A Two-Step Guessing Game

King King Li, Kang Rong

A Two-Step Guessing Game

Li King King and **Rong** Kang [*]

August 2023

Abstract

We propose a two-step guessing game to measure the depth of thinking. We apply this method to the P beauty contest game. Using our method, we find that 81% of subjects do not make choice following best response reasoning while the classical method would suggest only 12%. The result suggests that the classical method has the fundamental problem that it cannot distinguish if a submitted number is due to best response reasoning or not. It also suggests that traditional level *k* analysis falsely attributes some sophistication to random players, and that the degree of false attribution is large. Our procedure provides an alternative way to identify whether the individual has best response reasoning which is essential for any positive level of depth of thinking and differentiates between the depth of thinking and random choice, and hence provides a very different conclusion, which is suggestive of limitations of the classical method.

Keywords: P Beauty Contest; Best Response Reasoning; Experiment

JEL Classifications: D81; C7

[*] Li (Corresponding Author): Shenzhen Audencia Financial Technology Institute, Shenzhen University; Email: likingking@gmail.com. Rong: School of Economics, Shanghai University of Finance and Economics; Email: rong.kang@mail.shufe.edu.cn. We thank Rosemarie Nagel, Tanjim Hossain, Pablo Brañas Garza, Xu Zhibo, Li Lunzheng, and participants of ShanghaiTech SEM Experimental Workshop 2018 for helpful discussions and comments, and Leung Cheuk Kiu and So Wing Yeung for research assistance. We gratefully acknowledge the financial support from the Hong Kong Research Grants Council (Grant No. 21501915) and the National Natural Science Foundation of China (Grant No. 71973099).

1. Introduction

In the classical P beauty game (Nagel, 1995), each player submits a number between 0 and 100 (0 and 100 inclusive). The player whose submitted number is closest to p*average of the submissions wins the prize, where p is any number between 0 and 1 (0 and 1 inclusive). Thus, a player should not submit a number higher than p*100 as she will lose the game for sure. The unique equilibrium is to submit zero. A player who submits a number higher than p*100 can be considered to be bounded rational. Yet, in the classical P beauty game, when a player submits a number lower than p*100, it is unclear if the player chooses the number because he can formulate the best response but not sure about the level of the other players or because he can't formulate the best response and chooses the number for other reasons, e.g., random choice. It is plausible that a subject who does not understand the game may choose a number randomly.[1] If so, the submitted number will be likely to be larger than zero and hence being wrongly interpreted as reflecting some sort of ability to formulate the best response.

This paper proposes a two-step guessing game to accurately measure the depth of thinking (Keynes, 1936; Nagel, 1995). Our paper makes the important point that traditional level k analysis falsely attributes some sophistication to random players. Our two-step procedure is as follows. In the first step (human version), the classical P beauty contest game is played. In the second step (computer version), players are randomly matched to play the two-player P beauty game, where player 1 submits an integer between 0 and 100, while player 2's decision will be made by a computer that will submit a number which is equal to 0.9 times the number submitted by player 1. After the games, we elicit subjects' underlying reasons for the submitted numbers by asking them to write down the reasons which we will use for text analysis.

The unique feature of our design is that we can use the choice in the computer version of the beauty contest game to infer whether the player understands best response or not, and combine this information with his choice in step 1 to infer his depth of thinking. We illustrate the idea using the following example. Suppose a player submits 30 in step 1 and higher than 0 in step 2. If we only rely on the classical method of measuring the level of thinking, we may infer that the player exhibits some degree of depth of thinking (in fact, it is level 3 if the level is estimated using

[1] For paper on preference for randomization, see e.g., Li (2011), Gul et al. (2014), and Agranov and Ortoleva (2017).

elimination of iterated dominated strategies).[2] Yet, note that in step 2, the player should always submit 0, as submitting a number higher than 0 will lose the game for sure.[3] Hence, we know that this player, indeed, does not have strategic reasoning and should be classified as level zero (as in our method) rather than level 3 (in the classical method). Note that we assume that failure to best respond in step 2 implies failure to best respond in step 1. We admit that our measure of best responding may be narrow in the sense that we do not consider the case that players may have sophisticated strategic reasoning (i.e. generally understanding the logic of undercutting, being able to approximately calculate levels of arbitrary degree, etc.) and still make a mistake in best responding.

One may wonder why there should be any relationship between depth of thinking in step 1 and step 2 as the two games are not the same or subjects may not use the same reasoning when playing the games. This is, indeed, a valid concern, despite one may argue that both games are similar. As an attempt to address this concern, we analyze the reasons submitted using text analysis.[4] We find suggestive evidence that the reasons submitted indicate that significant proportion of subjects explicitly mention that they are using the same reasoning in both games.

Our two-step procedure has several attractive features. First, the computer version in step 2 can clearly identify the depth of thinking when there is *no uncertainty* about the depth of thinking of the other player (i.e., there is no strategic uncertainty). Thus, the number submitted does not depend on the belief on the number submitted by other players as in the classical beauty contest game. Hence, it can differentiate between a submission that differs from the equilibrium prediction that is due to uncertainty on the level of the other player, or due to bounded rationality or non-strategic choice of the player. Second, it allows us to classify players into best response reasoning type or non-best response reasoning type. Hence, combining the choice in the two steps provides an estimation of the distribution of the depth of thinking of the population, as compared to the conventional methods such as the P beauty contest game (Nagel, 1995) and the 11-20 game (Arad and Rubinstein, 2012). Third, the suggestive evidence from the text analysis reveals the underlying

[2] Number 30 would imply level 1 if we use 50 as the reference point.
[3] To see this, let x be the player's submitted number, then p*((x+0.9x)/2) = p*(0.95x). Since p<1, 0.9x is closer to p*(0.95x) than x unless x=0, in which case the player wins with 1/2 probability.
[4] A number of recent papers have adopted text analysis of reasoning in games, see e.g., Branas-Garza, et al. (2011).

reasons for the submitted numbers and hence can identify non-strategic choices accurately which cannot be achieved by looking at the numbers alone in the classical method.

Our method can also be applied to the 11-20 game. In the 11-20 game, the two players are randomly matched. Each player submits a number between 11 and 20. The player will receive a payoff that is equal to the number submitted plus the prize. The player wins the prize if the number submitted is one less than the other player. In this game, a player who submitted the number 20–x is classified as level x, where $0 \leq x \leq 9$ and x is an integer. The game has no pure strategy equilibrium, but there is a unique (symmetric) mixed strategy Nash equilibrium (see Arad and Rubinstein, 2012). Similar to the case of P beauty contest game, using this game alone cannot tell whether the player submits a number because he is strategic and believes this is the best response to the other player's submission or because he is non-strategic and choosing the number for other reasons.

Grosskopf and Nagel (2008) conduct a two-person beauty contest game.[5] In their game, two players are randomly matched and each player submits a number in the interval [0, 100], and the winner is one whose number is closest to two-thirds of the mean of the chosen numbers of the two players. The special feature of their game is that choosing zero is the weakly dominant strategy and is always the winning number regardless of the choice of the other player. They found that a small proportion of subjects chooses zero in the experiment, in particular, 9.85% for student subjects, and 36.92% for professionals who participated in economics and psychology decision-making conferences. As an explanation, Grosskopf and Nagel (2008) argue that subjects misunderstand the rule of the game and choose a number that is as closest as possible to 0.7*the expected average of the two players.

In other words, Grosskopf and Nagel (2008) argue that the fact that a large majority of subjects submit a number higher than zero is due to misunderstanding of the rule of the game. The key difference between our design (computer version) and Grosskopf and Nagel (2008) is that we

[5] Nagel et al. (2017) had a two-person beauty contest game where players are paid according to their distance to 2/3 times the average of both numbers. In their game, choosing zero is a unique equilibrium.

focus on inferring best response and combine it for analysis of the classical beauty contest game. Further, in our computer version, choosing any number higher than zero is strictly dominated.

Bosch-Rosa et al. (2018) conducted a one-player version of Grosskopf and Nagel's (2008) two-person guessing game. In the one-player game, each subject i picks two numbers $a_i \in [0,100]$ and $b_i \in [0,100]$ and subjects are paid for both choices, the payoff $1 - 0.05\left|a_i - \frac{2}{3}\frac{a_i+b_i}{2}\right|$ for a_i and $1 - 0.05\left|b_i - \frac{2}{3}\frac{a_i+b_i}{2}\right|$ for b_i. That is, the payoff depends on the absolute distance of each chosen number to the two thirds of the average of both numbers, and the payoff is maximized at $a_i = 0$ and $b_i = 0$. Thus, the player plays against himself. The one-player guessing game is similar to our computer version game.[6] In both games, there is just one player and the equilibrium is to choose zero.

There are three crucial differences between our two-step guessing game and the one-player guessing game of Bosch-Rosa et al. (2018). First, our two-step procedure contains the classical P beauty contest game, and thus our focus is on whether traditional level *k* analysis falsely attributes some sophistication to random players. Second, we elicit subjects' reasons underlying the submitted numbers. We conduct text analysis on the submitted reasons which allows us to check the reasons behind the choices, which is important to infer if subjects' submitted numbers are based on non-strategic consideration and also if subjects use the same type of reasoning across beauty contest games. Third, we believe our procedure to identify best response in the computer version is simpler and easier to understand as it requires submitting only one number instead of two, and the algorithm of the computer is clearly mentioned, which makes the subjects easier to figure out zero is the best response.[7]

Bosch-Rosa and Meissner (2020) conduct the one-player guessing game of Bosch-Rosa et al. (2018) and a modified version of the two-person guessing game by Grosskopf and Nagel's (2008). In their modified version of two-person guessing game, subjects are matched in pairs and

[6] We developed our research idea independently and conducted the experiment in 2018.

[7] Our computer version is essentially a decision problem as only one player makes the decision while the one-player guessing game by Bosch-Rosa et al. (2018) is a game with two selves making two decisions. Thus, it is automatically simpler and easier to figure out zero is best response (or optimal choice) in our problem. In the one-player guessing game, however, it needs some logic induction starting with a random value and then calculate the best responses of two selves iteratively to figure out (0,0) is an equilibrium as mentioned by Bosch-Rosa and Meissner (2020).

asked to pick a number $z_i \in [0,100]$, and the payoff is based on the absolute distance of each subject in the pair pick to 2/3 of the average of both numbers, and the best response is no longer zero but to choose 1/2 of the number a player believes the other player chooses. Note that, in the modified version of the two-person guessing game, one can no longer infer if a subject is making best response (as in our computer version). They find that majority of subjects fail to understand the structure of the one-player guessing game, and subjects with a better understanding submit choices closer to the Nash Equilibrium in the modified two-player guessing game.

Costa-Gomes and Crawford (2006) conduct two-person beauty contest games that are, however, very different from our design and Grosskopf and Nagel (2008). In Costa-Gomes and Crawford (2006), the games are asymmetric (each player has a lower and upper limit), with different values (0.5, 0.7, 1.3, or 1.5) of p for different players, and dominance-solvable in 3 to 52 rounds. Camerer et al. (2004) conducted a two-person contest game to test the cognitive hierarchy model. More importantly, in both Camerer et al. (2004) and Costa-Gomes and Crawford (2006), they cannot discriminate if a submitted number is due to a lack of best response or not.

Our method is also connected with two other strands of literature. The first strand of literature is about addressing measurement errors in experiments using multiple measures (see e.g., Gillen et al., 2019). In the highly influential paper, Gillen et al. (2019) demonstrate that experimental results on three classic experiments on overconfidence, risk, and ambiguity, change substantially when experimental measurement error is accounted for.

The second strand of literature is about testing level-k and related models (see e.g., Duffy and Nagel, 1997; Ho et al., 1998; Bosch-Domenech et al., 2002; Costa-Gomes and Crawford, 2006; Crawford and Iriberri, 2007) which relies on accurate identification of strategic reasoning and depth of thinking.[8] It is thus important to have an accurate measure of depth of thinking for accurately testing level-k models.

A few recent papers (Friedenberg et al., 2018; Jin, 2018; Alaoui et al., 2020) have investigated a related question in disentangling whether level-k behavior is due to cognitive limitations or beliefs about others. Our research question is different in the sense that it is more about how to determine whether the player uses best response using a very simple way. In our

[8] See Nagel et al. (2017) for an extensive review.

method, when a subject submits zero (positive number) in the computer version of the P beauty contest game, we can be sure that she is likely of having (no) best response, and hence her submitted number in the classical P beauty contest game is likely due to her beliefs about others (lack of best response), under the assumption that failure to best respond in step 2 implies failure to best respond in step 1. In this sense, our study is more about detecting whether the subject understands best response in the P beauty contest game. Note that in these papers, they focus on measuring cognitive limits on how many steps the players can think ahead. We, instead focus on investigating, arguably a more fundamental question, whether the player understands best response. Another difference is about the identification strategy and the "game" used. Alaoui et al. (2020) use the modified 11-20 game (Arad and Rubinstein, 2012) and a "tutorial method" to teach subjects about game theory. Friedenberg et al. (2018) and Jin (2018) both use ring games which was first introduced by Kneeland (2015) to study higher-order rationality. Our identification strategy relies on the novel design in the "computer version" and has the advantage that it can differentiate between best response and random choice, as well as being very simple and easy to understand.

The main findings of this study can be summarized as follows. We find that a significant proportion of players do not exhibit best response reasoning. In the computer version of the two-person beauty contest game, about 81% of players submit a number higher than 0 in the computer version, hence indicating non-understanding of best response. We combine the best response measure in the computer version with the human version of the beauty contest game, under the assumption that failure to best respond in step 2 implies failure to best respond in step 1. We find that about 81% of subjects are level zero, while the classical method would suggest only 12%. Taken together, our result suggests that the classical method falsely attributes some sophistication to random players, and the degree of false attribution is large. It should be noted that our main contribution is not about the absolute percentage of players exhibiting strategic reasoning, as this can differ substantially across different subject groups (Levitt, List, and Sadoff, 2011).[9]

The text analysis on the underlying reasons for the submitted numbers reveals suggestive evidence that the classical method will wrongly classify random choice as strategic choice. It also suggests that subjects tend to use the same reasoning across the games and hence supporting the

[9] See also Alaoui and Penta (2015) for a model of endogenous depth of thinking where the player's "depth of reasoning" is endogenously determined. Gill and Prowse (2016) found that participant's choices in the P beauty contest game respond positively with the cognitive ability of their opponents.

validity of our approach. More importantly, it suggests that our approach is able to identify most random choice as non-best response.

The rest of the paper is organized as follows. Section 2 reports the theoretical analysis, and section 3 reports the experimental design. Section 4 reports the experimental results. Section 5 concludes.

2. Theoretical Foundation

In general, a player's decision in a game is mainly influenced by two factors. One is the player's *belief* about other players' choices. The other is how the player *responds* to his belief about other players' choices. Based on the second factor, we can classify players into two categories. One is those who can best respond to their beliefs about opponents' choices, and the other is those who cannot. We call the players in the first category *strategic* players, and the players in the second category *non-strategic* players. We assume that whether a player is strategic or not is an inherent ability of the player, and thus, a player who is strategic in some game will also be strategic in some similar games and a player who is non-strategic in some game will also be non-strategic in some similar games.[10]

More specifically, our two games (the classic beauty contest game (N=2) and the computer version of beauty contest game) are quite similar, and we should expect that whether a player is strategic or not in a game implies whether the player is strategic or not in the other game. In addition, note that in the computer version of beauty contest game, there is no strategic uncertainty about the computer's choice (as the computer's choice is passively determined by the human player's choice). That is, for the human player in the computer version, there is no uncertainty about his belief about the opponent's choice. Thus, the only factor that determines a player's choice in the computer version is whether he is a strategic player or a non-strategic player. This implies that if a player chooses a number greater than zero in the computer version, then he is likely a non-strategic player, and if the player chooses a number equal to zero, then he is likely a strategic player. We can then use this information to get a better understanding of the player's choice and

[10] Using text analysis on the reasons behind the submitted numbers, we obtain supportive evidence that subjects use the same reasoning across the games.

depth of thinking in the classic version of the beauty contest game. In particular, if a player is strategic in the computer version, then, we assume that, he should also be strategic in the classic version, and thus his choice in the classic version indeed reflects how he thinks his opponent's choice and thus reflects the player's depths of thinking. On the other hand, if a player is non-strategic in the computer version, then, we assume that, he should also be non-strategic in the classic version, and thus the player's choice in the classic version is more like a random number and thus cannot be used to infer the player's depth of thinking.

Difference with Grosskopf and Nagel (2008)

Grosskopf and Nagel (2008) argue that in the P beauty contest game, players may misunderstand the rule of the game and choose a number that is as closest as possible to p*average of the two players. We adopt this assumption for our theoretical analysis. With this assumption, in the following, we first show that in the 2-person P beauty contest game (Grosskopf and Nagel, 2008), a player with the misunderstanding will submit a number greater than zero in equilibrium, and then show that in our computer version of the 2-person P beauty contest game, a (strategic) player will always submit zero in equilibrium.

Consider the 2-person P beauty contest game. We assume that there is α proportion of non-strategic players, who will randomly choose a number between 0 and 100.[11] The remaining $1 - \alpha$ proportion of players are "strategic", but they "misunderstand" the rule of the game, and choose a number that is as closest as possible to 0.7*the expected average of the two players (as in Grosskopf and Nagel, 2008). In particular, we use x to denote the number chosen by a strategic player. Then, the expected value of the number chosen by his opponent is $\alpha \times 50 + (1 - \alpha) \times x$ (noting that if his opponent is non-strategic (which occurs with probability α), then the average number of his opponent is 50, and if his opponent is strategic (which occurs with probability $1 - \alpha$), then the number chosen by his opponent will be x). So, we have:

$$x = 0.7 \times 0.5(x + \alpha \times 50 + (1 - \alpha) \times x)$$

[11] If we don't have non-strategic types (i.e., all players are strategic), then it can be shown that even if we have the assumption by Grosskopf and Nagel (2008), (strategic) players will submit zero in equilibrium.

Solve the equation, we have $x = 17.5\alpha/(0.3 + 0.35\alpha)$. It can be verified that if $\alpha = 0$, then $x \approx 0$; if $\alpha = 0.1$, then $x \approx 5$; if $\alpha = 0.5$, then $x \approx 18$; and if $\alpha = 1$, then $x \approx 27$.

Now consider the computer version of the 2-person P beauty contest game. If a player is non-strategic, then he still randomly chooses a number between 0 and 100. If a player is strategic, we still assume that he "misunderstands" the rule of the game, and choose a number that is as closest as possible to 0.7*the expected average of the two players. In particular, we use y to denote the number chosen by a strategic player, then we have:

$$y = 0.7 \times 0.5(y + 0.9y)$$

Solve the equation, we have $y = 0$. That is, in the computer version of 2-person P beauty contest game, a strategic player will always choose 0.

3. Experimental Design

We conduct an online experiment in 2018 with 192 subjects who are undergraduates in a major university in Hong Kong. The subjects are randomly recruited from an email announcement to approximately 3,000 subjects registered in the subject pool. Subjects receive HKD 20 for participation and payoff from a randomly drawn game.[12] In addition to the P beauty contest games, we also elicit subjects' reasons of choices for the respective beauty contest games, attitudes on investing in stock with price bubbles, and cognitive reflection test score (Frederick, 2005).[13]

Two-Step Guessing Game

Step 1 (Human version): Each player chooses a number between 0 and 100 (0 and 100 inclusive). The participant with the chosen number being closest to 0.7 times the average of all chosen numbers wins a prize of HKD 20. That is, the subject whose submitted number is closest to the average of all submitted numbers x 0.7 wins. If two or more participants win, the winner will be randomly chosen.

[12] 1 USD equals to about HKD 7.78.
[13] See online appendix A for the results on price bubbles and cognitive reflection test.

Step 2 (Computer version): Participants are randomly matched into groups. Each group is consisted of two participants, one is called participant 1, and the other is called participant 2. Both participants choose a number between 0 and 100 (0 and 100 inclusive).

However, participant 2's choice will be implemented by a computer which will always choose a number that is equal to the number chosen by participant 1 times 0.9. That is, participant 2's number = participant 1's number x 0.9

The participant with the chosen number (i.e., participant 1's chosen number and participant 2's number chosen by the computer) being closest to 0.7 times the average of the chosen numbers wins a prize of HKD20.

That is, the one whose submitted number is closest to the following number wins:(average of submitted numbers by participant 1 and participant 2) x 0.7. If two or more participants win, the winner will be randomly chosen.

We conduct two versions of the human version, n=2, and n>2. In the human (n>2) version, subjects play the classical P beauty game with all the subjects. [14] In the human (n=2) version, subjects play the p beauty contest game with another participant. Each subject plays all the three games, human version (n>2), human version (n=2), and computer version. We estimate the depth of thinking in the P beauty game in the human version in the following way. A player who submitted a number larger than 70 was classified as level 0. In general, level k player submits a number in the range of $(0.7^{k+1}100, 0.7^{k}100]$.

4. Experimental Results

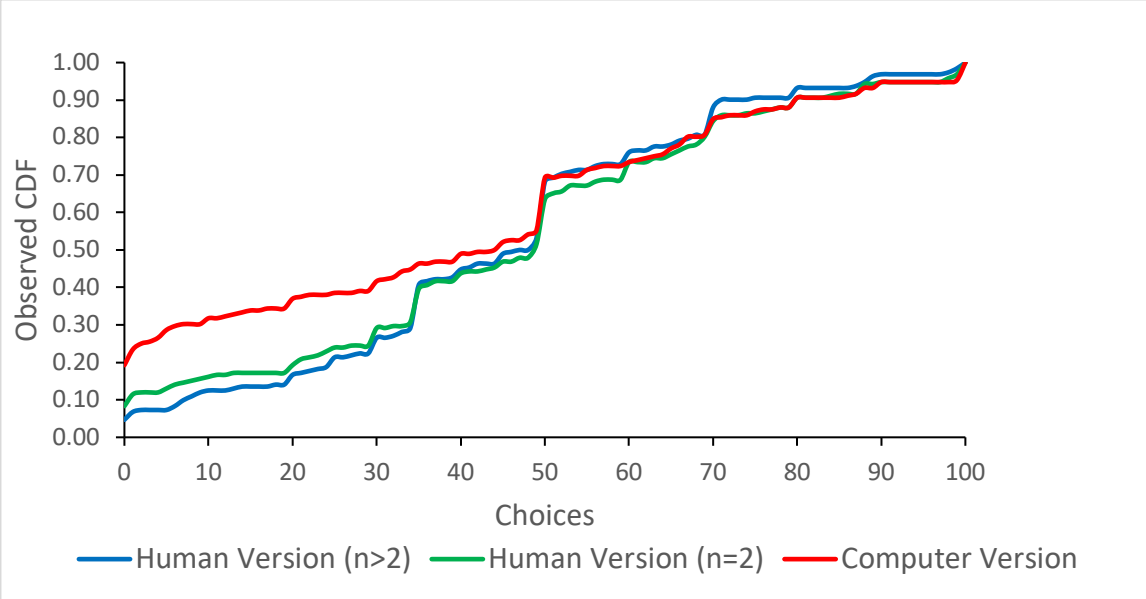
Figure 1 reports the cumulative distribution function (CDF) of the choices in the three versions of the beauty contest game. A striking pattern is that the CDF of computer version appears to be very different from the CDF of human version (n>2) and human version (n=2), while the latter two largely resemble each other. More specifically, there are more subjects with submitted numbers

[14] In our online experiment, a total of 9 games were conducted. These 9 games are (1) beauty contest game (n>2, human version), (2) beauty contest game (n=2, computer version), (3) beauty contest game (n=2, human version), (4) beauty contest game (70 human participants and 30 computers), (5) beauty contest game (90 human participants and 10 computers), (6) beauty contest game (99 human participants and 1 computer), and (7-9) three games on choosing 1-10. In the current paper, we focus on the game (1), (2), and (3).

less than 50 in the computer than in the human version (n>2) and human version (n=2). In fact, the mean of the numbers submitted in the computer version is 38.48, which is significantly lower than 44.54 of the human version (n>2) and 44.95 of the human version (n=2), with *p*-value equals to 0.003, and 0.003 under paired t-test, respectively. There is no significant difference between the two human versions, with *p*-value equals to 0.81 under paired t-test.[15] This supports the hypothesis that removal of strategic uncertainty in the computer version changes subjects' choices significantly. The distribution of choices is significantly different between computer version and human version (n>2), and human version (n=2), with *p*-value equal to 0.002, and 0.003, under the Wilcoxon Sign-rank test respectively. There is no significant difference on distribution of choices in the two human versions, with *p*-value equals to 0.95 under the Wilcoxon Sign-rank test.

Result 1: The mean of numbers submitted in the computer version of the beauty contest game is lower than in the human versions, while there is no difference between human version (n>2) and human version (n=2).

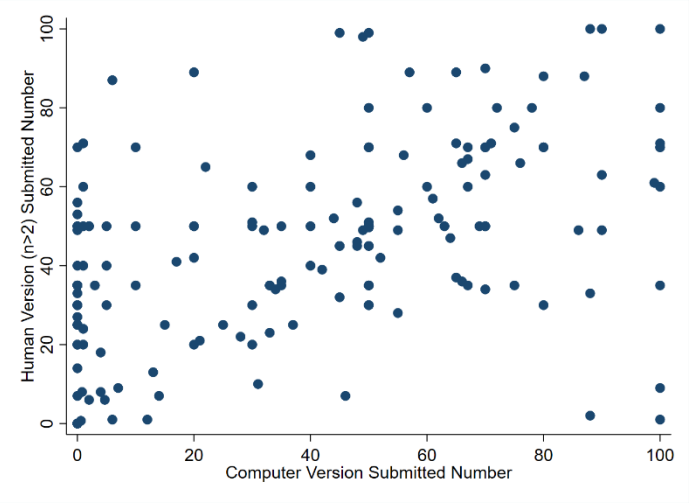
Figure 1. Cumulative Density Function of Choices



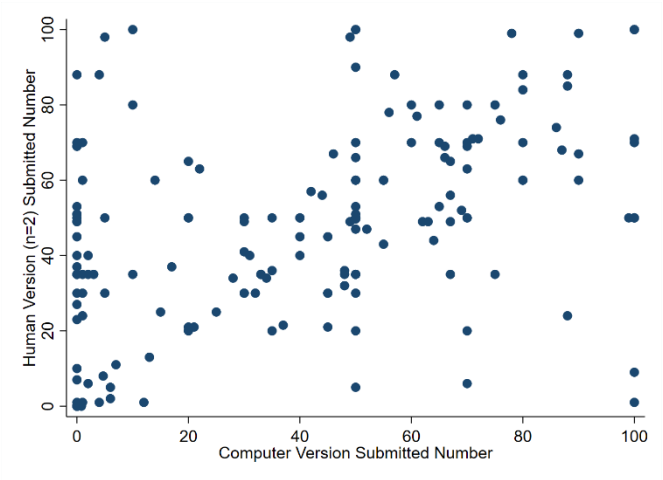
[15] A similar result is obtained when using Sign test. There is significant difference on median of choices in human version (n>2) and computer version, and human version (n=2) and computer version, with *p*-value equals 0.01, under Sign test, respectively. There is no significant difference between the two human version, with *p*-value equals to 1.00 under Sign test.

Figure 2. Submitted Numbers

a. Human Version (n>2) and Computer Version



b. Human Version (n>2) and Computer Version



We find that 19.17% of subjects submitted 0 in the computer version (n=2) of the beauty contest game. Under our framework, this implies that only 19.17% of subjects' choices are consistent with best response.[16] If a subject submits a number higher than 0 (i.e., no best response) in the computer version, she is classified as level 0 in the combined analysis.[17] However, it should

[16] We find that subjects with higher cognitive reflection test score are not more likely to exhibit best response., suggesting that ability to formulate best response is distinct from cognitive ability.

[17] We assume that the ability for formulating best response is a fixed trait. While the computer version and human version have different degree of complexity, we believe that it is reasonable to assume that a subject who chooses zero in the computer version would not choose above 70 in the human version. This is, indeed, what we found.

be noted that our approach assumes that when subjects fail to best respond in step 2, they also fail to best respond in step 1. We also assume that any number submitted other than zero in step 2 are considered to be not best responding. That is, we do not consider the possibility of mistakes. If a subject submits higher than 70 in the human versions, she is classified as level 0 in the respective treatments.

Result 2: Only 19% of subjects' choices are consistent with best response.

Figure 2a and Figure 2b plots the submitted numbers in the human version (n>2) against the computer version, and the human version (n=2) against the computer version, respectively. Conditional on subjects who submitted zero in the computer version, none submitted higher than 70 in the human version (n>2) and one (2.7%) submitted higher than 70 in the human version (n=2). Conditional on subjects who submitted higher than 70 in the human version (n>2), none submitted zero in the computer version. Conditional on subjects who submitted higher than 70 in the human version (n=2), only one (3.3%) submitted zero in the computer version. This provides the evidence that the game in step 2 is a good measure on best response, and the choices in the two steps are correlated.

Table 1 reports the estimated depth of thinking when using choices from the human version (n>2), and when combining choices from the human version (n>2) and the computer version, as well as combining choices from the human version (n=2) and computer version. It shows the striking finding that the distributions on the depth of thinking change dramatically when we combine the choices. In particular, combining the human version (n>2) and computer version, about 81% of subjects are classified as level 0 in the combined analysis, while it is only 12% when using choices from the human version (n>2) only. Combining the human version (n=2) and computer version, about 81% of subjects are classified as level 0 in the combined analysis, while it is only 16% when using choices from the human version (n=2). Note that the conclusion relies on a "narrow" view of level *k* models, one in which any failure of best response implies level 0. If we take a broader view of what it means to exhibit strategic reasoning (generally understanding the undercutting nature of the game, possibly making some mistakes in best responding, etc.), it is plausible that the percentage of subjects not exhibiting strategic thinking in the beauty contest game is less than 81%.

In summary, the combined analysis offers the advantage that random choice/choice based on lack of best response is identified, and it suggests that the classical method falsely attributes some sophistication to random players, and the degree of false attribution is large.

Result 3: Combining choices in the human versions and computer version leads to a substantially different conclusion on depth of thinking.

Table 1. Estimated Depth of Thinking in the P Beauty Contest Game

Level	0	1	2	3	4	5	6	7	8	9	10	11	≥12
Classic Method: Using Choices from Human Version (n>2)	0.12	0.38	0.21	0.10	0.06	0.01	0.02	0.04	0	0	0.05	0.02	0.05
Classic Method: Using Choices from Human Version (n=2)	0.16	0.37	0.17	0.08	0.06	0.01	0.02	0.02	0.01	0	0.01	0.03	0.08
Our Method: Combining Choices from Human (n>2) and Computer Version	0.81	0.03	0.06	0.04	0.01	0.01	0	0.01	0	0	0	0	0.05
Our Method: Combining Choices from Human (n=2) and Computer Version	0.81	0.05	0.03	0.01	0.01	0	0.01	0.01	0	0	0	0.01	0.08

For those submitted 0 in step 2, 40.5% in human version (n>2) and 24.3% in human version (n=2) submitted 0 in step 1, respectively. Among subjects who submitted 0 in step 2, in the human version (n=2), only 1 (2.7%) (7, 8.9%) chose above 70 (50); in the human version (n>2), no subject submitted more than 70, and 3 or 8.1% submitted more than 50. Thus, it suggests that for most subjects who choose zero in step 2, their choices in step 1 are consistent with best response (they may not choose zero as they may believe others are not choosing zero).

We find that subjects who understand best response (i.e., choosing zero) in the computer version tend to make a choice closer to the Nash equilibrium in the human version (n>2) and human version (n=2). In particular, the average number submitted in the human version (n>2) by those who understand best response is 26.46 which is significantly lower than 48.86 of those who don't understand best response, with *p*-value equal to 0.00 under two-sample t-test. Similarly, the average number submitted in the human version (n>2) by those who understand best response is

26.97 which is significantly lower than 49.72 of those who don't understand best response, with *p*-value equal to 0.00 under two-sample t-test.

Result 4: Subjects who understand best response (i.e., choosing zero) in the computer version tend to make a choice closer to the Nash equilibrium in the human version (n>2) and human version (n=2).

The implication of no-best response in the computer version is that this group of subjects are more likely to make no-best response choices in the human versions. This, indeed, what we find. Conditional on no-best response in the computer version, 14.8% submit a number higher than 70 in the human version (n>2) which is significantly higher than 0% observed with those follow best response, with *p*-value equal to 0.01 under two-sample test of proportions. On the other hand, conditional on no-best response in the computer version, 18.6% submit a number higher than 70 in the human version (n=2) which is significantly higher than 2.7% observed with those follow best response, with *p*-value equal to 0.01 under two-sample test of proportions.

An alternative approach to estimate depth of thinking is to use mixture model to analyze a large set of beauty contest data (see e.g., Stahl and Wilson, 1994; Bosch-Domènech et al., 2010; Georganas et al., 2015). We agree that the mixture model is appealing. However, we believe that our design has the unique advantage that we only need to have two steps of games which makes our game more practical for identifying individual levels, while the mixture model requires sufficiently high number of games (data) for the estimation. The mixture model is better for estimating levels in the population level, while our method can identity if the subject is level 0 at individual level. The step 2 of our game allows us to uniquely determine whether the subjects follow best response. In fact, this is the reason why our design does not require high number of games for estimation.

Robustness

An alternative way to compute the depth of thinking is to use 50 as the reference point.[18] Following Nagel (1995), we use the geometric mean to determine the boundaries of adjacent intervals

[18] Note that under this method, numbers higher than 50 are not assigned for any depth of thinking.

between two levels of depth of thinking.[19] Table 2 reports the estimated depth of thinking for human version (n>2) and human version (n=2), and for combined choices. It can be seen that the finding is similar with what is observed in Table 1 that the proportion of level zero subjects are substantially higher in the combined analysis than using the classic method.

Table 2. Estimated Depth of Thinking in the P Beauty Contest Game using 50 as Reference Point

Level	0	1	2	3	4	5	6	7	8	9	10	11	≥12	Unclassified
Classic Method: Using Choices from Human Version (n>2)	0.26	0.21	0.07	0.01	0.01	0.04	0.01	0	0	0.01	0	0.02	0.05	0.32
Classic Method: Using Choices from Human Version (n=2)	0.22	0.18	0.07	0	0.02	0.02	0.02	0	0	0.01	0	0.03	0.09	0.36
Our Method: Combining Choices from Human (n>2) and Computer Version	0.83	0.08	0.02	0.01	0	0.01	0	0	0	0	0	0	0.05	0
Our Method: Combining Choices from Human (n=2) and Computer Version	0.83	0.03	0.01	0	0.01	0.01	0	0	0	0	0	0.01	0.08	0.04

Some subjects may not perfectly best respond despite they understand the under-cutting logic of the beauty contest game. To address this concern, we add an analysis by allowing subjects to make mistakes in their response in step 2. We consider an alternative way to define whether the subject's choice follow best response in step 2 by taking the possibility of subjects may make some mistakes in choices into consideration. In particular, we set different level of "errors" such that choices fall within these error ranges are still considered to be consistent with best response. We consider error thresholds of 3, 5, 10. That is, subjects submitted a number equal or less than 3, 5, and 10 are considered to be exhibiting best response when error threshold=3, error threshold=5, error threshold=10, respectively. Using this method, the percentage of subjects who follow best response in step 2 are 25.5% when error threshold=3, 28.7% when error threshold=5, 31.8% when error threshold=10. Table 3 shows that there are still large proportions of subjects classified as level zero in the combined analysis under the respective error thresholds. However, note that a

[19] As a robustness check, we also use $50p^n$, where p=0.7 and n is 0, 1, 2, … referring to the depth of thinking, to compute the intervals. We find that the results are similar.

weakness of this robustness check is that there is no iterative structure to the computer game and choosing anything strictly greater than 0 yields the same low payoff.

Table 3. Estimated Depth of Thinking in the P Beauty Contest Game Under Different Error Thresholds of Best Response

Level	0	1	2	3	4	5	6	7	8	9	10	11	≥12
Error Threshold=3 Our Method: Combining Choices from Human (n>2) and Computer Version	0.75	0.05	0.07	0.04	0.02	0.01	0	0.02	0	0	0	0.01	0.05
Error Threshold=3 Our Method: Combining Choices from Human (n=2) and Computer Version	0.75	0.06	0.05	0.02	0.01	0	0.01	0.01	0	0	0	0.02	0.08
Error Threshold=5 Our Method: Combining Choices from Human (n>2) and Computer Version	0.72	0.06	0.07	0.04	0.03	0.01	0	0.03	0	0	0	0.01	0.05
Error Threshold=5 Our Method: Combining Choices from Human (n=2) and Computer Version	0.73	0.06	0.05	0.02	0.01	0	0.01	0.02	0	0	0	0.02	0.08
Error Threshold=10 Our Method: Combining Choices from Human (n>2) and Computer Version	0.69	0.07	0.08	0.04	0.03	0.01	0.01	0.03	0	0	0	0	0.05
Error Threshold=10 Our Method: Combining Choices from Human (n=2) and Computer Version	0.71	0.06	0.06	0.02	0.01	0	0.01	0.02	0.01	0	0.01	0.02	0.08

Text Analysis

In the questionnaire, we ask subjects to write down how they choose the number in each of the beauty contest games. This allows us to conduct *text analysis* on the written reasons to infer their intentions behind the numbers they submitted.

We first look for subjects who indicate they are choosing their numbers *randomly*. In the human version (n>2), 95.83% of subjects have written down their reasons, and among these subjects, 17.93% indicated that they choose the numbers randomly. For example, one subject said "I randomly select it", and another said, "Just randomly choose". Among these subjects, the average number submitted is 50.51, with 5 subjects (5%) submitting a number higher than 70, and 36% of subjects submitted a number less than 50. This implies that in the classical method these subjects are wrongly classified as exhibiting some positive degree of depth of thinking (as most of them submit a number less than 70).

Column 1 of Table 4 reports the OLS regression where the dependent variable is the estimated depth of thinking in human version (n>2), and the independent variable is whether the subject indicated choosing randomly in the human version (n>2). The coefficient on Random Reason Human (n>2) is significantly negative, suggesting that those choosing randomly are estimated to have lower level. A similar pattern is found with human version (n=2). It suggests that there is a strong correlation between stated reasoning and behavior. Column 3 of Table reports the marginal effect estimates of probit regression where the dependent variable is whether the subject's choice in computer version is consistent with best response, and the independent variables are random reason in the computer version, level of depth of thinking in human version (n>2), and level of depth of thinking in human version (n=2). While the coefficient on random reason is not statistically significant, the sign is consistent with our prediction. In fact, 10.7% of subjects who stated they chose randomly chose zero in the game, which is lower than 23.0% observed with subjects who didn't give random as a reason. The regression also shows that the probability of best response is positively correlated with levels of depth of thinking in human version (n>2) and human version (n=2), suggesting that subject's responses are correlated across games.

Table 4. Random Reasoning and Depth of Thinking

	(1) Level Human (n>2)	(2) Level Human (n=2)	(3) Best Response Computer Version
Random Reason Human (n>2)	-4.91***		
	(1.80)		
Random Reason Human (n=2)		-9.41***	
		(2.34)	
Random Reason Computer			-0.05
			(0.09)
Level Human (n>2)			0.01***
			(0.00)
Level Human (n=2)			0.01***
			(0.00)
Constant	7.17***	11.62***	
	(1.67)	(2.29)	
Observations	192	192	192
R-squared	0.00	0.01	0.33

Notes: Column 1 reports the OLS regression where the dependent variable is depth of thinking in human version (n>2). Column 2 reports the OLS regression where the dependent variable is depth of thinking in human version (n=2). Column 3 reports the marginal effect estimates of probit regression where the dependent variable is whether the subject exhibits best response in the computer version. Random reason human (n>2) is a dummy that equals 1 if the subject indicated choosing randomly in the human version (n>2), zero otherwise. Random reason human (n=2) is a dummy that equals 1 if the subject indicated choosing randomly in the human version (n=2), zero otherwise. Random reason computer is a dummy that equals 1 if the subject indicated choosing randomly in the computer version, zero otherwise. Level human (n>2) is the estimated depth of thinking in human version (n>2) under classic method. Level human (n=2) is the estimated depth of thinking in human version (n=2) under classic method. *, **, *** denotes significance at the 10%, 5%, and 1%, respectively.

Result 5: Significant proportion of subjects indicate that they chose randomly in the beauty contest games. Text analysis on the reasons behind the submitted numbers suggests that the classical method falsely attributes non-strategic (random choice) as exhibiting some positive degree of depth of thinking.

One way to verify if our best response measure is accurate is to check the proportion of subjects who indicated choosing randomly are classified as not exhibiting best response. If our measure is accurate, a high proportion of subjects who indicated choosing randomly should be classified as non-best response subjects under our method. This is, indeed, the case, 94.74% (human version, n>2), 89.29% (human version n=2), and 92.86% (computer version) of subjects who indicated choosing randomly are classified under our method as non-best response subjects in the respective treatments. This, again, shows that our method is better than the classical method in identifying non-best response subjects.

We also find that 91.9% of subjects who at least explicitly indicated once they are submitting the number randomly are classified as exhibiting non-best response, suggesting our measure performs well in identifying random response.

By looking at the reasons submitted, we can have a better judgment on whether the subjects' numbers reflect best response thinking as compared to the classical method. For subjects who are not exhibiting best response, some indicated that they do not know what to do. For example, a subject wrote when explaining his choice in the computer version "The game is very complicated. I do not know how to have the best strategy to win." Another wrote "Whatever I choose a number, I will lose the game, according to the game rule," indicating that the subject didn't understand the rule.

Similar Reasoning Across Games

We find that there is significant positive correlation between those indicating random as reasons in the three conditions. In particular, the correlation coefficients between random reasoning in human version (n>2) and human version (n=2) is 0.51 (p-value=0.00), human version (n>2) and computer version is 0.60 (p-value=0.00), human version (n=2) and computer version is 0.75 (p-value=0.00). This offers evidence that subjects use the same reasoning across the three

conditions.[20] In other words, if a subject is choosing the number randomly in the human version (n>2), she will likely be chosen randomly in the other two conditions also. This implies that the best response measure elicited in the computer condition is a good prediction of best response in the other two conditions, thus validating our method. In fact, 27.12% of subjects write down the same words when explaining their choices in the three conditions.

Result 6: Significant proportion of subjects use the same reasoning in the three versions of the beauty contest game.

5. Discussion

The classical beauty contest game (Nagel, 1995) is one of the most influential experimental protocols that have been adopted by numerous experimental papers to measure depth of thinking of decision-makers. As of Nov 2020, the study by Nagel (1995) alone has received 1754 citations in Google scholar.

This paper investigates whether traditional level *k* analysis falsely attributes some sophistication to random players. We propose a very simple two-step procedure to identify the depth of thinking in the beauty contest game. We show that by combining the choices in the two steps (two games), the method gives a very different conclusion, which is suggestive of limitations of the classical method. Using our method, we find that 81% of subjects do not have best response reasoning (which is essential for any positive level of depth of thinking) while the classical method would suggest only 12%. An interpretation of our results is that the classical method falsely attributes some sophistication to random players, and the degree of false attribution is large. More specifically, in the classical method, if the player submits any number equal or below 70, it is not clear the submitted number is based on best response or non-best response reasoning. In the computer version of our beauty contest game, if a subject submits a number higher than 0, he/she will lose the game for sure. As such, assuming that a subject who best respond in the computer version will also best respond in the human version, we can infer whether the subjects have best response reasoning from the choice in the computer version beauty contest game.

[20] Since subject give their reasoning in one game, and then immediately asked to give their reasoning in another, it is possible that there may be contamination of reasoning.

The text analysis of the underlying reasons for the submitted numbers reveals suggestive evidence that the classical method has the weakness of classifying random choice as strategic choice. The analysis suggests that our approach is able to identify most random choices as non-best response.

Our proposed method can identify random choice/non-strategic choice, and thus when combined with the classical beauty contest game will provide a very different conclusion as compared with the estimation of the depth of thinking using the classical game, which is suggestive of limitations of the classical method. Our method adds to the growing literature on addressing measurement errors in experiments using multiple measures (Gillen, et al., 2019). Our result also implies that it is important to obtain multiple measures when eliciting preferences in experiments. Our two-step procedure is also very general and can be easily applied to other games as well to measure whether subjects exhibit strategic reasoning as game theory assumes.

Our method can be applied to investigate a number of interesting questions for future research. For example, it would be interesting to use our method to estimate the proportion of retail investors as well as professional investors with strategic reasoning in the financial market. It would also be interesting to use our method to investigate if the propensity of strategic reasoning varies in different types of games.

Funding Information: This research received research grants support from the Research Grants Council, Hong Kong under grant no. ECS21501915 and the National Natural Science Foundation of China (Grant No. 71973099). The authors have no other relevant financial or non-financial interests to disclose.

Availability of data and materials: Data are available upon request.

References

Agranov, M. and P. Ortoleva. 2017. "Stochastic Choice and Preferences for Randomization." *Journal of Political Economy*, 125(1), 40-68.

Alaoui, L.; K. A. Janezic and A. Penta. 2020. "Reasoning About Others' Reasoning." *Journal of Economic Theory*, 189, 105091.

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

Arad, A. and A. Rubinstein. 2012. "The 11–20 Money Request Game: A Level-K Reasoning Study." *American economic review*, 102(7), 3561-73.

Bosch-Domenech, A.; J. G. Montalvo; R. Nagel and A. Satorra. 2002. "One, Two,(Three), Infinity,...: Newspaper and Lab Beauty-Contest Experiments." *American economic review*, 92(5), 1687-701.

Bosch-Domènech, A.; J. G. Montalvo; R. Nagel and A. Satorra. 2010. "A Finite Mixture Analysis of Beauty-Contest Data Using Generalized Beta Distributions." *Experimental Economics*, 13(4), 461-75.

Bosch-Rosa, C. and T. Meissner. 2020. "The One Player Guessing Game: A Diagnosis on the Relationship between Equilibrium Play, Beliefs, and Best Responses." *Experimental Economics*, 1-19.

Bosch-Rosa, C.; T. Meissner and A. Bosch-Domènech. 2018. "Cognitive Bubbles." *Experimental Economics*, 21(1), 132-53.

Branas-Garza, P.; M. P. Espinosa and P. Rey-Biel. 2011. "Travelers' Types." *Journal of Economic Behavior & Organization*, 78(1-2), 25-36.

Camerer, C. F.; T.-H. Ho and J.-K. Chong. 2004. "A Cognitive Hierarchy Model of Games." *The Quarterly Journal of Economics*, 119(3), 861-98.

Costa-Gomes, M. A. and V. P. Crawford. 2006. "Cognition and Behavior in Two-Person Guessing Games: An Experimental Study." *American economic review*, 96(5), 1737-68.

Crawford, V. P. and N. Iriberri. 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), 1721-70.

Duffy, J. and R. Nagel. 1997. "On the Robustness of Behaviour in Experimental 'Beauty Contest'games." *The Economic Journal*, 107(445), 1684-700.

Frederick, S. 2005. "Cognitive Reflection and Decision Making." *Journal of economic perspectives*, 19(4), 25-42.

Friedenberg, A.; W. Kets and T. Kneeland. 2018. "Is Bounded Rationality Driven by Limited Ability?" *working paper*.

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

Gill, D. and V. Prowse. 2016. "Cognitive Ability, Character Skills, and Learning to Play Equilibrium: A Level-K Analysis." *Journal of Political Economy*, 124(6), 1619-76.

Gillen, B.; E. Snowberg and L. Yariv. 2019. "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study." *Journal of Political Economy*, 127(4), 1826-63.

Grosskopf, B. and R. Nagel. 2008. "The Two-Person Beauty Contest." *Games and Economic Behavior*, 62(1), 93-99.

Gul, F.; P. Natenzon and W. Pesendorfer. 2014. "Random Choice as Behavioral Optimization." *Econometrica*, 82(5), 1873-912.

Ho, T.-H.; C. Camerer and K. Weigelt. 1998. "Iterated Dominance and Iterated Best Response in Experimental" P-Beauty Contests"." *The American Economic Review*, 88(4), 947-69.

Jin, Y. 2018. "Does Level-K Behavior Imply Level-K Thinking?" *working paper*.

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

Kneeland, T. 2015. "Identifying Higher‐Order Rationality." *Econometrica*, 83(5), 2065-79.

Levitt, S. D.; J. A. List and S. E. Sadoff. 2011. "Checkmate: Exploring Backward Induction among Chess Players." *American Economic Review*, 101(2), 975-90.

Li, K. K. 2011. "Preference Towards Control in Risk Taking: Control, No Control, or Randomize?" *Journal of Risk and Uncertainty*, 43(1), 39-63.

Nagel, R. 1995. "Unraveling in Guessing Games: An Experimental Study." *American economic review*, 1313-26.

Nagel, R.; C. Bühren and B. Frank. 2017. "Inspired and Inspiring: Hervé Moulin and the Discovery of the Beauty Contest Game." *Mathematical Social Sciences*, 90, 191-207.

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