasoning that should be expected, especially in economics field. Students rely on the Nash Equilibrium concept to solve games, in fact, is one of the first concepts to study in Game Theory. Nash Equilibrium is a very strong concept based on the common knowledge of rationality among players. But, in certain environments:

- Players are not fully rational
- Players might not be able to anticipate that other players are not fully rational
- Players might not be able to anticipate that other players are fully rational

Or, in other words, players might have difficulties in correctly guessing the type of rationality of other players. Here it arises the problem of overconfidence, in which agents have an overly positive view about himself or their skills (see Camerer and Lovallo (1999) or De la Rosa (2011)). Another interesting concept that arises in these types of games is the deviation from the Nash Equilibrium. Because I expect that subjects in my experiment will not be perfectly rational, they will probably deviate from Nash Equilibrium. As I said before, deviating from the Nash Equilibrium does not mean losing the game. In fact, several papers study deviations from the equilibrium, especially those focusing on the *Perfect Bayesian Nash Equilibrium* and sending signals to rule out the "bad" equilibrium (see Fudenberg and Tirole (1991)).

2.1 Objectives

I want to shed light on why students think in diverse ways, and how can we improve the quality of teaching such that they learn to reason, instead of mechanically thinking as computers. I want to check how diverse subjects in the academic environment reason differently. I also want to show if there are significant differences in the reasoning of subjects based on their studies, gender, teaching language and knowledge of game theory.

2.2 Structure of the Thesis

The thesis continues as follows: Chapter 3 provides a brief analysis of the literature of the game and the concepts that might arise. Chapter 4 explains the experimental design and how I ran the experiment. Chapter 5 states the hypotheses and how I planned to test each one. Chapter 6 states the results and statistical tests. Chapter 7 discusses results that arise as a function of other variables, Chapter 8 discusses a bit how further research could be conducted in this topic, and Chapter 9 concludes.

3 Previous literature

The main paper I base my experiment on is *One, two,(three), infinity,...: Newspaper and lab beauty-contest experiment* (Bosch-Domenech et al. (2002)). This is the first published paper that analyzes different levels of reasoning in subjects. Not only they check for levels of reasoning, but also for the process, divided into five categories. They ran the so-called Keynesian Beauty Contest in an experiment in three different journals of different countries. They observed a common pattern of answer in the three journals, with very similar winning numbers (13, 14.7 and 16.99). Another seminal paper of this theory is Stahl and Wilson (1995), where the authors developed a 3x3 symmetric game to test human behavior and classify it in one of five categories. Ho et al. (1998), Costa-Gomes and Crawford (2006) and Crawford and Iriberri (2007) are other papers where researchers stressed out the level-k framework, among others developing beauty contest models in related settings. Level-k tries to classify subjects in levels of reasoning. A level-0 subject corresponds to a random, non-strategic behavior. A level-1 subject has the belief that all other players are level-0, and best respond to this belief. A level-2 subject has the belief that all other players are level-1, and best respond to this belief, and so on. Most of these papers show a common pattern in responses, as a function of the levels of reasoning of the subjects playing the game. Agranov et al. (2015) and Burchardi and Penczynski (2014) found that, in their experiments, a large number of subjects were level 0. In fact, they found that, applied to their own experimental design, level-0 beliefs responses are not significantly different from 50. They also state were able to disentangle level-0 beliefs from particular assumptions or belief distributions. Further studies have been run within this framework to try to understand level-k behavior. Ye Jin proposes a strategy to identify ability-bounded subjects, who cannot reason more than a certain finite number of k steps, and therefore arrive to their maximum capacity (see Jin (2021)). Another game using level-k reasoning is the so-called *11-20 game* (Arad and Rubinstein (2012)). In this two player game, each subject has to choose to receive an amount of shekels between 11 and 20; but if he chooses exactly 1 shekel less than the other player, he will receive an additional 20 shekels bonus as an addition to the number of shekels chosen. With this game and variations of it, Rubinstein and Arad showed that subjects do not usually use more than three steps of reasoning.

Even though level-k framework might seem a simple concept, with some twists it helps to understand decision processes. For example, Shapiro et al. (2014) developed a theoretical model within the level-k framework to study how its predictive power changes when modifying two important factors of the came: Coordination and information symmetry. They found that introducing private information, or introducing coordination mechanisms, weakens level-k behavior in subjects. Moreover, Alaoui and Penta (2013) also developed a model that allows them to identify the levels of reasoning of a subject as a function of their own cognitive constraints and the beliefs of the cognitive constraints of their opponents. They also study how individuals change their behavior as a function of incentives.

It is also noteworthy to cite some psychology papers, trying to study the frontiers of human

thinking. Kinderman, Dubnar and Bentall show that most subjects do not understand long and complex sentences such as: Did X think that Y thought that Z thought...? (see Kinderman et al. (1998)). This directly relates to our level-k thinking framework, and how subjects face difficulties in arriving to further levels of reasoning. A paper from Whalen, Zunshine and Holquist shows how the reader faces difficulties in reading sentences with high numbers of embedments, that increase reading time and reduce the understanding of the reader. For example, the sentence "She felt she knew too much about Rudi to respect him" requires at least 3 levels of reasoning, which sometimes might be hard in literature (see Whalen et al. (2015)).

On the other hand, there are several articles with the idea to model human thinking using tools from economics. Alaoui and Penta introduce a model in which they can be able to differentiate the depth of reasoning of subjects and their cognitive bound as a function of his beliefs about the other player (Alaoui and Penta (2016)). They also studied jointly with Janezic up to which extent level-k behavior is affected by beliefs of subjects or by their cognitive bounds (see Alaoui et al. (2020)). Capra developed a model to try to understand the decision process of a subject in guessing games. She classified types of processes as a function of their elicited cognitive answers in the game (see Capra (2019)).

4 Experimental design

This chapter explains the experimental design with some discussion about the incentive problem and how I ran the experiment in my university, and then states the hypotheses and analyzes how each one is going to be tested.

4.1 The game

I ran an adapted version of the Keynesian Beauty Contest. Subjects were asked to choose a number with up to three decimal points, between 0 and 100 (both included). Their objective was to get as close as possible to 1/2 times the average of all the students in the class. 1/2 was chosen instead of the typical 2/3 because I thought that students would find it easier to do computations. Then I also asked for some demographic data of each student. Participation was voluntary. I chose students from the undergraduate degrees of business and economics from the *Universitat de Barcelona* from different subjects. A detailed table of the sociodemographic characteristics of students is available in section 4.2.

4.1.1 The incentive problem

One of the biggest challenges I faced developing the experiment was how to properly incentivize students to develop the task, and put some effort on it. In the Behavioral Economics field, it is well known to use money in experiments as an incentive for subjects to play. Usually, subjects can get a fixed amount of money plus a bonus monetary payoff that depends on their outcome in the game. In a Master Thesis, this is not possible, especially without any funding. Therefore, I needed a way to induce some sort of effort by students without providing them any monetary reward (nor rewards such as an increase in the final grade, or vouchers, which were not possible either).

The cleanest solution I arrived, was the following: If I wanted students to make an effort, the best way would be to give them the opportunity to be seen as smart in class. Therefore, I told students at the beginning of the experiment that if they win, I (or the corresponding professor) would publicly announce their name in the following lessons. Now, a collateral effect might be that the student is shy, and does not want to have his or her name publicly announced in class; so I also provided the option to be informed privately (by email). The experiment presented to students is available in the appendix.

The key point here is that, if students knew that the reward was this kind of "joy" of being smarter than the rest of the class, they would do the effort. Of course, this is not an infallible tool, but it was the best I could do with the resources of a Master's student.

4.1.2 A brief discussion: Private or public reward?

However, I have to mention that this procedure might generate some problems. First, the self-esteem of each student will probably affect the decision of being informed publicly or privately.

While the main point of my incentive scheme is to let the class know the winner, the subject might be shy, and do not want their name to be called publicly in class. Therefore, my incentive design loses power. If this kind of subjects do not want to be called publicly in class, they will not make an effort. In fact, this is a commonly studied phenomenon in psychologic research. For instance, Paz et al. (2017) shows how self-esteem can affect the results in economic tasks such as the ultimatum game. One of their main results is that low self-esteem women reported more anger towards unfair offers with respect to high-esteem women.
On the other hand, it might also generate an overconfidence problem. Overconfident students might have a wrong belief, both about their own type and the type of the rest of the players. They might think they are smarter than the rest, and the reward of publicly announcing their names in class might backfire, since they would follow a different strategy because they think they are smarter than they actually are. De la Rosa (2011) shows that in a second-best contract in the principal-agent theory, the power of incentives are decreasing in overconfidence.
But there is also evidence (in other areas) supporting that public incentives are better than private or instrumental incentives, such as money. For example, Handgraaf et al. (2013), shows that in an electric company, social rewards such as grade points with a description were outperforming private rewards in form of money. Pay-for-performance is also studied in the medical literature, especially in the context of the United States, where access to health services is limited and costly. Articles such as Himmelstein et al. (2014) or Hemenway et al. (1990) show how monetary incentives regarding the performance of subjects might actually worsen their performances or to increase their practice without any real benefit for the client.
Letting students choose how they wanted to be informed if they win the experiment was a matter of ethics. I did not want anybody to feel uncomfortable about his or her name being called publicly in class, that is why I also gave them the option to be informed privately. In fact, only 3 winning subjects chose to be informed publicly, while 5 wanted to be informed privately (there was a tie in a session).

4.2 Descriptive statistics

The experiment was conducted entirely in the *Universitat de Barcelona*. Firstly, we asked for permission from the *Cap d' estudis* (Director of the studies) of the faculty to use faculty students as subjects. I talked with professors from the Faculty of Economics and Business, and asked them for permission to perform the experiment in their classes. During April and May 2022 I conducted several sessions of the experiment. I went into the class and briefly presented myself and told them I was doing an experiment for my thesis. Students were allowed to use calculators, but they were not allowed to comment on the game among them. I gave them a paper with the experiment. They were asked about demographic data. Also, they were asked if they would like to receive any information about the outcome of the experiment. I was able to collect 227 observations out of 7 sessions; 135 of those belonging to economics students, while 95 were business students. The following tables report the statistics about the experiment:

Table 1: Descriptive statistics of the experiment

Session	1	2	3
Degree	Business	Economics	Business
Language	English	English	Spanish
Class	Industrial Organization	Microeconomics II	Industrial Organization
Year of the degree	Third	Second	Third
Winner has previous knowledge of Game Theory?	Not asked	No	No
Winner has Spanish Nationality?	Yes	Yes	Yes
How does the winner want to be informed?	Publicly	Privately	Privately
Number of observations	46	21	26
Result of the computation	14.477	7.958	12.665
Winning number	14.446	7.5	12

Table 2: Descriptive statistics of the experiment

Session	4	5	6	7
Degree	Business	Economics	Economics	Economics
Language	Spanish	Spanish	Spanish	Spanish
Class	Industrial Organization	Microeconomics II	Microeconomics II	Game Theory
Year of the degree	Third	Second	Second	Second/Third
Winner has previous knowledge of Game Theory?	No	No	No	Yes
Winner has Spanish Nationality?	Yes	Yes	Yes	Yes
How does the winner want to be informed?	Privately	Public	Private/Public [1]	Privately
Number of observations	23	44	32	35
Result of the computation	12.266	13.646	12.757	11.24
Winning number	12.5	12.5	13	11.2

[1] In this session there was a tie

5 Hypotheses

Because I expect students to think in diverse ways, hypotheses are formulated in such a way that I can statistically test if there is a difference between the two groups.

To classify the diverse types of thinking of students, I follow a similar procedure as in Bosch-Domenech et al. (2002). The first type of reasoning is **iterated elimination of weakly dominated strategies** (i.e. Iterated Dominance). This type of reasoning consists of using backward induction to solve the game: a rational player would never choose any number above $100 * p$, in my game, above 50. Why? If the maximum number possible to be chosen is 100, all numbers above $100 * 1/2$ are weakly dominated by 50. But if he believes that the other players are rational, he is going to choose numbers below $100 * p^2$ (25 in my game), and so on until arriving to the unique Nash Equilibrium of the game which is 0. This game is ideal to study how many steps individuals perform while thinking, and how much they iterate this process.

However, there is also another way of solving these Beauty-contest games, which is the **Iterated Best Reply of (non)degenerate beliefs**. In this case, every player generates a belief about the level of reasoning of other players and has the belief that he is one step further than the rest of the players, or in other words, he has the belief that he is smarter than the rest of the players. This might be easier to understand with an example. If a player has the belief that the rest of the players are of Level-0, he will be Level-1, and therefore he will randomly choose a number below $50 * p$, with the mean of the interval [0,100] being 50. However, if a player has the belief that the rest of the players are of Level-2, he will be Level-3, therefore he will choose $50 * p^3$. Particularly, in my experiment, a player that is following Iterated Best Reply of degenerate beliefs would play:

Table 3: Strategies played by each level-k player

Level-0	Level-1	Level-2	Level-3
Random number $\in [0, 100]$	25	12.5	6.25

Level-4	Level-5	Level-6	Level-k
3.125	1.562	0.781	$50 * p^k$

The only difference that the authors attribute between non-degenerate and degenerate beliefs is that by degenerate beliefs we understand that the player gives probability 1 to all the other players being at a certain level of reasoning, while in non-degenerate beliefs the player assigns a positive probability to the other players being at more than one level of reasoning. I am not going to use that difference, and I will classify them in the same group. From now on, I will refer to both of these groups as Iterated Best Reply of degenerate beliefs (IBRd).

Note that, the difference between iterated best reply and iterated dominance is that in the iterated best reply the starting point is 50, instead of 100.

The two hypotheses are as it follows:

- Hypothesis 1: Economics students will follow iterated best reply of degenerate beliefs and iterated best reply of non-degenerate beliefs. (i.e., their level of reasoning is around 2-3 levels of deepness)
- Hypothesis 2: Business students will arrive to level-k infinite reasoning, driven by solving the game by iterated dominance. Otherwise, they will be level-0 players.

So, the key question here is: Why did I choose economics students to be following iterated best reply of degenerate beliefs? Economics students are more familiar with concepts regarding game theory. Some of them have taken optional courses in game theory, moreover, they take more Microeconomics and Industrial Organization courses with respect to business students, which often imply this point of "strategical thinking". On the other hand, business students are not used to these types of games, in fact, most of them only take up to two courses in Microeconomics, and no courses in Industrial Organization or Game Theory. I also expect them to have different preferences and interests; for instance, economics students might be more interested in these kinds of topics regarding game theory than business students, that they usually are more interested in topics such as human resources or marketing. Not only that, but due to how I gathered the data, there will also be differences within the data set. Students are diverse in their game theory knowledge, the year of the degree and, of course, the background of subjects they have.
That's why I formulated my hypotheses like this. While business students are expected to mechanically solve the game (i.e., iterated dominance); economics students are expected to be able to anticipate the Nash Equilibrium of the game, and iterate some levels of thinking using best reply of degenerate beliefs. Business students are expected to either be level 0 (randomly choosing any number), or infinite level (solving the game by backward induction). I expect this group of students to play mostly the Nash Equilibrium of the game, which is 0, or numbers close to it; and if not, I expect them to just randomize. Economics students are expected to iterate some levels of reasoning, but finitely (2-3 levels of deepness). In other words, I think economics students will play the winning numbers in existing literature, which are usually between [10,30]; as a result of forming beliefs about their colleagues' reasoning. I believe them to iterate up to two or three levels of reasoning and then stop.
To test these hypotheses, I am going to use some statistical tests and graphs to show the differences between the two groups. The two first hypotheses will be tested together, using the Mann-Whitney U test. This is an easy non-parametric test to compare two independent samples, where the null hypotheses is that the means of both populations are equal. So the test will compare the results of the two groups, business and economics, and tell if there are significant differences in their responses. Not only I am going to test if there are differences in the answers of the subjects by their degree, but also by their gender, language of the class, if they have Spanish nationality or not and if they had previous game theory knowledge.
However, even though this might provide me some initial idea of the direction of my hypotheses,

it is not enough to test them, since I want to check if each group of students follows the reasoning process already mentioned. I am also going to group the observations in seven different processes of reasoning. In other words, I classify observations from level-0 to level-k (which in this case would be level-7), and create some histograms and graphs to see the distribution of answers. The hypotheses will be true if both the statistical tests reject the null hypotheses of the distributions of both populations being identical and the graphs showing a clear pattern of responses. The graphs should show a majority of responses from little numbers of business students, and the majority of responses of economics students being concentrated in the range [10,30].
Moreover, I also provide the graphs of cumulative frequencies of the responses of different groups, to help me see the differences in the answers.

6 Results

This section states the results of each test with an explanation. First, let's start with the statistical tests.

6.1 Discussing the problem with Mann-Whitney U test

Before going into deep detail about the statistical result, I have to discuss a well-known problem with the U test: the ties. Ties can modify the real value of this test, because, when doing the calculations, if ties are found they are canceled out.
Since I do not have a large number of observations, this might be an issue for me, because if a lot of observations are canceled out, the test would lose a lot of predictive power. Before running the test, I computed how many ties are in each group. I classified the ties by pairs of variables, since the Mann-Whitney test will compare those to run the analysis (for example, I counted the number of ties between the responses of males and responses of females). Ties are reported in the following table:

Table 4: Ties found per each pair of variables

Pair of Variables	Number of ties
Gender (Male/Female)	4
Game theory knowledge (Yes/No)	0
Language (Spanish/English)	0
Nationality (Spanish/Other)	4
Total	8

Because there is not a large number of ties by pairs (only a little number in gender and nationality), there will be no problem in performing the Mann-Whitney U test.

6.2 Studying the differences in degree: The Mann-Whitney U test

Because I do not expect the distribution of responses to be normal, I cannot use the t-test. The first step is to use the Mann-Whitney U test (sometimes called Wilcoxon test, or simply U test). This test allows me to distinguish if the mean of both populations is equal or not. The test is distributed as:

$$\text{H=} \begin{cases} H0 : \text{Mean of numbers in economics} = \text{Mean of numbers in business} \\ H1 : \text{Mean of numbers in economics} \neq \text{Mean of numbers in economics} \end{cases}$$

This is related to the first two hypotheses, and, to both of the hypotheses to be true, I would need to reject the null hypotheses of equality when distinguishing by degree.
Because the p-value > 0.05 (p-value $= 0.3809$), I have to reject the null hypothesis in favor of the alternative hypothesis. This means that both populations are not statistically different, and I would have to reject my first two hypotheses. Why? This test is basically showing that there is no significant difference between the responses of economics students and business students, therefore, their answers are very similar.
Since I was very surprised with the response of this test, I decided to perform another one called the Kolmogorov-Smirnov test. This is a very similar test that instead of comparing means, it tests the equality of distributions of two samples. In this case the p-value is 0.371 so, again, I cannot reject the null hypothesis of no difference between the two distributions.
However, this is not the end of the analysis. Let's check the distribution of responses by degree in a histogram.

6.3 Studying the differences in degree: Histogram

Even though the analysis of the statistical test points towards a similar distribution of the responses, there is a different pattern observed in the histogram by degree:

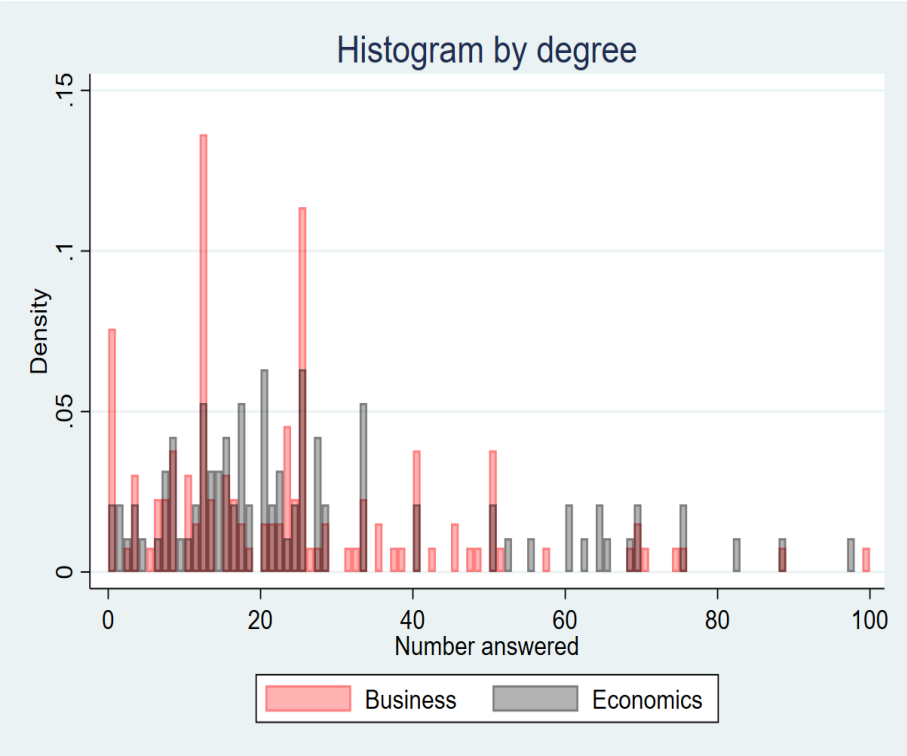

Figure 1: Histogram by degree of students

The analysis is pretty interesting at this point. Mann-Whitney U test did not allow me to reject the null hypothesis of similarity in means, however, there is a notable contrast in the histogram of responses. While the business students distribution follows a flatter distribution,

we observe some higher spikes of economics students in lower numbers. Particularly, there are spikes around the 0 and in the numbers 15 and 25; however, the answers of business students are more distributed.

Now, this generates some doubts about the validity of using the Mann-Whitney U test (even though I checked and discussed the ties problem). Because I still have some doubts about the main results, I am going to do the most basic analysis: Divide the responses in levels of reasoning.

6.4 Studying the differences in degree: Levels of reasoning

The following table shows the relative frequency of levels of reasoning by degree:

Table 5: Relative frequency of responses by level of reasoning

Level of reasoning	Economics	Business
Level 0 (>50)	19%	11%
Level 1 ([25,50])	20%	29%
Level 2 ([12.5,25])	39%	31%
Level 3 ([6.25,12.5])	14%	15%
Level 4 ([3.125,6.25])	3%	5%
Level 5 ([1.562,3.125])	1%	1%
Level 6 ([0.781,4.562])	2%	0%
Level 7 (<0.781)	2%	10%

I stopped classifying at level 7 since it is not worth it to divide more levels in such low numbers. While most of the levels of reasoning follow a similar pattern, there are two notable differences: In level 1, and in level 7.

I expected economics students to iterate 2 or 3 levels of reasoning, while business students to either be level-0, or level-k (in this case level-7). With some differences, the table shows that most of the observations are grouped up in levels from 1 to 3. In fact, this is consistent with existing literature. This is not consistent with my two first hypotheses: what we can see in this table is that both groups of students respond similarly, even though some differences.

6.5 An additional result: Level-k business students

However, another important thing to comment is the level-7 in the business group. There is a notable number of observations in the last level of reasoning (10) compared to the number of economics.

This is a notable fact. Not all the business students resolve the game by iterated dominance, in fact, there is not a large amount of observations around the 0 (in both groups), but 10 students

in business are at the last level of reasoning, compared with 2 in economics. I think that, with a bigger number of observations, this could give a clearer result about differences in the levels of reasoning in both groups.

If we group the observations from level-4 to level-7, we see that, while economics students only represent a 8% of the total, business students represent 16%. This additional result would support my first two hypotheses. While economics students are mostly grouped in levels 2-3, business students have two "peaks" in their distributions, at the beginning, level 0-1, and at the end, level-7. In other words, while economics students only have one peak, business students have two; they are different in the tails of the distribution. Recall that my second hypothesis was expecting business students to either be at one extreme or at the other of all the levels of reasoning.

6.6 Studying the differences in degree: Cumulative frequency

To shed more light about the distribution of responses in both degrees, I computed the table of cumulative frequencies distinguishing by degree:

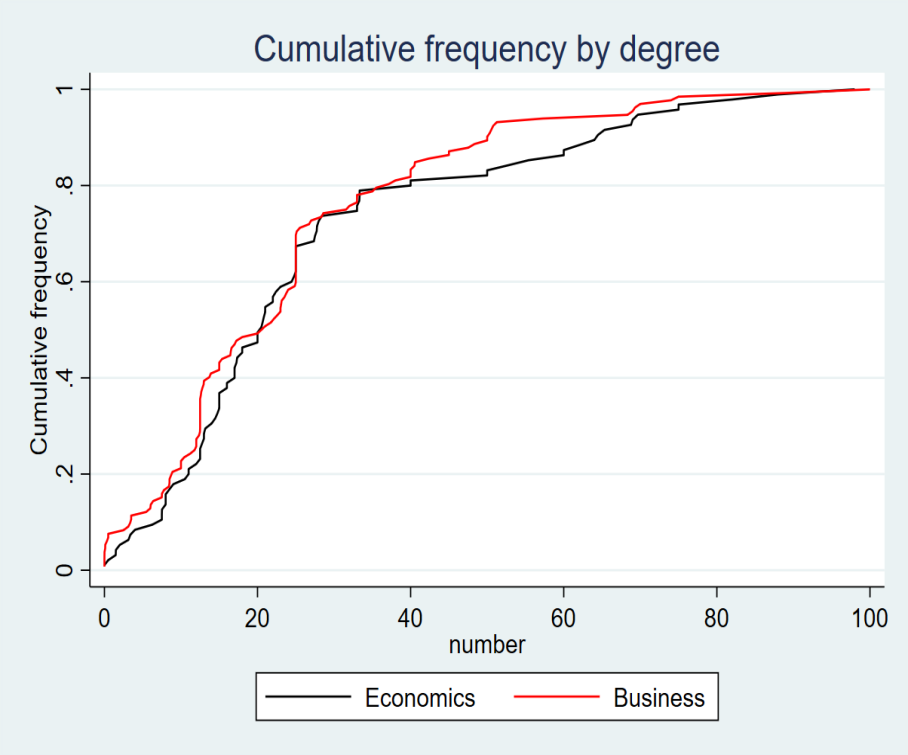

Figure 2: Cumulative frequency by degree of students

Again, and following similar patterns observed in the previous analysis, there are no clear differences in the cumulative frequencies of both groups. This reinforces the fact that I cannot accept the first two hypotheses.

6.7 Reasoning about students reasoning

Before discussing the hypotheses, I wanted to do one last analysis of the difference of responses by degree. Because level-k framework allows me to check for the steps of reasoning of each subject, I wanted to count how many observations are **exactly** in each step. This is, count how many responses are 50, 25, 12.5 and so on. The following table states the number of responses by exact level of reasoning:

Table 6: Exact responses

Number	Economics	Business
50	0	4
25	5	14
12.5	2	10
6.25	1	0
3.125	0	0
1.562	0	0
0.781	0	0

Against what I expected, this table shows that a higher number of business students are performing this kind of procedure. In fact, they are stopping at most of 3 levels of reasoning (which is consistent with existing literature) while economics students are not following this procedure. This again makes me think my two first hypotheses are not correct. I cannot state what were students thinking when playing this game. but a large number of responses in 25 and 12.5 suggests that students would be following backward induction, but up to a certain point.

6.8 Business against Economics: the game

Another interesting analysis to do is to perform the game with all the economics subjects and all the business subjects. The following table shows the result when grouping observations from all sessions in only one game:

Table 7: Game played with all subjects divided by degree

Degree	Economics	Business
Result of the computation	13.868	11.885
Winning number	14	11.78

Of course, results are pretty similar, and this supports the fact that I will not be able to confirm the two hypotheses I developed. Further discussion is provided in the results Chapter.

6.9 Discussing the results and hypotheses

Let's recall the two main hypotheses of my study:

- Hypothesis 1: Economics students will follow iterated best reply of degenerate beliefs and iterated best reply of non-degenerate beliefs. (i.e., their level of reasoning is around 2-3 levels of deepness)
- Hypothesis 2: Business students will arrive to level-k infinite reasoning, driven by solving the game by iterated dominance. Otherwise, they will be level-0 players.

The hypotheses are not accepted. In fact, in the first step of the analysis, the statistical tests (Mann-Whitney and Kolmogorov-Smirnov) are already pointing towards a similarity of distribution in the answers of both groups. Moreover, the relative frequencies table and cumulative frequencies graph support this idea. If we inspect the histogram, we see a different pattern in the responses, especially in the answers of the economics students, where we see some spikes in high levels of reasoning. However, this is not enough to statistically reject the null hypotheses of similarity in means and similarity of distributions, so, as explained above, I have to reject the first two hypotheses.
On the other hand, I expected business students to solve the game by iterated dominance since they would not be able to form beliefs about their opponents. In fact, 10% of business students are in the last level of reasoning (very close to 0), while only 2% of economics students arrived to this level. With this data, I cannot confirm my belief, but, there is some initial evidence regarding this fact. I leave for future research to run this experiment with higher observations, that could give more predictive power and therefore analyze the hypotheses more consistently.

7 Further results

Even though I already discussed the main hypotheses of the study, I collected a data set that allows me to analyze further differences between subjects. This section discusses those differences and if they might be affecting the responses of students.

7.1 A simple regression

The first step is to run a simple regression. I advance that this regression has multicollinearity problems, and it just serves for the purpose of checking how the variables correlate between them. I regressed the variable number (all the responses), against all the demographic variables.

	(1) number
degree	-1.853
	(-0.51)
gender	-5.752
	(-1.87)
gametheoryknowledge	-2.041
	(-0.57)
language	-11.28**
	(-3.32)
nationality	-5.221
	(-1.11)
constant	34.66***
	(7.05)

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: Regression of number answered against all variables

Only the variables of interest language and the constant are significant. Those who take the classes in English would, on average ceteris paribus, report a lower number. The dummy variable for gender takes value 1 for males and 0 for females, the dummy for the degree is equal to 1 for economics and 0 for business. The variable *gametheoryknowledge* is a dummy that takes value equal to 1 if the subject had prior knowledge of game theory, and 0 otherwise. The language variable takes value equal to 1 if the class taught in English and 0 in Spanish. Lastly, the dummy variable for nationality takes value equal to 1 if the subject has Spanish nationality, and 0 otherwise[2].

7.2 Analyzing the other variables: Gender

As I said before, the data collected allows me to also compare the responses in function of several variables, let's start with the gender.

[2]STATA do-file is available under request

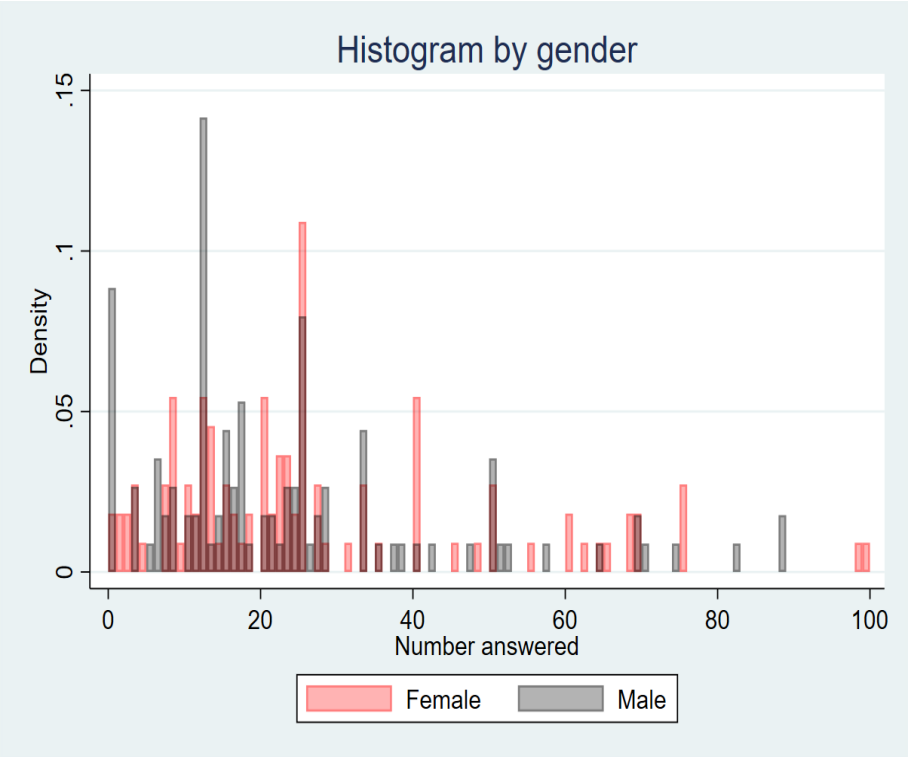

Figure 3: Histogram of responses by gender of subjects

The histogram shows a clear difference between the responses of males and females. While males tend to be more concentrated on low numbers, females are more distributed along the histogram. Their spike is at 25. Another very notable characteristic of the histogram is that there is a large spike of males in 0. The data set is pretty balanced (110 females and 113 males). However, 82 males were from business, and, since we found that a large number of business students were level-7, this could explain the spike at 0.

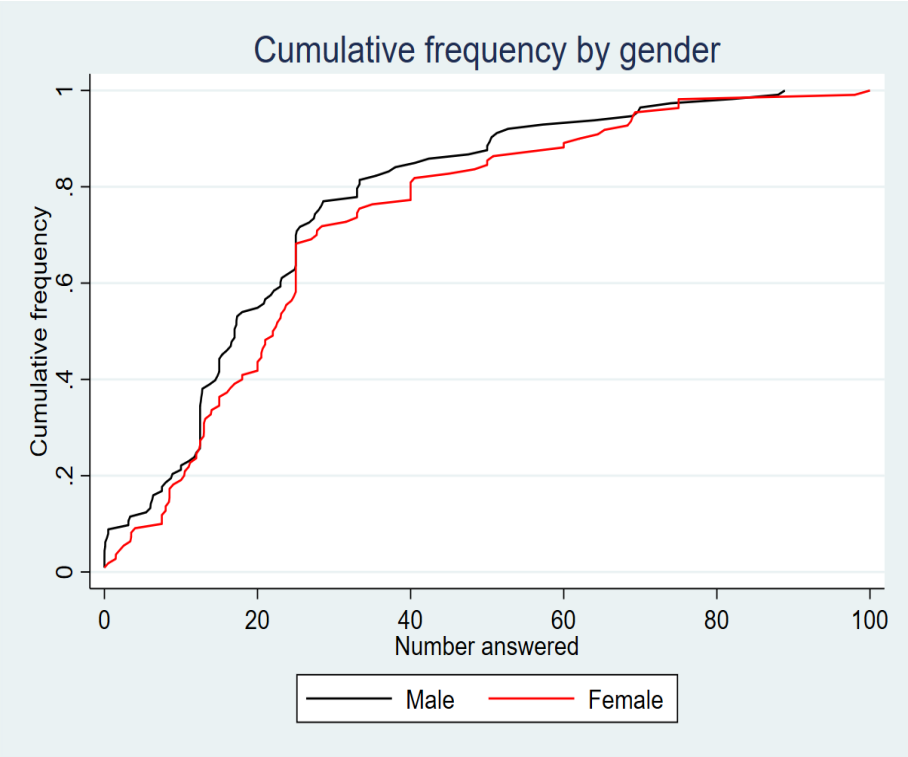

Figure 4: Cumulative frequency by gender of subjects

With the cumulative frequency graph we can observe that there are no large differences in their responses. Again, I computed both the Mann-Whitney (p-value=0.1913) test and the Kolmogorov-Smirnov (p-value=0.224) test, and I could not reject the null hypotheses of similarity in none of them.

7.3 Analyzing the other variables: Previous game theory knowledge

The hypotheses of this thesis arose when I was thinking if there would be any significant difference of subjects playing this game depending on whether they would have prior knowledge of game theory or not. This section analyzes it. However, this is subject to the limits of my data set, I was only able to gather data for 52 subjects who had previous game theory knowledge, while the rest did not.

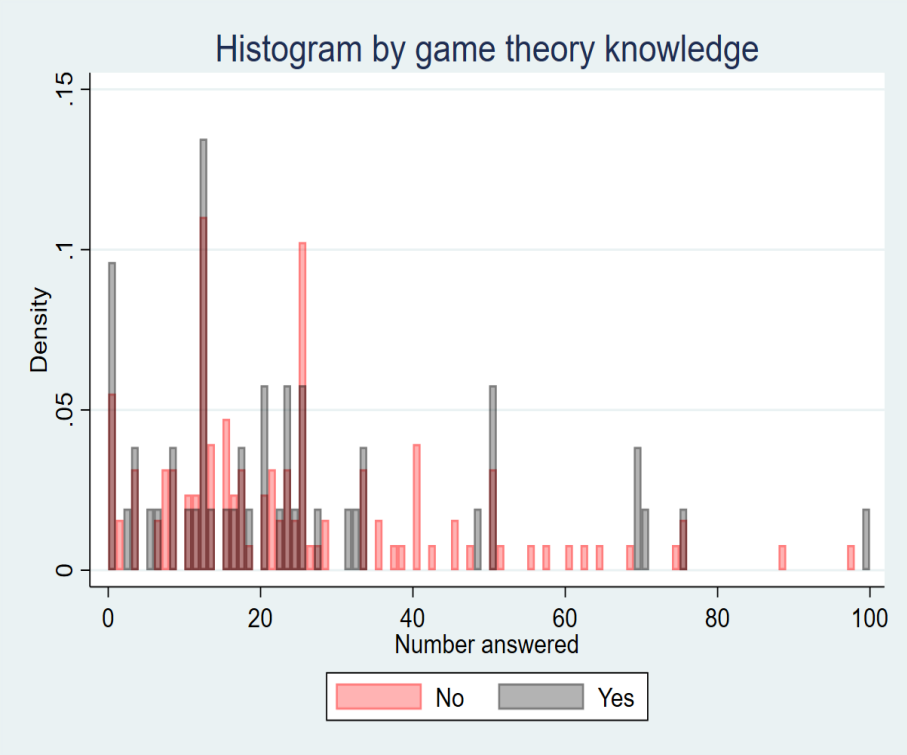

Figure 5: Histogram differentiating if subjects had prior game theory knowledge or not

In this case, we observe that, even though the cumulative frequency graph exhibits no difference, the pattern of the histogram shows clearly how most of the subjects with previous game theory knowledge are distributed in the lower numbers of the histogram. This is consistent with my initial intuition that subjects with previous knowledge of game theory would have a better performance in this game. Again, both of the statistical test point in the same direction: no difference in means and no difference in distributions (Mann-Whitney p-value=0.5465 and Kolmogorov-Smirnov p-value=0.965).

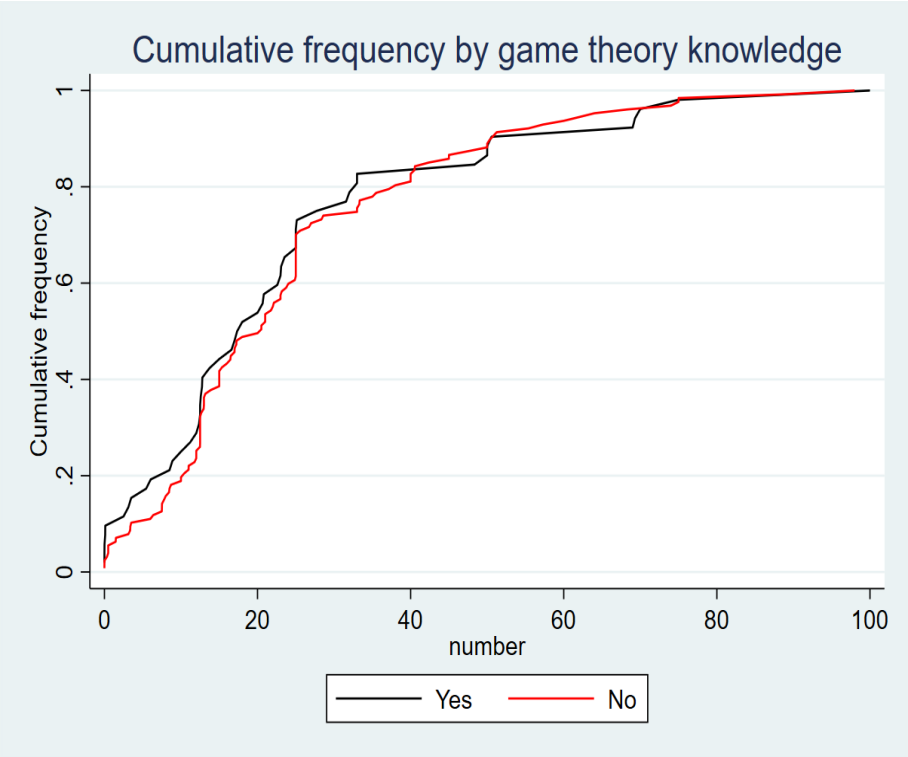

Figure 6: Cumulative frequency differentiating if subjects had prior game theory knowledge or not

7.4 Analyzing the other variables: Language

Now let's check if there is some significant difference of answers depending on the language of the class (English or Spanish).

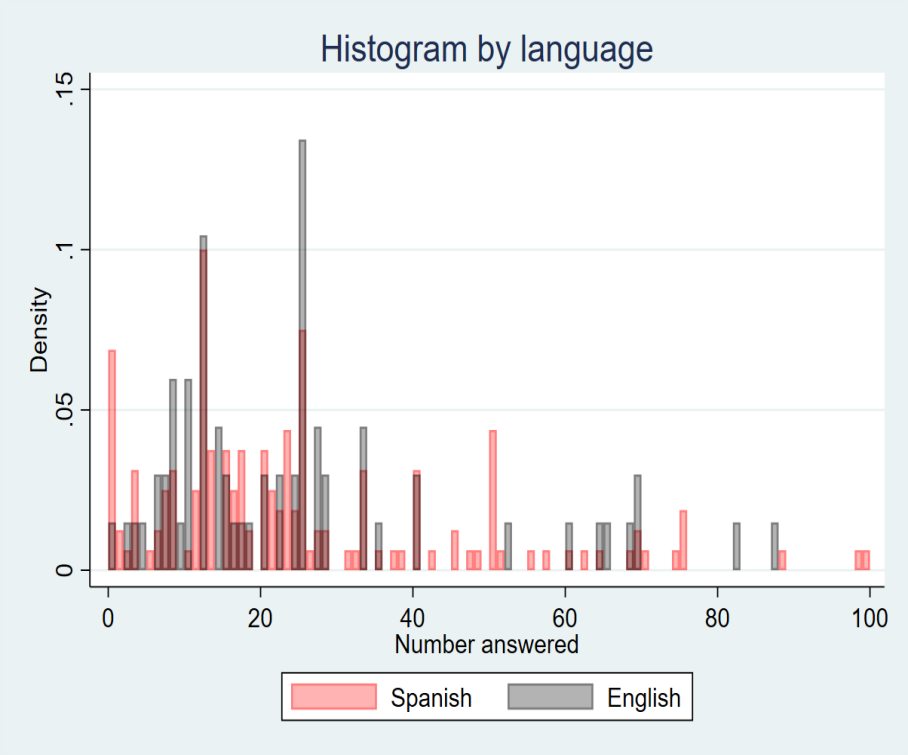

Figure 7: Histogram by language of the class

The histogram shows a spike of Spanish students at 0, while English students are more concentrated in the spike of 25. The cumulative frequency graph again shows no significant differences, as well as both statistical tests since I cannot reject the null hypothesis of similarity. (Mann-Whitney p-value=0.9493 and Kolmogorov-Smirnov p-value=0.904).

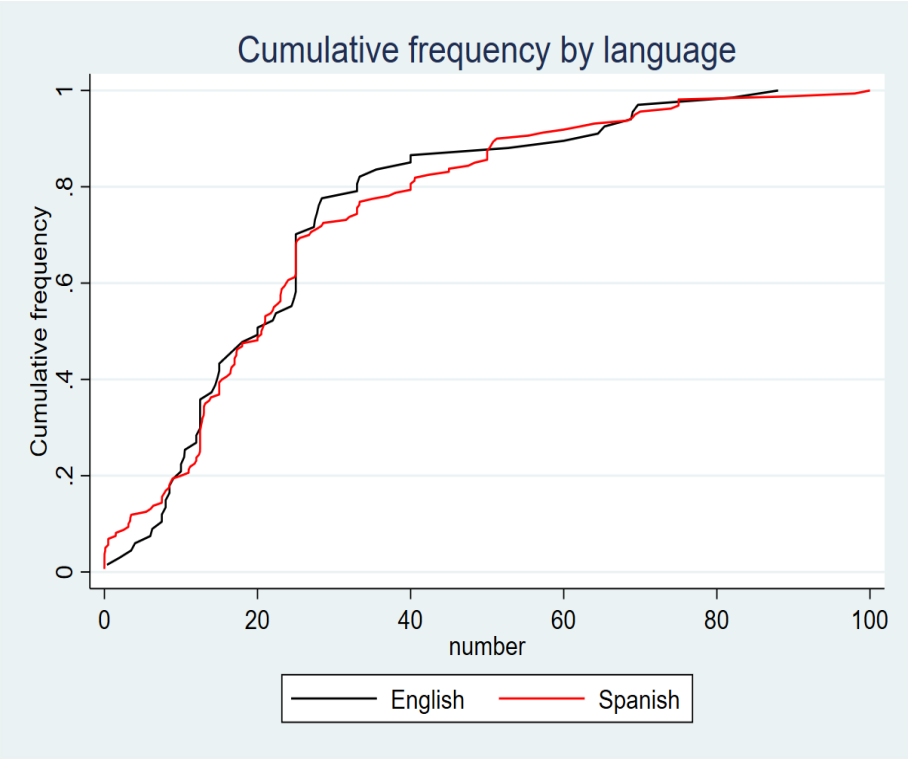

Figure 8: Cumulative frequency by the language of the class

7.5 Analyzing the other variables: Nationality

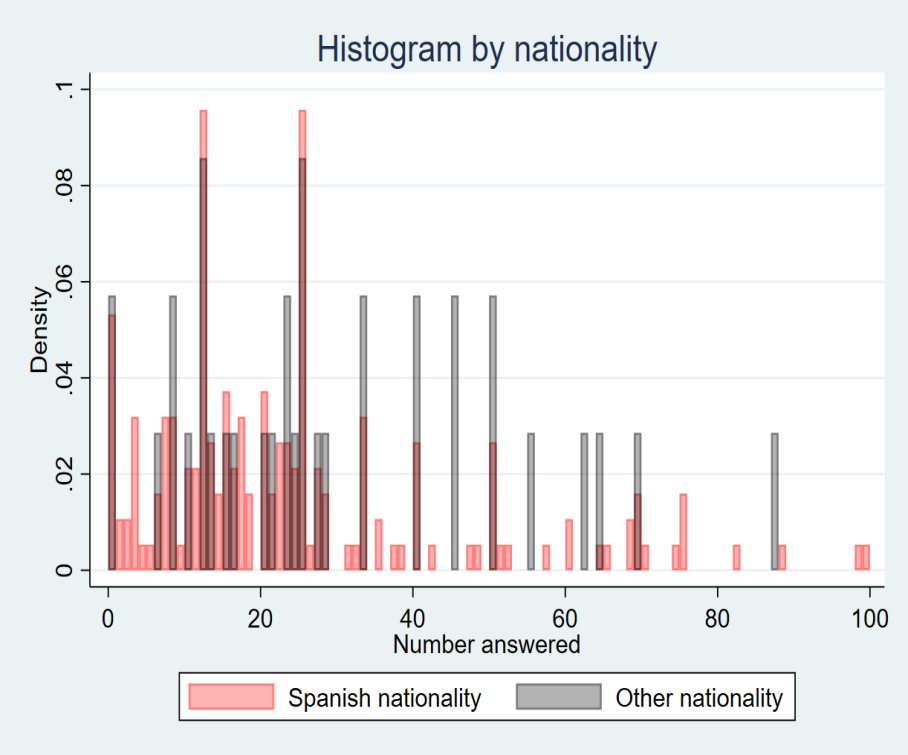

Figure 9: Histogram by the nationality of the subject

The histogram shows a very similar pattern of responses regarding nationality. In fact, we can observe that in the 3 spikes (0, 15, 25) there is almost the same density. However, subjects that do not have Spanish nationality seem to be responding higher numbers, while Spaniards are more concentrated in lower numbers. This can also be seen in the cumulative frequency graph; since the increase in the curve of Spaniards is faster than the curve from non-Spaniards. In fact, this second group starts to increase faster around the number 20. This could make me think that, maybe, Spaniards iterate more steps in their reasoning with respect to other nationalities, however, the statistical tests again show no difference in means, and no difference in distributions. (Mann-Whitney p-value=0.1294 and Kolmogorov-Smirnov p-value=0.367).

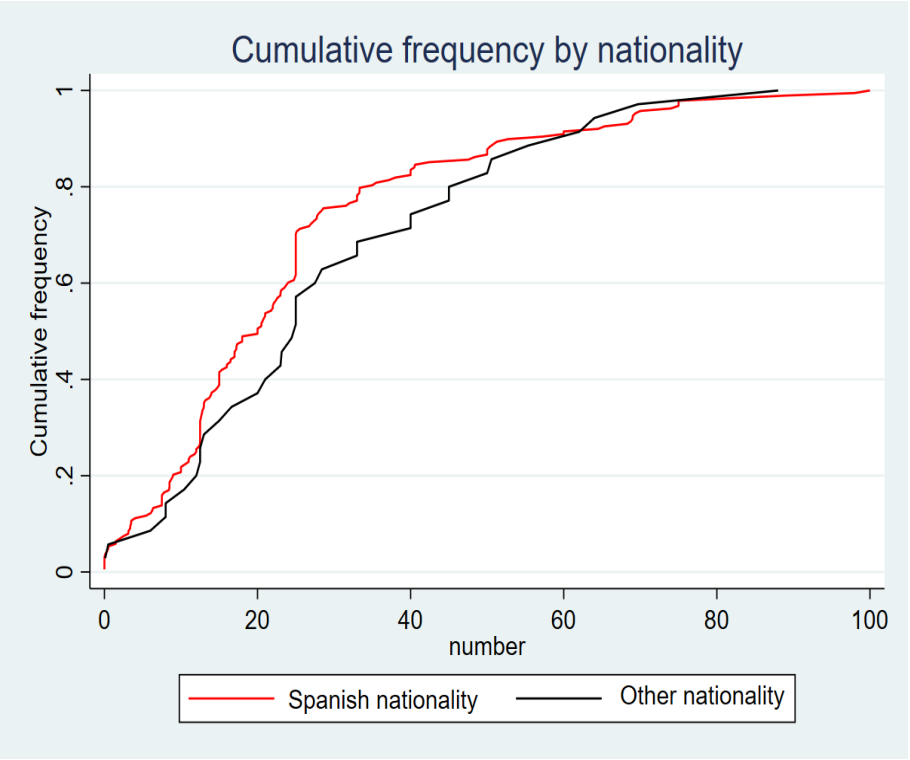

Figure 10: Cumulative frequency by the nationality of the subject

8 Further research

I leave for future research a key point that was in my original experimental design: ask the subjects why they chose that number and briefly explain it in no more than 50 words. This is the same procedure as Bosch-Domenech et al. (2002) did in the three newspapers they sent the experiment, and allowed them to classify the way of reasoning of each subject without any test. Moreover, I think that performing a proper experiment like this could really shed light in the differences between students. Because I had to gather all the data manually, I do not have as many observations as existing papers in level-k literature. Even though my data shows a little pattern of business students responding the Nash Equilibrium; I have no way to statistically test it. A theoretical model could be developed trying to test this idea. Furthermore, the Keynesian beauty contest is not the only game to test level-k reasoning, but there are other games that could be used, such as the 11-20 game.

9 Conclusions

Level-k framework is a very powerful tool to study human behavior. This theory, even with some limitations, helped explore different processes of thinking and iteration of humans. My main idea was to test if the process of reasoning of business and economics students is different. Within this framework, I developed an adapted version of the Keynesian Beauty Contest. I designed an experiment that I conducted in several sessions in classes of economics and business at the *Universitat de Barcelona*. I gathered data about 227 students of different years. I tested statistically and with graphs the appropriation of my two hypotheses. None of them was accomplished. However, I found an interesting additional result; business students were situated at the extremes of the levels of reasoning (i.e., level 0-1 or level-7). In fact, it was one of my expectations, that they would either be level-0, or solve the game by iterated dominance arriving to the last level of reasoning. I also provided a brief discussion about other characteristics that could influence the outcome of the game, such as gender, the language of the class or if the subject had prior knowledge of game theory, among others. I leave for future research to study this possible difference with a bigger data set.

References

Agranov, M., Caplin, A., and Tergiman, C. (2015). Naive play and the process of choice in guessing games. *Journal of the Economic Science Association*, 1(2):146–157.

Alaoui, L., Janezic, K. A., and Penta, A. (2020). Reasoning about others' reasoning. *Journal of Economic Theory*, 189:105091.

Alaoui, L. and Penta, A. (2013). *Level-k reasoning and incentives*. Citeseer.

Alaoui, L. and Penta, A. (2016). Endogenous depth of reasoning. *The Review of Economic Studies*, 83(4):1297–1333.

Arad, A. and Rubinstein, A. (2012). The 11-20 money request game: A level-k reasoning study. *American Economic Review*, 102(7):3561–73.

Bosch-Domenech, A., Montalvo, J. G., Nagel, R., and Satorra, A. (2002). One, two,(three), infinity,...: Newspaper and lab beauty-contest experiments. *American Economic Review*, 92(5):1687–1701.

Burchardi, K. B. and Penczynski, S. P. (2014). Out of your mind: Eliciting individual reasoning in one shot games. *Games and Economic Behavior*, 84:39–57.

Camerer, C. and Lovallo, D. (1999). Overconfidence and excess entry: An experimental approach. *American economic review*, 89(1):306–318.

Capra, M. (2019). Understanding decision processes in guessing games: a protocol analysis approach. *Journal of the Economic Science Association*, 5(1):123–135.

Costa-Gomes, M. A. and Crawford, V. P. (2006). Cognition and behavior in two-person guessing games: An experimental study. *American economic review*, 96(5):1737–1768.

Crawford, V. P. and Iriberri, N. (2007). Level-k auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions? *Econometrica*, 75(6):1721–1770.

De la Rosa, L. E. (2011). Overconfidence and moral hazard. *Games and Economic Behavior*, 73(2):429–451.

Fudenberg, D. and Tirole, J. (1991). Perfect bayesian equilibrium and sequential equilibrium. *journal of Economic Theory*, 53(2):236–260.

Handgraaf, M. J., de Jeude, M. A. V. L., and Appelt, K. C. (2013). Public praise vs. private pay: Effects of rewards on energy conservation in the workplace. *Ecological Economics*, 86:86–92.

Hemenway, D., Killen, A., Cashman, S. B., Parks, C. L., and Bicknell, W. J. (1990). Physicians' responses to financial incentives: evidence from a for-profit ambulatory care center. *New England journal of medicine*, 322(15):1059–1063.

Himmelstein, D. U., Ariely, D., and Woolhandler, S. (2014). Pay-for-performance: toxic to quality? insights from behavioral economics. *International Journal of Health Services*, 44(2):203–214.

Ho, T.-H., Camerer, C., and Weigelt, K. (1998). Iterated dominance and iterated best response in experimental" p-beauty contests". *The American Economic Review*, 88(4):947–969.

Jin, Y. (2021). Does level-k behavior imply level-k thinking? *Experimental Economics*, 24(1):330–353.

Kinderman, P., Dunbar, R., and Bentall, R. P. (1998). Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, 89(2):191–204.

Paz, V., Nicolaisen-Sobesky, E., Collado, E., Horta, S., Rey, C., Rivero, M., Berriolo, P., Díaz, M., Otón, M., Pérez, A., et al. (2017). Effect of self-esteem on social interactions during the ultimatum game. *Psychiatry Research*, 252:247–255.

Shapiro, D., Shi, X., and Zillante, A. (2014). Level-k reasoning in a generalized beauty contest. *Games and Economic Behavior*, 86:308–329.

Stahl, D. O. and Wilson, P. W. (1995). On players models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1):218–254.

Whalen, D. H., Zunshine, L., and Holquist, M. (2015). Increases in perspective embedding increase reading time even with typical text presentation: Implications for the reading of literature. *Frontiers in Psychology*, page 1778.

A Experiment

In this game you must choose a number with up to three decimals between 0 and 100. Your objective is to get as close as possible to 1/2 times the average of the numbers chosen by all the students in your classroom.

Example: there are 3 students, and the numbers chosen are 10, 20, 30. The average is 20, 1/2 of which is 10. The person who chooses 10 wins.

You must think carefully about your answer and think about it individually, without discussing it with your colleagues. If you win and you give your permission; your name will be announced publicly in the following weeks. If you prefer, you can be informed privately. In any case, your answers will be used only for the purpose of this analysis. You can also receive information about the results of the experiment.

Number chosen:

Name/NIUB:

Have you already taken any course on game theory?:

☐ Yes

☐ No

Course:

Degree

☐ Business

☐ Economics

If I win,

☐ I want to be informed publicly

☐ I want to be informed privately

☐ I do not want to be informed

Email (If you want to be informed):

Gender:

☐ Male

☐ Female

☐ I prefer not to specify

Age:

Spanish Nationality:

☐ Yes

☐ No