Strategic complexity and the value of thinking *

David Gill †
Victoria Prowse ‡

This version: April 15, 2022

Abstract

Response times are a simple low-cost indicator of the process of reasoning in strategic games. In this paper, we leverage the dynamic nature of response-time data from repeated strategic interactions to measure the strategic complexity of a situation by how long people think on average when they face that situation (where we categorize situations according to the characteristics of play in the previous round). We find that strategic complexity varies significantly across situations, and we find considerable heterogeneity in how responsive subjects' thinking times are to complexity. We also study how variation in response times at the individual level across rounds affects strategic behavior and success: when a subject thinks for longer than she would normally do in a particular situation, she wins less frequently and earns less. The behavioral mechanism that drives the reduction in performance is a tendency to move away from Nash equilibrium behavior. Finally, cognitive ability and personality have no effect on average response times.

Keywords: Response time; decision time; deliberation time; thinking time; complexity; level-k; game theory; strategic game; repeated games; beauty contest; cognitive ability; personality.

JEL Classification: C72; C91.

*The first version was titled "Using response times to measure strategic complexity and the value of thinking in games." We thank the Economic Science Laboratory at the University of Arizona (and, in particular, Andreas Blume, Eric Cardella, Ray Chiu, Martin Dufwenberg, James Fisher, Richard Kiser, Quyen Nguyen, and John Wooders) for hosting our experiment. We also thank Peter Wagner, Wei Zhang and Junya Zhou for excellent research assistance. Prowse gratefully acknowledges financial support to run the experiment from a Zvi Meitar – Oxford University Social Sciences Research Grant. The University of Oxford (SSD/CUREC1A/10-079) provided primary Institutional Review Board approval for the experimental data collection. An Online Appendix provides additional materials.

†Department of Economics, Purdue University; dgill.econ@gmail.com.

‡Department of Economics, Purdue University; vprowse@purdue.edu.

1 Introduction

Psychologists have long been interested in choice response or decision times (e.g., Stone, 1960, and Busemeyer and Townsend, 1993). Recently economists have become interested in how the speed of decision making affects behavior and outcomes. Response times are thought to be connected to decision-making style: fast thinking is linked to intuitive or instinctive decision making, while slower thinking is linked to a more deliberate or contemplative mode of thought (e.g., Rubinstein, 2016).[1] An alternative perspective finds that quicker decisions are more accurate in sequential sampling problems with binary decisions (e.g., Fudenberg et al., 2018). Following the psychology literature, one stream of research in economics looks at response times in non-strategic environments: Spiliopoulos and Ortmann (2018) provide a comprehensive survey, while also developing the different benefits and challenges of collecting and analyzing response-time data.[2]

In contrast to the literature that studies response times in non-strategic settings, our focus is on response times in strategic settings. Speed of decision making is of particular importance in strategic settings because of the complexity of the environment: players need to think about the payoff structure ('the rules of the game') and then form beliefs about how others will behave. For a sophisticated agent forming such beliefs involves iterative thinking about how others think about the agent herself, about how others think about how the agent thinks about them, and so forth. The type of reasoning required to perform well in games requires substantial cognitive effort, but with few exceptions (e.g., Alaoui and Penta, 2016a,b, forthcoming) standard theory is silent about how much cognitive effort people exert in strategic settings.[3] In reality, agents vary in their willingness to exert cognitive effort and in how useful cognitive effort is for their performance in strategic settings. According to Rubinstein (2007), response time is a simple low-cost indicator of the process of reasoning in strategic games, with more 'cognitive' choices taking longer than 'instinctive' ones.[4] The speed of decision making in strategic environments has received some recent interest. Experiments on the relationship between response times and strategic behavior have considered how response times relate to: the type of game being played; rates of cooperation in public goods games; behavior in contests; strategies designed to persuade; threshold strategies in global games; and private information in auctions and environments with social learning.[5]

[1] As noted by Rubinstein (2016), this distinction between instinctive and contemplative decision making is in some sense consistent with Kahneman (2011)'s distinction between 'system I' and 'system II' decision processes, although economists generally study response times measured in seconds while psychologists generally study response times measured in fractions of seconds.

[2] Wilcox (1993) finds that lottery choice response times respond positively to incentives. Piovesan and Wengström (2009) and Lohse et al. (2017) find that generosity is associated with longer response times (both across and within-individuals). Achtziger and Alós-Ferrer (2013) link response times to Bayesian updating. Clithero and Rangel (2013) use response times together with choice data to predict choice out of sample. Rubinstein (2013) finds that shorter response times are associated with more mistakes but not with the incidence of behavior inconsistent with standard models of decision making such as expected-utility theory. In a single-agent forecasting task, Moritz et al. (2014) find that very fast or very slow decisions compared to the prediction from a learning model with individual-level heterogeneity perform worse. Hutcherson et al. (2015), Cappelen et al. (2016), Chen and Fischbacher (2016) and Konovalov and Krajbich (2019) show that response times are associated with risk, time and social preferences. Caplin and Martin (2016) find that quick automatic decisions are of lower quality. Alekseev (2019) uses response time as a proxy for effort to separate the effects of cognitive ability and motivation on performance.

[3] A small literature in behavioral economics addresses this issue: see, e.g., Alaoui and Penta (2016a) who present a cost-benefit analysis of the endogenous depth of reasoning in games. Alaoui and Penta (forthcoming) extend the model, Alaoui and Penta (2016b) apply the extended model to Avoyan and Schotter (2020)'s data on the allocation of time across games, and Alaoui et al. (2020) provide experimental support for the model.

[4] In the one-shot p-beauty contest, Rubinstein (2007) argues that choosing 33-34 or 22 is more cognitive, and finds that these choices take longer.

[5] Kuo et al. (2009) and Polonio et al. (2015) find faster response times in coordination games than in dominance-solvable games, while Di Guida and Devetag (2013) find shorter response times in games with focal points and Rand

In particular, we aim to study response times in *repeated* strategic interactions. Strategic interactions of economic interest are often repeated: examples include repeated rounds of job hiring and searching, markets with repeated price or quantity competition, repeated selling of goods via auction and multiple rounds of competition for promotions within firms. As we explain in more detail below, we leverage the dynamic nature of response-time data from repeated strategic interactions to: (i) measure the strategic complexity of a situation by how long people think on average when they face that situation (where we categorize situations according to the characteristics of play in the previous round); and (ii) discover whether variation in thinking time across rounds in which an individual faces the same situation affects strategic behavior and success.

To study response times in repeated strategic interactions we use the experimental data collected by us in Gill and Prowse (2016).[6] In the experiment, 780 student subjects were matched into 260 groups of three players. Each group of three played the p-beauty contest for ten rounds with feedback and no rematching.[7] In each round, the subject whose chosen number was closest to seventy percent of the average of the three numbers chosen by the group members (i.e., $p = 0.7$) won six dollars. The subjects had ninety seconds to make their choice in each round. Gill and Prowse (2016) also measured the subjects' cognitive ability using the Raven test, and measured the personality of a subset of the subjects (the Big Five, grit and a measure of future orientation).

In the p-beauty contest the incentive to undercut the average of the choices drives the equilibrium to the lower bound of the action set: the unique Nash equilibrium is for all players to choose zero. With repetition, choices in experiments move toward the equilibrium (e.g., Nagel, 1995, Ho et al., 1998, Gill and Prowse, 2016). However, the game is well-suited to studying strategic thinking: players who expect others to select non-equilibrium actions often have an incentive to choose away from equilibrium themselves.[8] Real-world parallels of the p-beauty contest include timing games in financial and labor markets: during a bubble or in a job market, there is an advantage to trading or making job offers a little earlier than competitors, but moving too early is costly (in terms of lost profit on the upward wave of the bubble or missing out on new information about job candidates).[9] In some repeated games, some subjects behave in a forward-looking manner by choosing stage-game actions to influence outcomes beyond the current round: Online Appendix II provides different types of evidence that all

et al. (2015) find that response times vary according to whether decisions are implemented with error and intentions are observable. Gneezy et al. (2010) find that response times in the Race Game are longer in losing positions. Arad and Rubinstein (2012) study the relationship between response times and behavior in the Colonel Blotto contest. Glazer and Rubinstein (2012) study the association between response times and behavioral types in a game of persuasion. Rand et al. (2012), Lotito et al. (2013) and Nielsen et al. (2014) find that shorter response times are associated with more cooperation in public goods games, although Evans et al. (2015) find a U-shaped relationship, Krajbich et al. (2015) argue that the direction of the correlation is not robust to changing the relative attractiveness of the selfish and cooperative actions, Recalde et al. (2018) argue that the relationship may reflect mistakes and Nishi et al. (2017) find cross-cultural differences. Schotter and Trevino (2021) look at the relationship between response times and threshold strategies in a global game. Agranov et al. (2015) elicit incentivized choices at multiple points in time and find that 'sophisticated' players decrease their choice with thinking time in a p-beauty contest variant. Turocy and Cason (2015) consider the relationship between signals and response times in auctions. Frydman and Krajbich (forthcoming) find that in a social learning environment people can learn about others' private information from observing their response time. Nishi et al. (2016) and Nishi et al. (2017) find that in repeated social dilemmas response times correlate with the level of previous group cooperation. Spiliopoulos (2018) studies a repeated 2×2 constant-sum game and finds that the win-stay-lose-shift heuristic is associated with faster choices. Brañas-Garza et al. (2017) study the response times of ultimatum-game proposers.

[6]Gill and Prowse (2016) investigate how cognitive ability and personality influence the evolution of play toward Nash equilibrium. The paper does not study response times.

[7]See Nagel et al. (forthcoming) for a history of the beauty contest game.

[8]The group size of three in our data maximizes the number of independent observations, while ensuring that the game remains strategically interesting (when the group size is two, choosing zero is weakly dominant).

[9]Roth and Xing (1994) provide evidence of slow unraveling of the timing of offers in entry-level professional job markets.

support the conclusion that such forward-looking behavior is unlikely to be an important driver of behavior in our finitely repeated beauty contest setting.

We start by analyzing between-subject variation in response times. A few papers find a positive between-subject relationship between response times and success in strategic games.[10] We replicate this finding for our beauty contest game by showing that subjects who think for longer on average win more rounds and choose lower numbers (that is, numbers closer to equilibrium). We find no statistically significant relationship between cognitive ability or personality and average response times (see Proto et al., 2019, for related findings).[11] We also study the relationship between average response times and level-k behavior, finding that higher level-k types (who choose lower numbers) think for longer when the beauty contest is played repeatedly (as we describe in Section 3.4, this result is consistent with Alós-Ferrer and Buckenmaier (2021)'s data from the beauty contest played a single time).

Our first substantive contribution is to leverage the dynamic nature of repeated-game response times to develop a measure of strategic complexity. In our repeated-game setting with fixed groups, the subjects may perceive that the strategic complexity of the situation that they face varies with the characteristics of play in the previous round. Motivated by this observation, we categorize eleven different situations according to the particular subject's earnings in the previous round, the rank-order of the choices of the three group members in the previous round, and whether the group played the Nash equilibrium in that round.[12]

To clarify, we categorize situations using only characteristics of play in the previous round: in Online Appendix III we provide evidence that supports this methodology. For example, we use lasso regressions to show that choices in a particular round depend strongly on characteristics of the group of three subjects' play in the previous round, but depend little on characteristics of play in earlier rounds (after controlling for the characteristics of play in the previous round). Furthermore, our methodology that categorizes situations independently of the number of remaining rounds finds support from the evidence in Online Appendix II that forward-looking behavior is unlikely to be an important driver of behavior in our finitely repeated beauty contest setting. In Online Appendix III.4 we also show that our results are robust when we expand the set of situations (from 11 to 101) by further categorizing situations using the average choice of the three group members in the previous round.

We then measure the strategic complexity of a situation by how long subjects think on average when they face that particular situation. Having developed our measure of strategic complexity, we show that strategic complexity varies significantly across situations. Using average thinking time to measure the strategic complexity of a situation in our repeated-game setting is related to Rubinstein (2016)'s distinction between 'contemplative' and 'instinctive' actions in one-shot games according to

[10] Arad and Rubinstein (2012) find that longer response times are associated with winning more battles in the Colonel Blotto game. Brañas-Garza et al. (2017) find that ultimatum-game proposers who think longer earn more. With hypothetical payoffs, Rubinstein (2016) finds that in a 2×2 zero-sum game, the more contemplative action (that is, the action associated with more thought on average) yields a higher expected payoff (in non-zero-sum games the relationship between contemplative actions and payoffs is not as clear).

[11] We are not aware of other work that measures the relationship between personality and response times in strategic games, except for Proto et al. (2019) who find mostly null results (although groups of high conscientious subjects choose faster). In a public goods game Nielsen et al. (2014) find: (i) a positive relationship between scores in the three-question Cognitive Reflection Test and response time; and (ii) a marginally significant negative relationship between scores in a twenty-question Raven test and response time. However, these relationships were no longer significant when the same game was framed as 'taking' rather than 'giving'.

[12] To give a flavor, we define three of the eleven situations here. Situation 1: the subject won in the previous round with the lowest choice, and the two other subjects chose the same higher number. Situation 2: the subject won in the previous round with an 'intermediate' choice (that is, one of the other subjects lost with a lower number and the other lost with a higher number). Situation 3: the subject lost in the previous round with the highest choice, and the other two subjects chose lower numbers different from each other.

how long subjects think on average before choosing the action: according to Rubinstein (2016), more contemplative actions require more strategic reasoning.[13] The link between complexity and response time is also motivated by recent experimental evidence: Alós-Ferrer and Buckenmaier (2021) find that response times increase in cognitive effort, Avoyan and Schotter (2020) find that the time allocated to a particular game (from a fixed time budget across games) depends on characteristics of the game, and Proto et al. (2019) find that response times vary with strategic considerations.[14]

We also find considerable between-subject heterogeneity in how responsive subjects' thinking times are to changes in strategic complexity: we estimate a two-type mixture regression model and find that one type of subject varies her response times substantially with the strategic complexity of the situation that she faces, while the other type hardly varies her response times at all. Interestingly, strategic sophistication, as measured by the level-k model of boundedly rational thinking, predicts responsiveness to strategic complexity: an increase of one level in the level-k typology predicts an increase of around three percentage points in the probability of being a responsive type-1 subject. However, neither cognitive ability nor personality predict responsiveness to strategic complexity.

These findings shed new light on how subjects allocate cognitive resources in games. Our finding that, in repeated games, thinking time varies across situations provides evidence that subjects respond to the characteristics of the situation that they face when deciding how much cognitive effort to allocate to the situation. However, only one type of subject responds to strategic complexity, which highlights the importance of taking seriously across-subject heterogeneity.

Our second substantive contribution is to study how within-subject variation in response times across rounds affects behavior and success in a repeated-game strategic setting. Specifically, we often observe the same subject facing the same situation more than once, and we can measure whether thinking for longer or for less long than the subject would normally do in that situation affects the subject's choices and her probability of winning the round.[15,16] We find that when a subject thinks for longer than she would normally do in a particular situation, she wins less frequently and earns less. The behavioral mechanism that drives the reduction in performance is a tendency to move away from Nash equilibrium behavior: when the subject thinks for longer than normal she is more likely to increase her choice relative to the previous round and she is less likely to choose the equilibrium number. Interestingly, these results based on within-subject variation contrast with those from between-subject variation noted above, which reveal that subjects who think for longer on average (across situations and rounds) perform better on average.

[13] Rubinstein (2016) further defines a subject's 'contemplative index' to be her propensity to choose contemplative actions across different games.

[14] Proto et al. (2019) find that: (i) in the repeated prisoner's dilemma, defect choices take longer (with similar results in a modified battle-of-the-sexes game; the effect is stronger for higher cognitive ability subjects); and (ii) in the stag-hunt game choices of hare take longer.

[15] Recall that we categorize situations according to the characteristics of play in the previous round (specifically, according to the particular subject's earnings in the previous round, the rank-order of the choices of the three group members in the previous round, and whether the group played the Nash equilibrium in that round).

[16] Some existing papers constrain thinking time (see the survey by Spiliopoulos and Ortmann, 2018). However, in the context of strategic games, constraining players' thinking time does not give a clean measure of how a player's strategic behavior varies in her own thinking time because a player's behavior could also change in anticipation of the time constraint on the behavior of the other players. Furthermore, under a time constraint experimental subjects are not choosing how long to think and the imposed time constraint is often very short. Most closely related to our competitive strategic setting: (i) with a time limit of just fifteen seconds Kocher and Sutter (2006) find slower convergence to equilibrium in a modified beauty contest; (ii) with a time limit of fifteen seconds to both read the instructions and then decide Lindner and Sutter (2013) find fewer sophisticated choices in the 11-20 game (but also choices closer to the mixed Nash equilibrium); and (iii) in a design where subjects have just twenty seconds to search payoff boxes using their mouse and then decide Spiliopoulos et al. (2018) find that time pressure reduces payoff search and induces simpler heuristics in normal-form games.

This negative relationship between response times and performance is consistent with individual subjects finding some decisions harder than others, after controlling for the systematic relationship between situations and average response times that we study in our analysis of the complexity of situations. As we explain below, in the context of Alaoui and Penta (2016a, forthcoming)'s model, this negative relationship is consistent with the higher cost of reasoning associated with decisions that a subject finds harder increasing the subject's response time, while lowering the number of steps of reasoning that the subject completes, which in turn lowers understanding and performance. As a result, this negative relationship between response times and performance does not imply that forcing subjects to think for a shorter amount of time would improve their performance! The negative relationship is also consistent with Fudenberg et al. (2018)'s model of sequential sampling in which quicker decisions are more accurate because they follow clearer evidence about which alternative is best (see also, e.g., Chabris et al., 2009, Krajbich et al., 2014, Echenique and Saito, 2017, and Clithero, 2018, on the negative relationship between response times and utility differences in binary choice decisions, while Woodford, 2014, builds on rational inattention models like that of Sims, 2010, to make predictions about response times in sequential sampling problems).

Broadly speaking, our results are consistent with Alaoui and Penta (2016a, forthcoming)'s model of endogenous depth of reasoning, in which the number of steps of reasoning is determined by a comparison of the cost and value of each additional step. In Alaoui and Penta (forthcoming)'s extension to response times, a higher per-step cost of reasoning increases the response time associated with each step of reasoning, while weakly lowering the number of steps; the total effect on response time is thus indeterminate, but under a "gross substitute" condition total response time increases with the cost of reasoning (we thank a referee for noting this gross substitute condition). First, we find a systematic relationship between the characteristics of situations and the average response time associated with those situations, which is consistent with characteristics of situations changing the cost of reasoning according to the effect of those characteristics on complexity. Second, within situation we find that longer response times are associated with worse performance and a tendency to move away from equilibrium behavior, which is consistent with a higher per-step cost of reasoning increasing response times while lowering the number of steps of reasoning (and hence lowering understanding). Our findings thus fit into a broader experimental literature that supports Alaoui and Penta (2016a, forthcoming)'s model (Goeree and Holt, 2001; Alaoui and Penta, 2016a; Alaoui et al., 2020; Avoyan and Schotter, 2020; Alós-Ferrer and Buckenmaier, 2021; see Alaoui and Penta, forthcoming, for a detailed summary).

We conclude with hope that our findings will spur further empirical and theoretical research on the important topic of response times in games. A better understanding of how and when subjects allocate time and cognitive resources in games will help to refine existing models of boundedly rational thinking in games, such as level-k thinking (Stahl and Wilson, 1994; Nagel, 1995), while also helping to build new models that yield better predictive power.

The paper proceeds as follows: Section 2 describes the experimental design; Section 3 provides descriptive statistics on response times; Section 4 uses response times to measure strategic complexity and explores between-subject heterogeneity in responsiveness to complexity; Section 5 considers whether subjects' characteristics predict heterogeneous responses to complexity; Section 6 studies how individual-level variation in thinking time across rounds relates to behavior and performance; and Section 7 concludes.

2 Experimental design

As explained in the introduction, we use the experimental data collected by us in Gill and Prowse (2016).[17] We ran thirty-seven experimental sessions at the University of Arizona's Experimental Science Laboratory (ESL). Each session lasted approximately seventy-five minutes. In total, 780 student subjects participated in our experiment, with eighteen or twenty-four subjects per session. On average subjects earned twenty United States dollars, on top of a show-up fee of five dollars (subjects were paid privately in cash). Online Appendix I.1 provides the experimental instructions.[18]

Subjects played ten rounds of the p-beauty contest game in fixed groups of three without rematching. In every round each group member privately chose an integer $x \in \{0, 1, ..., 100\}$. We implemented the beauty contest with $p = 0.7$: the group member whose chosen number was closest to seventy percent of the mean of the three numbers chosen by the group members (the 'target') was paid six dollars and the other group members received nothing. In the case of ties, the six dollars was split equally among the subjects who tied. The unique Nash equilibrium is for all players to choose zero.

The subjects had ninety seconds to make their choice in each round. If a subject failed to make a choice within ninety seconds, then a flashing request prompted an immediate choice.[19] While making their choice, the subjects could see a reminder of the rules. At the end of the ninety seconds, all groups advanced together to a feedback stage that lasted thirty seconds. We provided feedback about the group members' choices in that round, seventy percent of the mean of the choices, and the earnings of the group members in that round.

Before the start of the first round, we measured the subjects' cognitive ability using a thirty-minute computerized Raven test, and subjects were matched into groups of three to play the p-beauty contest according to their Raven test score.[20,21] We also measured the personality of 270 of our 780 subjects using an eight-minute questionnaire that was administered before the test of cognitive ability. For these 270 subjects, we measured the Big Five (openness, conscientiousness, extraversion, agreeableness and emotional stability), as well as grit and a measure of future orientation called Consideration of Future of Consequences (CFC).[22] We find a high degree of correlation between our seven measures of personality, which justifies the construction of a smaller number of uncorrelated personality factors. Varimax rotation (Jolliffe, 1995) generates three factors: Factor 1 mainly captures conscientiousness, grit and CFC, Factor 2 mainly captures agreeableness and emotional stability, while Factor 3 mainly captures openness, extraversion and CFC. Gill and Prowse (2016) provide further details, including details of the subject matching by Raven test score and of the personality factor loadings.

[17] See footnote 6.

[18] We drew the participants from the ESL subject pool (which is managed using a bespoke online recruitment system) and we excluded graduate students in economics. We randomized seating positions. We provided the experimental instructions to each subject on their computer screen and we read the instructions aloud (questions were answered privately). The experiment was programmed in z-Tree (Fischbacher, 2007).

[19] We code choices that occurred during the flashing request as having taken exactly ninety seconds. Such choices make up just 0.8% of our observations.

[20] The Raven test is recognized as a leading measure of analytic or fluid intelligence (Carpenter et al., 1990; Gray and Thompson, 2004, Box 1, p.472). We used the Standard Progressive Matrices Plus version of the Raven test, which consists of sixty questions. We did not provide any monetary incentives for completing the Raven test.

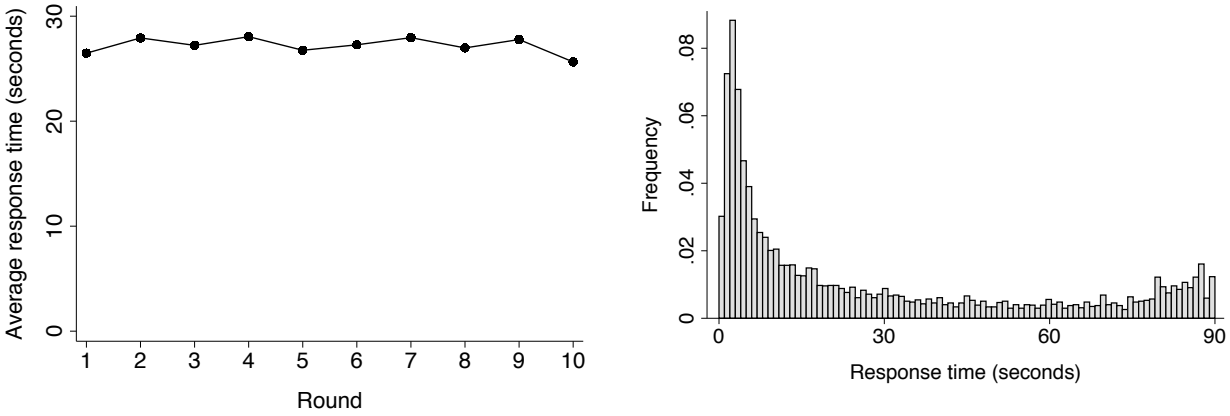
[21] We matched subjects according to whether their Raven test score was in the top half (H) or the bottom half (L) of the test scores in the session, and our design created all possible combinations (groups of three H, groups of three L, mixed groups with two H and one L, mixed groups with one H and two L). Before playing the p-beauty contest, we told each subject whether their own test score was in the top or bottom half of the test scores in the session, and whether each of the subject's opponents' scores was in the top or bottom half. Gill and Prowse (2016) provide full details.

[22] We measured the Big Five using the forty-four-item Big Five Inventory (John et al., 1991; John et al., 2008), grit using the twelve-item Grit Scale (Duckworth et al., 2007), and CFC using the twelve-item CFC Scale (Strathman et al., 1994). 0.3% of the responses are missing (57 of $270 \times 68 = 18,360$). For each question, we replaced any missing responses by the sample average of the non-missing responses to that particular question.

3 Descriptive statistics on response times

3.1 Time trend and heterogeneity in response times

Our sample consists of 780 subjects observed for ten rounds, which gives 7,800 subject-round observations. Figure 1(a) shows that, despite subjects gaining experience with the game, the average response time varies little across rounds of the experiment.[23] Figure 1(b) shows the distribution of the subject-round observations of response time (this captures both within-subject heterogeneity, which arises when a subject varies her response time from one round to the next, and across-subject heterogeneity, which originates from systematic differences in average response times across subjects). We see considerable heterogeneity in the subject-round observations of response time: while around one-quarter of responses occur within the first three seconds of the response window, around five percent of responses occur during the final five seconds of the ninety-second-long response window. Online Appendix IV.1 and Online Appendix IV.2 expand upon these descriptive statistics.



(a) Average response time in rounds 1–10

(b) Distribution of subject-round observations of response time

Figure 1

3.2 Between-subject relationship between response times and success

As we discussed in the introduction, a few papers find a positive between-subject relationship between response times and success in strategic games (see footnote 10). We replicate this finding for our beauty contest game. Specifically, we run a between-subject analysis in which we regress measures of success in the experiment on the subject-level average response time. We consider three measures of success: the fraction of rounds won; earnings per round; and log earnings per round. The first three columns of Table 1 show that a subject's average response time is a predictor of her success: subjects who think for longer on average are more likely to win and earn more. To understand the behavior that underlines these patterns, we regress p-beauty contest choices on the subject-level average response time. The last column of Table 1 shows that subjects who think for longer on average choose lower numbers in the beauty contest, i.e., they choose numbers closer to Nash equilibrium. The results in Table 1 show stability from the first half to the second half of the experiment (see Online Appendix IV.3).

[23] The round-by-round averages all lie within approximately one second of the across-round average of twenty-seven seconds. We test the significance of the trend in the average response time over the experiment by regressing response time on a linear round variable. The two-sided p-value for the coefficient on the linear round variable is 0.656.

	Fraction of rounds won	Earnings per round (cents)	Log earnings per round (cents)	Average choice
Average response time (minutes)	0.060***	38.118***	0.160***	-2.002**
	(0.019)	(9.027)	(0.042)	(0.872)
Intercept	0.402***	182.716***	4.811***	19.929***
	(0.011)	(4.153)	(0.021)	(0.645)
Subjects	780	780	780	780

Notes: Estimates are from OLS regressions. All averages are taken at the subject level. When calculating the fraction of rounds won, a subject is considered to be a winner if she won all or part of the prize. Log earnings per round is calculated by taking the log of earnings at the round level and then averaging at the subject level over rounds. When taking the log of earnings, we add fifty cents to earnings in each round (the show-up fee of five dollars divided by the number of rounds) to avoid taking the log of zero. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table 1: Average response time, success and choices

3.3 Response times and personal characteristics

We also explore how response times are related to cognitive ability and personality by regressing the subject-level average response time on Raven test scores and the three factors that measure personality (Section 2 explained how we constructed the personality factors). Table 2 shows that there is no significant relationship between average response times and Raven test scores or the three personality factors. A subject's average thinking time, therefore, is explained by characteristics or skills that are not captured by our measures of cognitive ability or personality. The results in Table 2 show stability from the first half to the second half of the experiment (see Online Appendix IV.3).

	Average response time (minutes)		
Raven test score	-0.007		0.012
(cognitive ability)	(0.013)		(0.019)
Personality factor 1		-0.027	-0.026
(conscientiousness, grit and future orientation)		(0.022)	(0.022)
Personality factor 2		0.006	0.005
(agreeableness and emotional stability)		(0.018)	(0.019)
Personality factor 3		-0.015	-0.015
(openness, extraversion and future orientation)		(0.021)	(0.021)
Intercept	0.453***	0.442***	0.442***
	(0.013)	(0.019)	(0.019)
Subjects	780	270	270

Notes: Estimates are from OLS regressions. Averages are taken at the subject level. The Raven test score and personality factors 1–3 have been standardized to have means of zero and standard deviations of one (Gill and Prowse, 2016, describe the construction of the personality factors). Personality was measured for 270 of our 780 subjects. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table 2: Raven test score, personality and average response time

3.4 Response times and level-k thinking

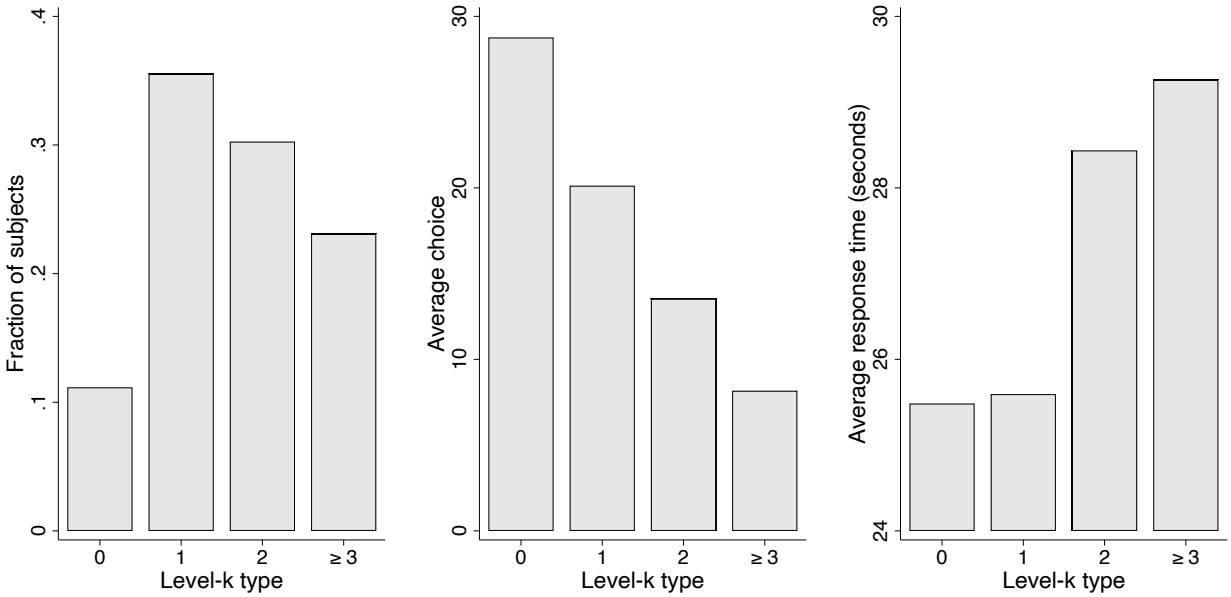
Finally, we study the relationship between response times and level-k thinking. In the typology of the level-k model of boundedly rational thinking, level-0 agents choose according to some strategically unsophisticated rule, while level-$k > 0$ agents choose as if all other agents are of level $k-1$ (Stahl and Wilson, 1994; Nagel, 1995). In a setting where subjects play the beauty contest game a single time, Alós-Ferrer and Buckenmaier (2021) find that higher level-k types think for longer.

Alós-Ferrer and Buckenmaier (2021) use a quadratic distance measure based on choices to allocate each subject to a level-k type. We extend Alós-Ferrer and Buckenmaier (2021)'s quadratic distance methodology to our repeated beauty contest setting by allowing level-k types to learn over time. In particular, building on Nagel (1995) and Gill and Prowse (2016), level-0 agents learn in a strategically unsophisticated adaptive manner, while agents with level-$k > 0$ understand how lower-level agents learn across rounds.[24] We describe the details of our level-k model in Online Appendix V.1.

The left-hand-side panel of Figure 2 shows the proportion of subjects that we allocate to each level-k type, using an aggregate category for levels 3 and above: we find that nearly two-thirds of subjects are level-1 or level-2 types. The middle panel of Figure 2 shows that higher level-k types choose lower numbers: this relationship follows mechanically from the fact that, in the beauty contest game, level-$k > 0$ types undercut the choices of types one level below them. The right-hand-side panel of Figure 2 shows that higher level-k types think for longer: in particular, types of level-2 and above think for around three seconds longer on average than level-0 and level-1 types. The second and third

[24] See also Stahl (1996), Duffy and Nagel (1997), and Ho et al. (1998). We need to include learning in the model in order to capture the tendency of choices to move toward zero across rounds (see Figure OA.1 in Online Appendix IV.2).

columns of Table 3 show that this relationship is statistically significant ($p < 0.05$ when we control for cognitive ability in the third column). Our findings here that higher level-k types choose lower numbers and think for longer help to explain why subjects who think for longer on average choose lower numbers (see the last column of Table 1).



Notes: See Online Appendix V.1 for details of the level-k learning model. As we explain there, we allocate each subject to a level-k type using quadratic distances in rounds 2 to 10, and so here we use data from those same rounds.

Figure 2: Frequency and behavior of level-k types

	Level≥2	Average response time (minutes)		
Raven test score (cognitive ability)	0.102***		-0.013	-0.014
	(0.019)		(0.014)	(0.020)
Level≥2		0.054*	0.059**	0.059**
		(0.029)	(0.029)	(0.029)
Level≥2 × Raven test score				0.002
				(0.027)
Intercept	0.533***	0.426***	0.423***	0.423***
	(0.023)	(0.019)	(0.019)	(0.019)
Subjects	780	780	780	780

Notes: As in Table 2: estimates are from OLS regressions; averages are taken at the subject level; and Raven test scores are standardized. "Level≥2" is an indicator for the subject being classified as a level-2 or higher type. See Online Appendix V.1 for details of the level-k learning model: as we explain there, we allocate each subject to a level-k type using quadratic distances in rounds 2 to 10, and so here we use data from those same rounds. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table 3: Level-k type, Raven test score and average response time

The subjects that we allocate to lower level-k types might choose faster because they engage in less complex reasoning or follow simple heuristics. Indeed, Fehr and Huck (2016) find evidence that some subjects fail to think strategically in the beauty contest: they choose numbers from the entire interval, and their choices are not associated with their beliefs about the cognitive ability of others (these subjects also score lower on the Cognitive Reflection Test).

Turning now to cognitive ability, the first column of Table 3 shows a strong positive relationship between cognitive ability and level-k: a one-standard-deviation increase in cognitive ability is associated with a ten-percentage-point increase in the probability of being classified as a level-2 or higher type. Since higher level-k types think for longer on average (see the previous paragraph), higher cognitive ability indirectly increases response times through the effect of cognitive ability on level-k. However, the third column of Table 3 shows that the direct effect of cognitive ability on average response time, holding level-k fixed, is negative (although not statistically significant), which explains why the total effect of cognitive ability on average response time is small and not statistically significant (see the first column of Table 2). Finally, the close-to-zero coefficient on the interaction term in the final column of Table 3 shows that the effect of level-k on average response time does not depend on cognitive ability.

Our model of level-k learning is simpler than that of Gill and Prowse (2016) (who allow subjects to shift levels over time), but we confirm their result that cognitive ability predicts level-k behavior when the beauty contest is played repeatedly. As we note above, in their setting where subjects play the beauty contest a single time, Alós-Ferrer and Buckenmaier (2021) find that higher level-k types think for longer (since higher level-k types mechanically choose lower numbers, they also find a negative correlation between choices and response times). In this section, we show that this relationship between level-k thinking and response times also holds when the beauty contest is played repeatedly.[25]

Finally, we consider personality. Table OA.13 in Online Appendix V.2 shows that: (i) the effects of cognitive ability and level-k on average response time in Table 3 are robust when we control for personality; and (ii) the result from Table 2 in Section 3.3 that there is no statistically significant relationship between average response time and personality is robust when we control for level-k.

[25] Alós-Ferrer and Buckenmaier (2021) measure cognitive ability using the Cognitive Reflection Test: like us, they do not find a statistically significant effect of cognitive ability on response times when controlling for level-k, but unlike us, they find a positive interaction between cognitive ability and level-k. Remarkably, in a one-shot beauty contest, Coibion et al. (2020) find that subjects who think for at least 20 seconds clump at the exact choices predicted by the level-k model (33, 22, and so on). In a repeated beauty contest setting, Chen et al. (2014) regress level-k on working memory and response times, and find a positive coefficient on response times; however, they do not report the unconditional correlation between levels and response times.

4 Strategic complexity

We now turn to our first substantive contribution: we use the dynamic nature of the repeated response times in our beauty contest game to measure the strategic complexity of the situations that our subjects face and to explore the heterogeneity in how subjects adjust their thinking time in response to changes in strategic complexity. We proceed in three stages: first, we categorize situations according to the characteristics of play in the previous round; second, we measure the strategic complexity of each situation by how long subjects think on average when they face that situation; and finally, we explore heterogeneity in how subjects respond to changes in strategic complexity.

4.1 Categorizing situations

In our repeated-game setting with fixed groups, subjects may perceive that the strategic complexity of the situation that they face in a given round varies with the characteristics of play in the previous round. Motivated by this observation, we categorize situations according to: (i) the subject's earnings in the previous round; (ii) the rank-order of the choices of the three group members in the previous round; and (iii) whether the group played the Nash equilibrium in the previous round. Let n denote the number chosen by the subject in the previous round, and let $\underline{n}^o$ and $\overline{n}^o$ denote the numbers chosen in the previous round by the subject's opponents. Without loss of generality, we order the opponents such that $\underline{n}^o \leq \overline{n}^o$. The subject's situation in a given round is determined by: (i) the subject's earnings in the previous round, (ii) the ordering of n, $\underline{n}^o$ and $\overline{n}^o$; and (iii) whether all three choices were zero (the Nash equilibrium) in the previous round. We focus on the eleven situations described in the first column of Table 4. The second column of Table 4 shows that we have between 123 and 1,739 subject-round observations for each situation, giving a total of 7,012 subject-round observations across the eleven situations.[26] We order the situations by strategic complexity: the third column of Table 4 reports our measure of strategic complexity, which we explain in Section 4.2.

4.2 Measuring strategic complexity

We measure the strategic complexity of a situation by how long subjects think on average when they face that particular situation. To identify the influence of the situation on average response times we cannot simply measure the average response time in each situation because subjects vary systematically in how long they think and subjects who tend to think for longer might be more likely to face certain situations. Thus, we leverage the repeated observations of response times to separate the component of response times that is attributable to the situation from the component of response times that is due to systematic differences in thinking times between subjects. We do this by running the following fixed-effects regression for subject i's response time in seconds in round r:

$$\text{ResponseTime}_{i,r} = \sum_s \theta_s D^s_{i,r} + \alpha_i + \tau_r + \varepsilon_{i,r} \quad \text{for} \quad r = 2, ..., 10, \tag{1}$$

[26] We categorize situations using only specific characteristics of play in the previous round: Online Appendix III provides evidence that supports this methodology. Given our methodology for categorizing situations, these eleven situations represent all situations with two exceptions: the situation where $n < \underline{n}^o < \overline{n}^o$ and the subject earned three dollars (the subject tied with the opponent who chose $\underline{n}^o$); and the situation where $\underline{n}^o < n < \overline{n}^o$ and the subject earned three dollars (again the subject tied with the opponent who chose $\underline{n}^o$). We omit these two situations because we have only four subject-round observations from each of the two situations, giving eight omitted observations. Subtracting these eight observations from the 7,020 subject-round observations from the second round onward gives our total of 7,012. We cannot use the 780 response-time observations from the first round since we categorize situations according to the characteristics of play in the previous round.

where $D^s_{i,r}$ is an indicator for subject i being in situation s in round r, α_i is a subject fixed effect that absorbs systematic between-subject difference in response times, τ_r is a round fixed effect that captures any trends in response times over the experiment (round five is the reference category), and $\varepsilon_{i,r}$ is an error term.

When estimating (1), one situation is arbitrarily chosen to be the reference category ($s = 1$; $\theta_1 = 0$), and the parameter θ_s then measures the pure effect on response time of being faced with situation s instead of this reference situation. Using estimates from the fixed-effects regression described in (1), we calculate the strategic complexity of situation s as follows:

$$\text{StrategicComplexity}_s \equiv \hat{\theta}_s + \overline{\hat{\alpha}_i}, \tag{2}$$

where $\hat{\theta}_s$ denotes the estimate of θ_s and $\overline{\hat{\alpha}_i}$ is the average of the estimates of the subject fixed effects.

The third column of Table 4 reveals that strategic complexity, measured by average thinking time as detailed in (2), varies substantially across the eleven situations. The most complex situation is where the subject won the entire prize of six dollars in the previous round with a choice that was between that of her opponents; in this situation subjects think for an average of thirty seconds. The least complex situation is where the group was at the Nash equilibrium in the previous round (that is, all three group members chose zero); in this situation subjects think for an average of fifteen seconds. The variation in strategic complexity across situations is statistically significant: the null hypothesis that strategic complexity is constant across the eleven situations is strongly rejected (an F-test returns $p = 0.000$); and this result holds whether or not we include the situation in which the group was at the Nash equilibrium in the previous round (see Section 4.4). Online Appendix VI provides further discussion of the results in Table 4.

Situation	Subject-round observations	Strategic Complexity (Average response time in seconds)
$n = \underline{n}^o = \overline{n}^o = 0$ ($\therefore$ subject earned \$2)	318	15.398
$\underline{n}^o < \overline{n}^o < n$ ($\therefore$ subject earned \$0)	1,739	25.249
$\underline{n}^o = \overline{n}^o < n$ ($\therefore$ subject earned \$0)	273	25.426
$n < \underline{n}^o < \overline{n}^o$ & subject earned \$0	633	26.064
$\underline{n}^o < n < \overline{n}^o$ & subject earned \$0	1,102	26.373
$n = \underline{n}^o = \overline{n}^o > 0$ ($\therefore$ subject earned \$2)	123	26.680
$\underline{n}^o < n = \overline{n}^o$ ($\therefore$ subject earned \$0)	362	26.775
$n < \underline{n}^o = \overline{n}^o$ ($\therefore$ subject earned \$6)	181	27.027
$n < \underline{n}^o < \overline{n}^o$ & subject earned \$6	1,102	27.984
$n = \underline{n}^o < \overline{n}^o$ ($\therefore$ subject earned \$3)	546	28.447
$\underline{n}^o < n < \overline{n}^o$ & subject earned \$6	633	30.284

Notes: n denotes the number chosen by the subject in the previous round, and $\underline{n}^o$ and $\overline{n}^o$ denote the numbers chosen in the previous round by the subject's opponents (with $\underline{n}^o \leq \overline{n}^o$). Earnings refer to the subject's earnings in the previous round. We have a total of 7,012 subject-round observations across the eleven situations (see footnote 26). The third column reports strategic complexity as described in (2).

Table 4: Strategic complexity

4.3 Heterogeneity

We now study whether subjects vary in how they adjust thinking time in response to changes in strategic complexity. We postulate that some subjects may be sensitive to changes in strategic complexity and choose to devote more time to thinking in more complex situations. Other subjects may not tailor their thinking time to the complexity of their situation, either because they fail to recognize that situations vary in complexity or because they lack the self-control needed to think for longer.[27]

We explore the empirical support for this reasoning by estimating the following two-type mixture regression model, which captures heterogeneity in how thinking time responds to changes in strategic complexity:

$$\text{ResponseTime}'_{i,r} = \beta_i \text{StrategicComplexity}'_{i,r} + \upsilon_r + \sigma_i \epsilon_{i,r} \quad \text{for} \quad r = 2, ..., 10, \tag{3}$$

where the primes in (3) denote variables expressed in deviation form (i.e., differenced relative to the average for subject i) with the deviation form of any variable $X_{i,r}$ defined as

$$X'_{i,r} \equiv X_{i,r} - \frac{\sum_{r=2}^{10} X_{i,r}}{9}, \tag{4}$$

and where

$$\text{StrategicComplexity}_{i,r} \equiv \sum_s \text{StrategicComplexity}_s \times D^s_{i,r} \tag{5}$$

denotes the strategic complexity of the situation facing subject i in round r, υ_r is a round fixed effect (round five is the reference category), and $\epsilon_{i,r}$ is an independent error term with a standard normal distribution (see Section 4.2, and in particular (1) and (2), for the definitions of $\text{StrategicComplexity}_s$ and $D^s_{i,r}$). The parameter β_i describes how the subject's response time changes when strategic complexity deviates from the average strategic complexity faced by the subject, and σ_i measures the standard deviation of the component of the subject's response time that is unresponsive to changes in strategic complexity. We distinguish two types of subjects: type 1 subjects have $[\beta_i, \sigma_i] = [\beta_1, \sigma_1]$ and type 2 subjects have $[\beta_i, \sigma_i] = [\beta_2, \sigma_2]$. The probability of a subject being of type j is π_j for $j = 1, 2$.

Table 5 reports the parameter estimates for this two-type mixture regression model. We find that the thinking times of the two types respond very differently to changes in strategic complexity. Around sixty percent of subjects are type-1 subjects who increase thinking time substantially when strategic complexity increases (specifically, when the strategic complexity of the situation increases by one second, the response time of type-1 subjects increases by around 1.3 seconds, and this effect is significant at the one-percent level). In contrast, type-2 subjects, who make up around forty percent of the subject population, hardly respond at all to changes in the complexity of their situation: the estimated effect size for type-2 subjects is around one-eighth of that for type-1 subjects. Interestingly, the estimate of σ is larger for type-1 subjects, which means that type-1 subjects also have larger variations in response times for reasons that are unrelated to changes in strategic complexity.

Table 6 reports the parameter estimates for the three-type version of the two-type mixture regression model described above. We continue to find that thinking time responds strongly to changes in strategic complexity for the majority of subjects, while thinking time responds little to complexity for the other subjects.

[27] Since we measure the strategic complexity of a situation by how long subjects think on average when they face that particular situation, and since we find that complexity varies across situations, we must find that the average subject's response time is sensitive to changes in complexity. However, our specification does not imply anything about the extent to which subjects vary in how they adjust thinking time to changes in strategic complexity. Thus, there is no inherent circularity in our definition of strategic complexity or in our analysis of heterogeneity.

	Type 1	Type 2
β	1.311***	0.165***
	(0.179)	(0.057)
σ	23.093***	5.844***
	(0.419)	(0.326)
π (type probability)	0.634***	0.366***
	(0.023)	(0.023)

Notes: Parameter estimates were obtained by applying Maximum Likelihood estimation to the sample of 7,012 subject-round observations described in Table 4. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests for β and one-sided tests elsewhere).

Table 5: Heterogeneous effect of strategic complexity on response time: two-type model

	Type 1	Type 2	Type 3
β	1.427***	0.240**	0.022
	(0.207)	(0.108)	0.036
σ	23.868***	7.671***	2.140***
	(0.402)	(0.423)	(0.202)
π (type probability)	0.581***	0.313***	0.106***
	(0.023)	(0.019)	(0.014)

Notes: Parameter estimates were obtained by applying Maximum Likelihood estimation to the sample of 7,012 subject-round observations described in Table 4. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests for β and one-sided tests elsewhere).

Table 6: Heterogeneous effect of strategic complexity on response time: three-type model

4.4 Robustness of results on strategic complexity

Table 4 shows that the complexity of the situation in which a subject's group played the Nash equilibrium in the previous round (the 'equilibrium' situation) is substantially lower than that of the other ten situations (the 'non-equilibrium' situations). In the Appendix we show that our results on strategic complexity are not driven by differences between the behavior of subjects in the equilibrium situation and subjects in the non-equilibrium situations. In particular, Table A.1 shows that our estimates of the complexities of the ten non-equilibrium situations reported in Table 4 are essentially unchanged if we re-estimate these complexities using a sample that excludes the 318 subject-round observations where the subject's group played the equilibrium in the previous round. The notes to Table A.1 also report statistically significant variation in strategic complexity across the ten non-equilibrium situations. Tables A.2 and A.3 show that our results on the heterogeneous effect of strategic complexity on thinking time reported in Section 4.3 are also robust to excluding the 318 subject-round observations where the subject's group played the equilibrium in the previous round.

5 Predicting heterogeneous responses to complexity

In Section 4.3 we found considerable heterogeneity in the degree to which subjects respond to strategic complexity: the response times of type-1 subjects increase substantially with the strategic complexity of the situation that they face, while type-2 subjects hardly respond to changes in complexity, with the responsive type-1 subjects making up around sixty percent of the subject population.

In Table 7 we study whether subjects' level-k type or cognitive ability predict the likelihood of being a responsive type-1 subject or an unresponsive type-2 subject (Section 3.4 introduces our model of level-k boundedly rational thinking). Panel A shows that when we include level-k type and cognitive ability in the mixture model from Section 4.3, the parameters that describe the two types are stable (compared to Table 5). Panel B shows that level-k thinking predicts responsiveness to strategic complexity: an increase of one level in the level-k typology predicts an increase of around three percentage points in the probability of being a responsive type-1 subject (this result holds whether or not we control for cognitive ability). By contrast, we also see from Panel B that cognitive ability does not predict responsiveness to strategic complexity.

	Model 1		Model 2		Model 3	
	Type 1	Type 2	Type 1	Type 2	Type 1	Type 2
Panel A: Heterogeneous effect of strategic complexity on response time						
β	1.311***	0.165***	1.308***	0.166***	1.308***	0.166***
	(0.179)	(0.057)	(0.178)	(0.057)	(0.179)	(0.057)
σ	23.094***	5.844***	23.090***	5.839***	23.093***	5.842***
	(0.420)	(0.327)	(0.416)	(0.321)	(0.416)	(0.321)
Average π	0.633***	0.367***	0.634***	0.366***	0.634***	0.366***
(average type probability)	(0.023)	(0.023)	(0.022)	(0.022)	(0.022)	(0.022)
Panel B: Average marginal effects on type probabilities						
Raven test score	0.000	0.000			-0.010	0.010
(cognitive ability)	(0.367)	(0.367)			(0.019)	(0.019)
Level-k type			0.031**	-0.031**	0.033**	-0.033**
			(0.014)	(0.014)	(0.014)	(0.014)

Notes: Here we extend the two-type mixture regression model from Table 5 by allowing the type probability to depend on the subject's traits (Raven test score in Model 1, level-k type in Model 2, or both in Model 3) according to a logistic distribution function (see Online Appendix V.1 for details of how we allocate subjects to level-k types; Raven test scores are standardized). The average type probability was obtained by computing the type probability for each subject, conditional on the subject's trait(s), and then averaging over the subjects. The average marginal effects are averages of the individual-level marginal effects. As in Table 5, parameter estimates were obtained by applying Maximum Likelihood estimation to the sample of 7,012 subject-round observations described in Table 4. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests for β and the average marginal effects, and one-sided tests elsewhere).

Table 7: Level-k type, Raven test score, and the heterogeneous effect of complexity on response time

We conclude that strategic sophistication, as measured by the level-k model of boundedly rational thinking, predicts responsiveness to strategic complexity. This relationship holds true even though we allocate each subject to a single level-k type based on their choices in the repeated p-beauty contest, while responsiveness to complexity is measured by within-subject variation in response times across situations of varying strategic complexity. However, neither cognitive ability nor personality predict responsiveness to strategic complexity (Table OA.14 in Online Appendix V.2 shows that: (i) there is no statistically significant relationship between personality and responsiveness; and (ii) the results in Table 7 are robust when we control for personality).

In Table 8 we study whether choices predict the likelihood of being a responsive type-1 subject or an unresponsive type-2 subject. Panel A shows that when we include choices in the mixture model from Section 4.3, the parameters that describe the two types are stable (compared to Table 5). Panel B shows that choices predict responsiveness to strategic complexity: a one-standard-deviation increase in the choice metrics predict a decrease of around three to five percentage points in the probability of being a responsive type-1 subject. The effect of a subject's average choice is not quite statistically significant ($p = 0.149$ in Model 1). However, when we normalize a subject's choice in a particular round relative to the behavior of others, the effects become statistically significant (in Model 2 relative to the round-specific mean of all 780 subjects' choices ($p = 0.088$); in Model 3 relative to the entire distribution of other subjects' choices in that round ($p = 0.013$)).

In Table OA.21 in Online Appendix X we find no evidence that success predicts responsiveness to strategic complexity (all the coefficients in Panel B are far from statistical significance). However, when interpreting this result we should bear in mind that success is a noisy measure of strategic ability.

	Model 1		Model 2		Model 3	
	Type 1	Type 2	Type 1	Type 2	Type 1	Type 2
Panel A: Heterogeneous effect of strategic complexity on response time						
β	1.309***	0.166***	1.309***	0.165***	1.307***	0.166***
	(0.178)	(0.057)	(0.178)	(0.057)	(0.178)	(0.057)
σ	23.088***	5.837***	23.084***	5.834***	23.085***	5.834***
	(0.418)	(0.323)	(0.418)	(0.323)	(0.416)	(0.321)
Average π	0.634***	0.366***	0.634***	0.366***	0.634***	0.366***
(average type probability)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)
Panel B: Average marginal effects on type probabilities						
Average choice (standardized)	-0.026	0.026				
	(0.018)	(0.018)				
Average relative choice			-0.030*	0.030*		
(standardized)			(0.018)	(0.018)		
Average percentile of choice					-0.046**	0.046**
(standardized)					(0.018)	(0.018)

Notes: Here we extend the two-type mixture regression model from Table 5 by allowing the type probability to depend on metrics of the subject's choices according to a logistic distribution function. All choice metrics are subject-level averages of the relevant variable over rounds 2–10 (this matches the estimation of the mixture model that uses response times only from round 2 onward because situations, and thus complexity, are defined only from round 2 onward). A subject's relative choice in round r is the subject's choice divided by the mean of all 780 subjects' choices in that round. A subject's percentile of choice in round r is the percentile of the subject's choice in the distribution of all 780 subjects' choices in that round. The average type probability was obtained by computing the type probability for each subject, conditional on the relevant metric of the subject's choices, and then averaging over the subjects. The average marginal effects are averages of the individual-level marginal effects. As in Table 5, parameter estimates were obtained by applying Maximum Likelihood estimation to the sample of 7,012 subject-round observations described in Table 4. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests for β and the average marginal effects, and one-sided tests elsewhere).

Table 8: Choices and the heterogeneous effect of complexity on response time

6 Thinking for longer than normal

In Section 4 we categorized eleven situations according to the characteristics of play in the previous round and we measured the strategic complexity of a situation by how long subjects think on average when they face that particular situation. In Section 3 we showed that subjects who think for longer on average (across situations and rounds) are more successful. We now turn to a more subtle question: is a subject more successful when she thinks for longer than she would normally do in a particular situation?

6.1 Estimation strategy and results

We explore whether a subject is more successful when she thinks for longer than she would normally do in a particular situation by exploiting within-subject variation in response times across rounds in which the subject faces the same situation. We often observe the same subject facing the same situation more than once, and we can measure whether thinking for longer or for less long than the subject would normally do in that situation affects the subject's success.

Mirroring our analysis in Section 3, we consider three measures of subject i's success in round r: being a winner of the round; earnings in the round; and the log of earnings in the round. Specifically, for each measure of success we run the following fixed-effects regression, which describes how a subject's success in a particular round depends on her thinking time in that round:

$$\text{Success}_{i,r} = \lambda \text{ResponseTime}_{i,r} + \eta_{i,s} + \gamma_r + e_{i,r}, \quad \text{for} \quad r = 2, ..., 10, \tag{6}$$

where $\eta_{i,s}$ is a subject-situation fixed effect, γ_r is a round fixed effect (round five is the reference category) and $e_{i,r}$ is an error term. The subject-situation fixed effects absorb subject-specific systematic differences in success across situations. In particular, they absorb the effect of systematic differences in thinking times across situations at the subject level that are correlated with the subject's success. The parameter λ is thus identified from within-subject variation in response times across rounds in which the subject faced the same situation. It follows that we can interpret our estimate of λ as the effect of a subject thinking for longer than she ordinarily does in a particular situation (and $-\lambda$ as the effect of thinking for less long than normal).

Table 9 reports parameter estimates from the fixed-effects regression described in (6). The table shows that thinking for longer than normal is associated with worse performance. In particular, when a subject thinks for one minute longer than she would normally do in a particular situation, the probability that the subject wins the round decreases by almost six percentage points. This translates into a reduction in earnings in the round of around twenty-seven cents or thirteen percent.

As we discuss in the introduction, this negative relationship between response times and performance is consistent with individual subjects finding some decisions harder than others, after controlling for the systematic relationship between situations and average response times that we study in our analysis of the complexity of situations.

	Winner of round	Earnings in round	Log earnings in round
Effect of thinking longer than normal (minutes)	-0.059** (0.028)	-27.262* (15.003)	-0.132** (0.067)
Subject-round observations	4,774	4,774	4,774

Notes: We start with the sample of 7,012 subject-round observations described in Table 4. From this sample, we then use the subject-round observations for which the subject is observed in the same situation in at least one other round (this ensures that subject-round observations are not fully absorbed by the fixed effects). This gives us 4,774 subject-round observations and 1,763 subject-situation fixed effects. A subject is considered to be a winner if she won all or part of the prize. When taking the log of earnings, we add fifty cents to earnings in each round (the show-up fee of five dollars divided by the number of rounds) to avoid taking the log of zero. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table 9: Effect of thinking for longer than normal on success

We explore the behavioral mechanism that drives the reduction in performance by running fixed-effects regressions as described by (6), but with dependent variables based on the subject's choice in the round. The behavioral mechanism that drives the reduction in performance is a tendency to move away from Nash equilibrium behavior: Table 10 shows that when a subject thinks for longer than normal, that subject is more likely to increase her choice relative to her choice in the previous round and less likely to choose the equilibrium action of zero. In more detail, when a subject thinks for one minute longer than normal, the probability that her choice increases relative to the previous round goes up by six percentage points and the probability that she chooses the equilibrium action falls by three percentage points. Thinking more than normal also increases the subject's choice in the round, but this effect is noisy and therefore insignificant.

	Choice in round	Choose zero in round	Increased choice in round
Effect of thinking longer than normal (minutes)	0.990 (1.025)	-0.027** (0.013)	0.057*** (0.019)
Subject-round observations	4,774	4,774	4,774

Notes: 'Choose zero in round' is an indicator for the subject having chosen the equilibrium action of zero. 'Increased choice in round' is an indicator for the subject's choice in the round being greater than her choice in the previous round. Also see notes to Table 9. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table 10: Effect of thinking for longer than normal on strategic behavior

In Section 4, we measure the strategic complexity of situations by measuring how response times vary across situations on average, controlling for systematic differences in response times between subjects. In this section, instead, the results are identified from within-subject variation in response times across rounds in which the subject faced the same situation. As a result, the results in this section do not follow directly from our analysis of the strategic complexity of situations in Section 4.

6.2 Robustness of results on thinking for longer than normal

A potential concern is that our results could be biased in the presence of a relationship between thinking time and experience. For example, subjects who have little experience of a situation might also think for longer and perform worse. We address this concern by showing that our results are in fact robust to including controls for the subject's experience of her current situation. Tables A.4 and A.5 in the Appendix present the details and results of this robustness exercise.

Even though we have included round fixed effects that control for trends over rounds, a second potential concern is that our results could be biased in the presence of heterogeneous round trends. For example, some subjects might exhibit declining response times and improving performance over rounds, while other subjects might exhibit constant or increasing response times and worsening performance over rounds. We address this concern by showing that our results are in fact robust to including controls for heterogeneous round trends. Tables A.6 and A.7 in the Appendix present the details and results of this second robustness exercise.

In Section 4.4 we showed that our results on strategic complexity are not driven by differences between behavior in the equilibrium situation (that is, after the subject's group played the equilibrium in the previous round) and behavior in the ten non-equilibrium situations. Tables A.8 and A.9 in the Appendix show that our results on the effects of thinking for longer than normal are also robust to excluding the equilibrium situation.

7 Conclusion

In this paper we extend a recent line of research that uses response times to better understand economic behavior and preferences. To give two examples, Clithero and Rangel (2013) use response times to help predict out-of-sample behavior, while Hutcherson et al. (2015) relate response times to social preferences. We study response times in strategic interactions. Response times in games can allow researchers to gain insight into subjects' reasoning processes that choices alone cannot (Rubinstein, 2016), and subjects themselves can potentially use information about the response times of others to infer information that choices alone do not reveal (Frydman and Krajbich, forthcoming).

We use experimental data on response times from repeated games to develop a measure of the strategic complexity of a situation based on how long subjects think on average when they face that situation, where situations are defined according to the characteristics of play in the previous round. Our finding that strategic complexity varies significantly across situations provides evidence that subjects respond to the characteristics of the situation that they face when deciding how much cognitive effort to allocate to their decision. But not all subjects do this: one type of subject responds strongly to strategic complexity, while another type hardly responds at all. We also leverage our response-time data from repeated strategic interactions to show that when subjects think for longer than they would normally do in a particular situation, they perform less well. The behavioral mechanism that drives the reduction in performance is a tendency to move away from Nash equilibrium behavior.

A nascent theoretical literature attempts to model how agents allocate cognitive resources in strategic situations (e.g., Alaoui and Penta, 2016a). Our experimental results suggest that the willingness to engage in strategic reasoning varies systematically with the characteristics of the situation that agents face. We hope that our findings will help to improve the predictive power of existing models of boundedly rational thinking in games, while also inspiring new empirically-grounded models that incorporate explicitly the choice of how hard to think in strategic interactions.

Appendix

Situation	Subject-round observations	Strategic Complexity (Average response time in seconds)
$\underline{n}^o < \overline{n}^o < n$ ($\because$ subject earned \$0)	1,739	25.092
$\underline{n}^o = \overline{n}^o < n$ ($\because$ subject earned \$0)	273	25.526
$n < \underline{n}^o < \overline{n}^o$ & subject earned \$0	633	26.028
$\underline{n}^o < n < \overline{n}^o$ & subject earned \$0	1,102	26.286
$n = \underline{n}^o = \overline{n}^o > 0$ ($\because$ subject earned \$2)	123	26.295
$\underline{n}^o < n = \overline{n}^o$ ($\because$ subject earned \$0)	362	26.806
$n < \underline{n}^o = \overline{n}^o$ ($\because$ subject earned \$6)	181	27.333
$n < \underline{n}^o < \overline{n}^o$ & subject earned \$6	1,102	27.998
$n = \underline{n}^o < \overline{n}^o$ ($\because$ subject earned \$3)	546	28.277
$\underline{n}^o < n < \overline{n}^o$ & subject earned \$6	633	30.272

Notes: We start with the sample described in Table 4. From this sample, we exclude the equilibrium situation: that is, we exclude the 318 subject-round observations where the subject's group played the equilibrium in the previous round. This gives us a total of 6,694 subject-round observations across the ten non-equilibrium situations. The third column reports strategic complexity as described in (2). See the notes to Table 4 for notational definitions. The null hypothesis that strategic complexity is constant across the ten non-equilibrium situations is strongly rejected (an F-test returns $p = 0.001$).

Table A.1: Robustness of the results in Table 4 to excluding the equilibrium situation

	Type 1	Type 2
β	1.585***	0.132
	(0.287)	(0.091)
σ	23.099***	5.865***
	(0.443)	(0.337)
π (type probability)	0.627***	0.373***
	(0.023)	(0.023)

Notes: Parameter estimates were obtained by applying Maximum Likelihood estimation to the sample of 6,694 subject-round observations described in Table A.1. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests for β and one-sided tests elsewhere).

Table A.2: Robustness of the results in Table 5 to excluding the equilibrium situation

	Type 1	Type 2	Type 3
β	1.737***	0.116	0.006
	(0.312)	(0.137)	(0.065)
σ	23.894***	7.697***	2.151***
	(0.398)	(0.376)	(0.201)
π (type probability)	0.573***	0.317***	0.109***
	(0.022)	(0.019)	(0.014)

Notes: Parameter estimates were obtained by applying Maximum Likelihood estimation to the sample of 6,694 subject-round observations described in Table A.1. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests for β and one-sided tests elsewhere).

Table A.3: Robustness of the results in Table 6 to excluding the equilibrium situation

	Winner of round	Earnings in round	Log earnings in round
Effect of thinking longer than normal (minutes)	-0.057** (0.028)	-26.800* (15.086)	-0.129* (0.067)
Subject-round observations	4,774	4,774	4,774

Notes: We run the same regressions on the same sample as in Table 9, adding controls for the subject's experience of her current situation by including the indicator functions $\mathbf{1}_{\{j\}}(\chi_{i,r})$ for $j = 1, ..., 6$, where $\chi_{i,r} \in \{0, 1, ..., 6\}$ is equal to the number of times that subject i experienced her current situation prior to round r (zero experience is the omitted category). No subject experienced her current situation in eight previous rounds (the maximum). One subject experienced her current situation in seven previous rounds: we set $\chi_{i,r}$ equal to six for this observation to ensure that the observation continues to contribute to the estimation. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table A.4: Robustness of the results in Table 9 to controlling for experience

	Choice in round	Choose zero in round	Increased choice in round
Effect of thinking longer than normal (minutes)	0.908 (1.026)	-0.027** (0.013)	0.058*** (0.019)
Subject-round observations	4,774	4,774	4,774

Notes: We run the same regressions on the same sample as in Table 10, adding controls for the subject's experience of her current situation as described in the notes to Table A.4. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table A.5: Robustness of the results in Table 10 to controlling for experience

	Winner of round	Earnings in round	Log earnings in round
Effect of thinking longer than normal (minutes)	-0.090** (0.039)	-40.807* (21.069)	-0.201** (0.093)
Subject-round observations	4,774	4,774	4,774

Notes: We run the same regressions on the same sample as in Table 9, controlling for heterogeneous round trends by including a separate linear round trend for each of the 780 subjects in the sample (formally, we augment (6) by adding the term $\rho_i \times r$ and treat ρ_i for $i = 1, ..., 780$, as parameters to be estimated). Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table A.6: Robustness of the results in Table 9 to controlling for heterogeneous round trends

	Choice in round	Choose zero in round	Increased choice in round
Effect of thinking longer than normal (minutes)	1.197	-0.020	0.080***
	(1.355)	(0.015)	(0.025)
Subject-round observations	4,774	4,774	4,774

Notes: We run the same regressions on the same sample as in Table 10, controlling for heterogeneous round trends as described in the notes to Table A.6. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table A.7: Robustness of the results in Table 10 to controlling for heterogeneous round trends

	Winner of round	Earnings in round	Log earnings in round
Effect of thinking longer than normal (minutes)	-0.060**	-28.407*	-0.136**
	(0.029)	(15.600)	(0.069)
Subject-round observations	4,510	4,510	4,510

Notes: We start with the sample of 6,694 subject-round observations described in Table A.1. From this sample, we then use the subject-round observations for which the subject is observed in the same situation in at least one other round (as explained in the notes to Table 9, this procedure ensures that subject-round observations are not fully absorbed by the fixed effects). This gives us 4,510 subject-round observations and 1,682 subject-situation fixed effects. We then run the same regressions as in Table 9. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table A.8: Robustness of the results in Table 9 to excluding the equilibrium situation

	Choice in round	Choose zero in round	Increased choice in round
Effect of thinking longer than normal (minutes)	1.026	-0.027**	0.058***
	(1.066)	(0.013)	(0.019)
Subject-round observations	4,510	4,510	4,510

Notes: The sample and subject-situation fixed effects are as described in the notes to Table A.8. We run the same regressions as in Table 10. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table A.9: Robustness of the results in Table 10 to excluding the equilibrium situation

References (for main text and all appendices)

Achtziger, A. and **Alós-Ferrer, C.** (2013). Fast or rational? A response-times study of Bayesian updating. *Management Science*, 60(4): 923–938

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.** (2016a). Endogenous depth of reasoning. *Review of Economic Studies*, 83(4): 1297–1333

Alaoui, L. and **Penta, A.** (2016b). Endogenous depth of reasoning and response time, with an application to the attention allocation task. *Mimeo, University of Wisconsin*

Alaoui, L. and **Penta, A.** (forthcoming). Cost-benefit analysis in reasoning. *Journal of Political Economy*

Alekseev, A. (2019). Using response times to measure ability on a cognitive task. *Journal of the Economic Science Association*, 5(1): 65–75

Alós-Ferrer, C. and **Buckenmaier, J.** (2021). Cognitive sophistication and deliberation times. *Experimental Economics*, 24(2): 558–592

Arad, A. and **Rubinstein, A.** (2012). Multi-dimensional iterative reasoning in action: The case of the Colonel Blotto game. *Journal of Economic Behavior & Organization*, 84(2): 571–585

Avoyan, A. and **Schotter, A.** (2020). Attention in games: An experimental study. *European Economic Review*, 124: 103410

Brañas-Garza, P., **García-Muñoz, T.**, and **González, R.H.** (2012). Cognitive effort in the beauty contest game. *Journal of Economic Behavior & Organization*, 83(2): 254–260

Brañas-Garza, P., **Meloso, D.**, and **Miller, L.M.** (2017). Strategic risk and response time across games. *International Journal of Game Theory*, 46(2): 511–523

Burnham, T.C., **Cesarini, D.**, **Johannesson, M.**, **Lichtenstein, P.**, and **Wallace, B.** (2009). Higher cognitive ability is associated with lower entries in a p-beauty contest. *Journal of Economic Behavior & Organization*, 72(1): 171–175

Busemeyer, J.R. and **Townsend, J.T.** (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100(3): 432–459

Camerer, C. and **Ho, H.T.** (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, 67(4): 827–874

Camerer, C.F., **Ho, T.H.**, and **Chong, J.K.** (2002). Sophisticated experience-weighted attraction learning and strategic teaching in repeated games. *Journal of Economic Theory*, 104(1): 137–188

Camerer, C.F., **Ho, T.H.**, and **Chong, J.K.** (2004). Behavioural game theory: Thinking, learning and teaching. In S. Huck, editor, *Advances in Understanding Strategic Behaviour: Game Theory, Experiments and Bounded Rationality*, 120–180. Palgrave Macmillan

Caplin, A. and **Martin, D.** (2016). The dual-process drift diffusion model: Evidence from response times. *Economic Inquiry*, 54(2): 1274–1282

Cappelen, A.W., **Nielsen, U.H.**, **Tungodden, B.**, **Tyran, J.R.**, and **Wengström, E.** (2016). Fairness is intuitive. *Experimental Economics*, 19(4): 727–740

Carpenter, P.A., **Just, M.A.**, and **Shell, P.** (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. *Psychological Review*, 97(3): 404–413

Chabris, C.F., **Laibson, D.**, **Morris, C.L.**, **Schuldt, J.P.**, and **Taubinsky, D.** (2009). The allocation of time in decision-making. *Journal of the European Economic Association*, 7(2-3): 628–637

Chen, D.L., **Schonger, M.**, and **Wickens, C.** (2016). oTree–An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9: 88–97

Chen, F. and **Fischbacher, U.** (2016). Response time and click position: Cheap indicators of preferences. *Journal of the Economic Science Association*, 2(2): 109–126

Chen, S.H., **Du, Y.R.**, and **Yang, L.X.** (2014). Cognitive capacity and cognitive hierarchy: A study based on beauty contest experiments. *Journal of Economic Interaction and Coordination*, 9(1): 69–105

Clithero, J.A. (2018). Improving out-of-sample predictions using response times and a model of the decision process. *Journal of Economic Behavior & Organization*, 148: 344–375

Clithero, J.A. and **Rangel, A.** (2013). Combining response times and choice data using a neuroeconomic model of the decision process improves out-of-sample predictions. *Mimeo, California Institute of Technology*

Coibion, O., **Gorodnichenko, Y.**, **Kumar, S.**, and **Ryngaert, J.** (2020). Do you know that I know that you know...? Higher-order beliefs in survey data. *Mimeo, UT Austin*

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1): 87–114

Dal Bó, P. and **Fréchette, G.R.** (2011). The evolution of cooperation in infinitely repeated games: Experimental evidence. *American Economic Review*, 101(1): 411–29

Dal Bó, P. and **Fréchette, G.R.** (2018). On the determinants of cooperation in infinitely repeated games: A survey. *Journal of Economic Literature*, 56(1): 60–114

Dal Bó, P. and **Fréchette, G.R.** (2019). Strategy choice in the infinitely repeated prisoner's dilemma. *American Economic Review*, 109(11): 3929–52

Di Guida, S. and **Devetag, G.** (2013). Feature-based choice and similarity perception in normal-form games: An experimental study. *Games*, 4(4): 776–794

Duckworth, A.L., **Peterson, C.**, **Matthews, M.D.**, and **Kelly, D.R.** (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, 92(6): 1087–1101

Duffy, J. and **Nagel, R.** (1997). On the robustness of behaviour in experimental 'Beauty Contest' games. *Economic Journal*, 107(445): 1684–1700

Echenique, F. and **Saito, K.** (2017). Response time and utility. *Journal of Economic Behavior & Organization*, 139: 49–59

Elsner, B. and **Isphording, I.E.** (2017). A big fish in a small pond: Ability rank and human capital investment. *Journal of Labor Economics*, 35(3): 787–828

Embrey, M., **Fréchette, G.R.**, and **Yuksel, S.** (2018). Cooperation in the finitely repeated prisoner's dilemma. *Quarterly Journal of Economics*, 133(1): 509–551

Evans, A.M., **Dillon, K.D.**, and **Rand, D.G.** (2015). Fast but not intuitive, slow but not reflective: Decision conflict drives reaction times in social dilemmas. *Journal of Experimental Psychology: General*, 144(5): 951–966

Fe, E., **Gill, D.**, and **Prowse, V.L.** (forthcoming). Cognitive skills, strategic sophistication, and life outcomes. *Journal of Political Economy*

Fehr, D. and **Huck, S.** (2016). Who knows it is a game? On strategic awareness and cognitive ability. *Experimental Economics*, 19(4): 713–726

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2): 171–178

Frydman, C. and **Krajbich, I.** (forthcoming). Using response times to infer others' private information: An application to information cascades. *Management Science*

Fudenberg, D., **Rand, D.G.**, and **Dreber, A.** (2012). Slow to anger and fast to forgive: Cooperation in an uncertain world. *American Economic Review*, 102(2): 720–49

Fudenberg, D., **Strack, P.**, and **Strzalecki, T.** (2018). Speed, accuracy, and the optimal timing of choices. *American Economic Review*, 108(12): 3651–84

Gill, D. and **Prowse, V.** (2016). Cognitive ability, character skills and learning to play equilibrium: A level-k analysis. *Journal of Political Economy*, 124(6): 1619–1676

Gill, D. and **Rosokha, Y.** (2020). Beliefs, learning, and personality in the indefinitely repeated prisoner's dilemma. *CAGE Working Paper 489*

Glazer, J. and **Rubinstein, A.** (2012). A model of persuasion with boundedly rational agents. *Journal of Political Economy*, 120(6): 1057–1082

Gneezy, U., **Rustichini, A.**, and **Vostroknutov, A.** (2010). Experience and insight in the Race game. *Journal of Economic Behavior & Organization*, 75(2): 144–155

Goeree, J.K. and **Holt, C.A.** (2001). Ten little treasures of game theory and ten intuitive contradictions. *American Economic Review*, 91(5): 1402–1422

Gray, J. and **Thompson, P.** (2004). Neurobiology of intelligence: Science and ethics. *Nature Reviews Neuroscience*, 5(6): 471–482

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1): 114–125

Grubb, M.D. and **Osborne, M.** (2015). Cellular service demand: Biased beliefs, learning, and bill shock. *American Economic Review*, 105(1): 234–71

Handel, B. and **Schwartzstein, J.** (2018). Frictions or mental gaps: What's behind the information we (don't) use and when do we care? *Journal of Economic Perspectives*, 32(1): 155–78

Hastie, T., **Tibshirani, R.**, and **Wainwright, M.** (2019). Statistical learning with sparsity: The lasso and generalizations. In F. Bunea, V. Isham, N. Keiding, T. Louis, R. Smith, and H. Tong, editors, *Monographs on Statistics and Applied Probability*, 143, 1–335. Chapman and Hall

Ho, T.H., **Camerer, C.**, and **Weigelt, K.** (1996). Iterated dominance and iterated best-response in experimental "p-beauty contests". *Social Science Working Paper 974, CalTech*

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

Ho, T.H., **Camerer, C.F.**, and **Chong, J.K.** (2007). Self-tuning experience weighted attraction learning in games. *Journal of Economic Theory*, 133(1): 177–198

Ho, T.H., **Park, S.E.**, and **Su, X.** (2021). A Bayesian level-k model in n-person games. *Management Science*, 67(3): 1622–1638

Hutcherson, C.A., **Bushong, B.**, and **Rangel, A.** (2015). A neurocomputational model of altruistic choice and its implications. *Neuron*, 87(2): 451–462

John, O.P., **Naumann, L.P.**, and **Soto, C.J.** (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O.P. John, R.W. Robins, and L.A. Pervin, editors, *Handbook of personality: Theory and research*, 114–158. New York, NY: Guilford Press

John, O.P., **Donahue, E.M.**, and **Kentle, R.L.** (1991). The Big Five Inventory: Versions 4a and 54. *Institute of Personality and Social Research, University of California, Berkeley*

Jolliffe, I.T. (1995). Rotation of principal components: Choice of normalization constraints. *Journal of Applied Statistics*, 22(1): 29–35

Kahneman, D. (1973). *Attention and effort*. Prentice-Hall

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan

Kivetz, R., **Netzer, O.**, and **Srinivasan, V.** (2004). Alternative models for capturing the compromise effect. *Journal of Marketing Research*, 41(3): 237–257

Kocher, M.G. and **Sutter, M.** (2006). Time is money? Time pressure, incentives, and the quality of decision-making. *Journal of Economic Behavior & Organization*, 61(3): 375–392

Konovalov, A. and **Krajbich, I.** (2019). Revealed strength of preference: Inference from response times. *Judgment and Decision Making*, 14(4): 381–394

Krajbich, I., **Bartling, B.**, **Hare, T.**, and **Fehr, E.** (2015). Rethinking fast and slow based on a critique of reaction time reverse inference. *Nature Communications*, 6(7455): 1–9

Krajbich, I., **Oud, B.**, and **Fehr, E.** (2014). Benefits of neuroeconomic modeling: New policy interventions and predictors of preference. *American Economic Review: Papers and Proceedings*, 104(5): 501–06

Kuo, W.J., **Sjöström, T.**, **Chen, Y.P.**, **Wang, Y.H.**, and **Huang, C.Y.** (2009). Intuition and deliberation: Two systems for strategizing in the brain. *Science*, 324(5926): 519–522

Lindner, F. and **Sutter, M.** (2013). Level-k reasoning and time pressure in the 11–20 money request game. *Economics Letters*, 120(3): 542–545

Lohse, J., **Goeschl, T.**, and **Diederich, J.H.** (2017). Giving is a question of time: Response times and contributions to an environmental public good. *Environmental and Resource Economics*, 67(3): 455–477

López, R. (2001). On p-beauty contest integer games. *UPF Economics and Business Working Paper No. 608*

Lotito, G., **Migheli, M.**, and **Ortona, G.** (2013). Is cooperation instinctive? Evidence from the response times in a public goods game. *Journal of Bioeconomics*, 15(2): 123–133

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2): 81–97

Moritz, B., **Siemsen, E.**, and **Kremer, M.** (2014). Judgmental forecasting: Cognitive reflection and decision speed. *Production and Operations Management*, 23(7): 1146–1160

Murphy, R. and **Weinhardt, F.** (2020). Top of the class: The importance of ordinal rank. *Review of Economic Studies*, 87(6): 2777–2826

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review*, 85(5): 1313–1326

Nagel, R., **Bühren, C.**, and **Frank, B.** (forthcoming). Inspired and inspiring: Hervé Moulin and the discovery of the beauty contest game. *Mathematical Social Sciences*

Nielsen, U.H., **Tyran, J.R.**, and **Wengström, E.** (2014). Second thoughts on free riding. *Economics Letters*, 122(2): 136–139

Nishi, A., **Christakis, N.A.**, **Evans, A.M.**, **O'Malley, A.J.**, and **Rand, D.G.** (2016). Social environment shapes the speed of cooperation. *Scientific Reports*, 6(29622): 1–10

Nishi, A., **Christakis, N.A.**, and **Rand, D.G.** (2017). Cooperation, decision time, and culture: Online experiments with American and Indian participants. *PloS ONE*, 12(2): 1–9

Piovesan, M. and **Wengström, E.** (2009). Fast or fair? A study of response times. *Economics Letters*, 105(2): 193–196

Polonio, L., **Di Guida, S.**, and **Coricelli, G.** (2015). Strategic sophistication and attention in games: An eye-tracking study. *Games and Economic Behavior*, 94: 80–96

Proto, E., **Rustichini, A.**, and **Sofianos, A.** (2019). Intelligence, personality, and gains from cooperation in repeated interactions. *Journal of Political Economy*, 127(3): 1351–1390

Proto, E., **Rustichini, A.**, and **Sofianos, A.** (forthcoming). Intelligence, errors and cooperation in repeated interactions. *Review of Economic Studies*

Rand, D.G., **Fudenberg, D.**, and **Dreber, A.** (2015). It's the thought that counts: The role of intentions in noisy repeated games. *Journal of Economic Behavior & Organization*, 116: 481–499

Rand, D.G., **Greene, J.D.**, and **Nowak, M.A.** (2012). Spontaneous giving and calculated greed. *Nature*, 489(7416): 427–430

Recalde, M.P., **Riedl, A.**, and **Vesterlund, L.** (2018). Error-prone inference from response time: The case of intuitive generosity in public-good games. *Journal of Public Economics*, 160: 132–147

Rigdon, M.L., **McCabe, K.A.**, and **Smith, V.L.** (2007). Sustaining cooperation in trust games. *Economic Journal*, 117(522): 991–1007

Roth, A. and **Xing, X.** (1994). Jumping the gun: Imperfections and institutions related to the timing of market transactions. *American Economic Review*, 84(4): 992–1044

Rubinstein, A. (2007). Instinctive and cognitive reasoning: A study of response times. *Economic Journal*, 117(523): 1243–1259

Rubinstein, A. (2013). Response time and decision making: An experimental study. *Judgment and Decision Making*, 8(5): 540–551

Rubinstein, A. (2016). A typology of players: Between instinctive and contemplative. *Quarterly Journal of Economics*, 131(2): 859–890

Schotter, A. and **Trevino, I.** (2021). Is response time predictive of choice? An experimental study of threshold strategies. *Experimental Economics*, 24(1): 87–117

Sims, C.A. (2010). Rational inattention and monetary economics. In B.M. Friedman and M. Woodford, editors, *Handbook of Monetary Economics*, volume 3, 155–181. Elsevier

Sivakumar, K. (2016). A unified conceptualization of the attraction effect. *AMS Review*, 6(1): 39–58

Spiliopoulos, L. (2018). The determinants of response time in a repeated constant-sum game: A robust Bayesian hierarchical dual-process model. *Cognition*, 172: 107–123

Spiliopoulos, L. and **Ortmann, A.** (2018). The BCD of response time analysis in experimental economics. *Experimental Economics*, 21(2): 383–433

Spiliopoulos, L., **Ortmann, A.**, and **Zhang, L.** (2018). Complexity, attention, and choice in games under time constraints: A process analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(10): 1609–1640

Stahl, D.O. (1996). Boundedly rational rule learning in a guessing game. *Games and Economic Behavior*, 16(2): 303–330

Stahl, D.O. and **Wilson, P.W.** (1994). Experimental evidence on players' models of other players. *Journal of Economic Behavior & Organization*, 25(3): 309–327

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, 25(3): 251–260

Strathman, A., **Gleicher, F.**, **Boninger, D.S.**, and **Scott Edwards, C.** (1994). The consideration of future consequences: Weighing immediate and distant outcomes of behavior. *Journal of Personality and Social Psychology*, 66(4): 742–752

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1): 267–288

Turocy, T.L. and **Cason, T.N.** (2015). Bidding in first-price and second-price interdependent-values auctions: A laboratory experiment. *Mimeo, University of East Anglia*

Wilcox, N.T. (1993). Lottery choice: Incentives, complexity and decision time. *Economic Journal*, 103(421): 1397–1417

Wilson, B.J. (2014). The meaning of deceive in experimental economic science. In G.F. DeMartino and D. McCloskey, editors, *Oxford Handbook of Professional Economic Ethics*, 317–326. Oxford University Press

Woodford, M. (2014). Stochastic choice: An optimizing neuroeconomic model. *American Economic Review: Papers and Proceedings*, 104(5): 495–500

Online Appendix

(Intended for Online Publication)

Online Appendix I Experimental Instructions

Online Appendix I.1 Instructions for the experiment described in Section 2

Please look at your screen now. I am reading from the instructions displayed on your screen. Please now turn off cell phones and any other electronic devices. These must remain turned off for the duration of this session. Please do not use or place on your desk any personal items, including pens, paper, phones etc. Please do not look into anyone else's booth at any time. Thank you for participating in this experimental session on economic decision-making. You were randomly selected from the Economic Science Laboratory's pool of subjects to be invited to participate in this session. There will be a number of pauses for you to ask questions. During such a pause, please raise your hand if you want to ask a question. Apart from asking questions in this way, you must not communicate with anybody in this room or make any noise.

You will be paid a show-up fee of $5 together with any money you accumulate during this session. The amount of money you accumulate will depend partly on your actions and partly on the actions of other participants. You will be paid privately in cash at the end of the session.

{*Further instructions in the 15 sessions that included questionnaires to measure personality:* Please raise your hand if you have any questions. Before we start the experimental session, I would like you to complete a pre-experimental survey. The survey is made up of 68 questions. There are 17 pages with 4 questions on each page. For each question, please enter your answer in the column to the right of the question. There are no right or wrong answers to the questions. You will have 8 minutes to complete the pre-experimental survey. During the 8 minutes, you can move back and forth between the 17 pages and you can change your previous answers. The top right-hand corner of the screen will display the time remaining (in seconds). Before we start the pre-experimental survey, please raise your hand if you have any questions. During the pre-experimental survey, please raise your hand if you have a problem with your computer. [Subjects complete questionnaire] Thank you for completing the pre-experimental survey. I will now describe the experimental session.}

The session is made up of 2 parts. In the first part you will complete a test. Right at the end of the session you will find out your own test score, but you will not be paid for completing the test. I will describe the second part of the session after you have completed the test. Please raise your hand if you have any questions.

I will now describe the test which makes up the first part of the session. The test is made up of 60 questions, divided into parts A, B, C, D and E. Each of these parts is made up of 12 questions. For every question, there is a pattern with a piece missing and a number of pieces below the pattern. You have to choose which of the pieces below is the right one to complete the pattern. For parts A and B of the test, you will see 6 pieces that might complete the pattern. For parts C, D and E you will see 8 pieces that might complete the pattern. In every case, one and only one of these pieces is the right one to complete the pattern.[28] For each question, please enter your answer in the column to the right of the pattern. You will score 1 point for every right answer. You will not be penalized for wrong answers. You will have 3 minutes to complete each of parts A and B, and you will have 8 minutes to complete each of parts C, D, and E. During each part, you can move back and forth between the 12 questions in that part and you can change your previous answers. The top right-hand corner of the screen will display the time remaining (in seconds). Before we start the test, please raise your hand if you have any questions. During the test, please raise your hand if you have a problem with your

[28] The wording of this description follows the standard Raven test convention.

computer. [Subjects complete test]

Your screen is now displaying whether your test score was in the top half of the test scores of all participants in the room or was in the bottom half of the test scores of all participants. [30 second pause] [Example (not read aloud): Your test score was in the top half of the test scores of all participants in the room.] At the end of the session you will find out your own test score.

I will now describe the second and final part of the session. This second part is made up of 10 rounds. You will be anonymously matched into groups of 3 participants. You will stay in the same group for all 10 rounds. In each round, you and your other 2 group members will separately choose a whole number between 0 and 100 (0, 100 or any whole number in between is allowed). The group member whose chosen number is closest to 70% of the average of all 3 chosen numbers will be paid $6 for that round and the other 2 group members will be paid nothing. If more than one group member chooses a number which is closest to 70% of the average of all 3 chosen numbers, the $6 will be split equally among the group members who chose the closest number or numbers. Your total payment will be the sum of your payments in each round together with your show-up fee of $5. In each round you will have 90 seconds to choose your number. If you choose your number early you will still have to wait until the end of the 90 seconds. The top right-hand corner of the screen will display the time remaining (in seconds). The screen will also include a reminder of the rules.

At the end of each round you will discover: (i) the numbers chosen by all your group members; (ii) the average of all 3 chosen numbers; (iii) what 70% of the average of all 3 chosen numbers was; and (iv) how much each group member will be paid for the round. Please raise your hand if you have any questions.

You will stay in the same group of 3 for all 10 rounds. Each group member has been randomly allocated a label, X, Y or Z. Your screen is now displaying your label and whether the test scores of the members of your group were in the top half or the bottom half of the test scores of all participants in the room. [60 second pause] [Example (not read aloud): You are group member Y. Your test score was in the top half of the test scores of all participants in the room. You have been matched with 2 participants (group member X and group member Z). Group member X was randomly selected from those whose test scores were also in the top half. Group member Z was randomly selected from those whose test scores were in the bottom half.] Please raise your hand if you have any questions. There will be no further opportunities for questions.

[10 rounds of beauty contest with feedback as described in Section II.C of Gill and Prowse (2016)]

[Screen asks subjects to report their gender]

[Screen reports the subject's score in the Raven test]

The session has now finished. Your total cash payment, including the show-up fee, is displayed on your screen. Please remain in your seat until you have been paid. Thank-you for participating.

Online Appendix I.2 Instructions for the supplemental experiment described in Online Appendix II.6.2

[Screen 1] Introduction. Please now turn off cell phones and any other electronic devices. These must remain turned off for the duration of this session. Please do not use or place on your desk any personal items, including pens, paper, phones etc. Please do not look into anyone else's booth at any time. Thank you for participating in this experimental session on economic decision-making. You were randomly selected from the Vernon Smith Experimental Economics Laboratory's pool of subjects to be invited to participate in this session. There will be a number of points in these instructions where

you will be asked to raise your hand if you have any questions. Apart from asking questions in this way, you must not communicate with anybody in this room or make any noise. You will be paid a show-up fee of $5 together with any money you accumulate during this session. The amount of money you accumulate will depend partly on your actions and partly on the actions of other participants. You will be paid privately in cash at the end of the session. Please raise your hand if you have any questions. [Button: Click to continue.]

[Screen 2] Introduction. This session is made up of 2 parts. In the first part of the session, you will participate in an economic interaction that includes opportunities to earn money. In the second part of the session, you will be asked to complete two tests and a short questionnaire. We will pay you $3 for each test (irrespective of your scores on the tests). You will not be paid for completing the questionnaire. Please raise your hand if you have any questions. [Button: Click to continue.]

[Screen 3] Instructions on Part 1 of the session. We now describe the economic interaction that makes up the first part of the session. The economic interaction will last for up to 10 rounds. You will be anonymously matched into groups of 3 participants. You will stay in the same group for all rounds. In each round, you and your other 2 group members will separately choose a whole number between 0 and 100 (0, 100 or any whole number in between is allowed). The group member whose chosen number is closest to 70% of the average of all 3 chosen numbers will be paid $6 for that round and the other 2 group members will be paid nothing. If more than one group member chooses a number which is closest to 70% of the average of all 3 chosen numbers, the $6 will be split equally among the group members who chose the closest number or numbers. Your total payment from the economic interaction will be the sum of your payments in each round. In each round you will have 90 seconds to choose your number. If you choose your number early you will still have to wait until the end of the 90 seconds. The screen will display the time remaining (in seconds). The screen will also include a reminder of the rules. At the end of each round you will discover: (i) the numbers chosen by all your group members; (ii) the average of all 3 chosen numbers; (iii) what 70% of the average of all 3 chosen numbers was; (iv) how much each group member will be paid for the round. Please raise your hand if you have any questions. [Button: Click to continue.]

[Screen 4] Instructions on Part 1 of the session. Recall, the economic interaction will last for up to 10 rounds. Specifically: (1) The economic interaction will last for at least 5 rounds. (2) Starting from the 5th round, at the end of each round there is a [depending on treatment: 50%, 75%, or 90%] chance that the economic interaction continues to the next round and a [depending on treatment: 50%, 25%, or 10%] chance that the economic interaction ends. (3) If the economic interaction reaches the 10th round, then the interaction ends for sure at the end of that 10th round. You will stay in the same group of 3 for all rounds of the economic interaction. Each group member has been randomly allocated a label, X, Y or Z. Your label is shown below. [Example: You are group member X.] Please raise your hand if you have any questions. There will be no further opportunities for questions on this part of the session. [Button: Click to continue.]

[Between 5 and 10 rounds of beauty contest with feedback as described in Online Appendix II.6.2]

[Subjects complete two cognitive ability tests and a demographic questionnaire (see footnotes 34 and 38)]

[Final screen] The session has now finished. Your total cash payment, including the show-up fee, is [total payment in dollars]. Please remain seated in your booth. The lab assistant will come to your booth to give you your cash payment. Thank-you for participating.

Online Appendix II Forward-looking behavior

Online Appendix II.1 Introduction

In some repeated games, some subjects behave in a forward-looking manner by choosing stage-game actions to influence outcomes beyond the current round. For example, in the indefinitely repeated prisoner's dilemma, subjects' myopic best-response is to defect in every round, but experimental evidence shows that subjects cooperate more when the continuation probability is higher (Dal Bó and Fréchette, 2011, 2018), which allows the play of efficient cooperative subgame-perfect Nash equilibria to sometimes emerge in the laboratory.

In this appendix, we provide different types of evidence that all support the conclusion that such forward-looking behavior is unlikely to be an important driver of behavior in our finitely repeated beauty contest setting.

Online Appendix II.2 Thinking for longer than normal has no effect on success in the following rounds

In Table 9 we showed that when a subject thinks for longer than normal, she performs worse in the same round. We now provide evidence that these subjects are not trading off worse performance in the current round for better performance in later rounds. Specifically, Table OA.1 shows that thinking for longer than normal in a particular round has no effect on success across the next two rounds (the effect sizes are positive but small and far from statistical significance). Tables OA.2 and OA.3 show that thinking for longer has a negative (but statistically insignificant) effect on success in the very next round and a positive (but again statistically insignificant) effect two rounds in the future. When we consider the effect of thinking for longer than normal in round r on choices (instead of on success) in round $r+1$ or $r+2$, we find effects that are positive but statistically insignificant at the ten-percent level.

	Winner	Earnings	Log earnings
Effect of thinking longer than normal (minutes)	0.008 (0.022)	2.535 (11.787)	0.015 (0.052)
Subject-round observations	3,394	3,394	3,394

Notes: This table reports parameter estimates from the regression described in (6) but with the dependent variable, $Success_{i,r}$, replaced by the mean of subject i's success across rounds $r+1$ and $r+2$ (see the notes to Table 9 for details about our measures of success). We start with the sample of 7,012 subject-round observations described in Table 4 and exclude observations from rounds 9 and 10 (since we do not observe success in both rounds $r+1$ and $r+2$ for these observations). From this sample, we then use the subject-round observations for which the subject is observed in the same situation in at least one other round (as explained in the notes to Table 9, this procedure ensures that subject-round observations are not fully absorbed by the fixed effects). This gives us 3,394 subject-round observations and 1,340 subject-situation fixed effects. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table OA.1: Effect of thinking for longer than normal in round r on average success in rounds $r+1$ and $r+2$

	Winner	Earnings	Log earnings
Effect of thinking longer than normal (minutes)	-0.031 (0.029)	-14.225 (14.527)	-0.068 (0.066)
Subject-round observations	4,088	4,088	4,088

Notes: This table reports parameter estimates from the regression described in (6) but with the dependent variable, $Success_{i,r}$, replaced by subject i's success in round $r+1$ (see the notes to Table 9 for details about our measures of success). We start with the sample of 7,012 subject-round observations described in Table 4 and exclude observations from round 10 (since we do not observe success in round $r+1$ for these observations). From this sample, we then use the subject-round observations for which the subject is observed in the same situation in at least one other round (as explained in the notes to Table 9, this procedure ensures that subject-round observations are not fully absorbed by the fixed effects). This gives us 4,088 subject-round observations and 1,565 subject-situation fixed effects. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table OA.2: Effect of thinking for longer than normal in round r on success in round $r+1$

	Winner	Earnings	Log earnings
Effect of thinking longer than normal (minutes)	0.034 (0.031)	14.879 (16.388)	0.075 (0.073)
Subject-round observations	3,394	3,394	3,394

Notes: This table reports parameter estimates from the regression described in (6) but with the dependent variable, $Success_{i,r}$, replaced by subject i's success in round $r+2$ (see the notes to Table 9 for details about our measures of success). We use the same sample as in Table OA.1. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table OA.3: Effect of thinking for longer than normal in round r on success in round $r+2$

Online Appendix II.3 Theoretical considerations

In the data from Gill and Prowse (2016) that we study in this paper, subjects played ten rounds of the p-beauty contest game (with $p = 0.7$) in fixed groups of three without rematching. The stage game is a constant-sum game with a unique Nash equilibrium in which all three players choose zero.[29] By backward induction, the unique subgame-perfect Nash equilibrium of the repeated game is for all players to choose zero in every round.

In this setting, forward-looking behavior is unlikely. Below, we summarize the theoretical considerations that underpin this reasoning, but we note that Ho et al. (2021) study a similar dataset (also with ten rounds, $p = 0.7$, and fixed groups of three) and argue that: "Reputation building is unlikely because the p-beauty contest game is a constant-sum game where players' interests are strictly opposed."

Since the stage game is constant-sum, players have no incentive to behave in a forward-looking manner to try to coordinate play to improve efficiency (e.g., by coordinating play onto a cooperative

[29] Because the game is discrete with integer choices, the equilibrium is not unique for all values of p (López, 2001), but it is unique for $p = 0.7$ (see footnote 12 in Gill and Prowse, 2016).

subgame-perfect Nash equilibrium), or to try to manipulate others by pretending to be a cooperative type who cares about efficiency. Thus, the type of forward-looking behavior studied in the experimental literature on the finitely or indefinitely repeated prisoner's dilemma (e.g., Dal Bó and Fréchette, 2011, Embrey et al., 2018) cannot occur here. Furthermore, since the stage-game equilibrium is unique, players have no incentive to behave in a forward-looking manner to try to coordinate play in later rounds onto an egoistically preferred stage-game equilibrium. Thus, the type of forward-looking behavior that leads to 'aggressive' or 'stubborn' early-round choices in the repeated hawk-dove game or the repeated battle-of-the-sexes game cannot occur here.

Having said this, in the repeated beauty contest a player could try to manipulate her opponents by choosing a very high number in a particular round in an attempt to induce her opponents to believe that she will continue to choose a high number in the next round. If the opponents respond by moving their own choices up in the next round, the manipulating player might then be able to win the next round.

However, from a theoretical perspective this manipulative strategy is unattractive. First, by following this strategy, the player effectively guarantees herself a payoff of zero in one round in exchange for only the possibility of successful manipulation in the next round. Indeed, the manipulation is predicated on both opponents responding by raising their choices: Camerer et al. (2002) call forward-looking behavior in repeated games "strategic teaching", and note that across different repeated games such behavior is difficult in groups of three, because success depends on the behavior of the "least teachable" player. Second, this type of manipulation is unlikely to work more than once. Third, a sophisticated player who accurately predicts her opponents' behavior does not generally need to engage in such risky and costly forward-looking behavior in the repeated beauty contest; instead she can simply best-respond to her opponents' predicted choices in each round, given her understanding of how her choice and that of her opponents determine the target (seventy percent of the average of the three choices).[30]

Online Appendix II.4 Structural models of beauty contest behavior

The workhorse structural models that have been used to understand and predict behavior in the repeated beauty contest do not include forward-looking behavior, even though they allow sophisticated players: (i) to anticipate how others learn across repetitions of the game; and (ii) to then use their sophisticated forecasts to best-respond to predicted behavior in the current round. Camerer et al. (2004) succinctly summarize the nature of this sophisticated behavior: "Sophisticated players believe that others are learning and anticipate how those others will change in deciding what to do. In learning to shoot at a moving target, for example, soldiers and fighter pilots learn to shoot ahead, towards where the target will be, rather than shoot at the target where it is when they aim. They become sophisticated."

One class of models used to analyze repeated beauty contest data builds on the static level-k model by including learning across repetitions of the game (Nagel, 1995; Stahl, 1996; Duffy and Nagel, 1997; Ho et al., 1998; Gill and Prowse, 2016). In these models, level-0 agents learn in a strategically unsophisticated adaptive manner, while agents with level-$k > 0$ are highly sophisticated in the sense that they understand how lower-level agents learn across rounds and use their forecasts to best-respond in the current round. Ho et al. (2021)'s cutting-edge analysis of repeated beauty contest data goes

[30] When choices have converged to the equilibrium of zero, best-responding leads only to a tie. However most groups never converge, and those that do tend to do so toward the end of the ten rounds (see Gill and Prowse, 2016).

one step further by allowing sophisticated players to learn in a Bayesian way about others' levels, but the players in Ho et al. (2021)'s model still do not look forward beyond the current round.

Another class of models used to analyze repeated beauty contest data builds on Camerer and Ho (1999)'s experience-weighted attraction (EWA) framework that nests a range of adaptive learning behaviors including reinforcement learning and fictitious play (Camerer and Ho, 1999; Ho et al., 2007; Gill and Prowse, 2016). Camerer et al. (2002) extend the EWA model to include sophisticated players who anticipate the learning of the adaptive EWA players, and then best-respond to the predicted behavior of these adaptive players. Camerer et al. (2002) use repeated beauty contest data to estimate this extension of EWA, but once again the sophisticated players in this model do not look forward beyond the current round.[31]

Online Appendix II.5 Analysis of 'spoilers'

In Online Appendix II.3, we noted that in the repeated beauty contest a player could try to manipulate her opponents by choosing a very high number in a particular round in an attempt to induce her opponents to believe that she will continue to choose a high number in the next round. In the final paragraph of Online Appendix II.3 we explained why, from a theoretical perspective, this manipulative strategy is unattractive.

Ho et al. (1996, 1998) provide evidence that high choices in the repeated beauty contest are unlikely to be driven by forward-looking manipulation. Empirically, subjects sometimes choose 100, the highest available number; these choices are called 'spoilers' in the literature. In a similar setting to ours (also with ten rounds, with fixed groups of three or seven, and with $p = 0.7$ or $p = 0.9$), Ho et al. (1996, 1998) find that 1.9% of choices are spoilers. Importantly, Ho et al. (1996, 1998) provide evidence that these spoilers are not forward-looking attempts to manipulate opponents. In particular, Ho et al. (1996, 1998) report that choices of 100: (i) are evenly distributed across rounds; (ii) tend to follow low-payoff rounds; and (iii) do not raise average payoffs (comparing post-spoiling payoffs to pre-spoiling payoffs). From this analysis, Ho et al. (1996, 1998) conclude that the choices of 100 are likely driven by frustration or confusion, rather than forward-looking manipulation or altruism.

Our data exhibit similar patterns to those found by Ho et al. (1996, 1998). We find that 1.4% of choices (106 of 7,800) are spoilers (compared to 1.9%, or 52 of 2,770, in Ho et al., 1996, 1998). Similarly to Ho et al. (1996, 1998), in our dataset: (i) choices of 100 are evenly distributed across the first and second halves of the experiment, with 58.5% in rounds 1-5 and 41.5% in rounds 6-10; (ii) average per-round payoffs in the rounds before a choice of 100 are lower than the expected payoff from the beauty contest ($2) (although the difference is not statistically significant); and (iii) average per-round payoffs in the rounds after a choice of 100 are not statistically significantly different (at the ten-percent level) from average per-round payoffs in the rounds before a choice of 100.[32]

[31] Camerer et al. (2002) also study forward-looking behavior in repeated games, which they call "strategic teaching". However, despite using repeated beauty contest data earlier in the same paper to estimate the EWA model with sophistication, they do not use their repeated beauty contest data to study forward-looking behavior.

[32] The unit of analysis is subject-round observations where the subject chose 100. In (ii), we use all subject-round observations from rounds 2-10 where the subject chose 100 (to allow a comparison of earnings before the choice of 100 to the expected payoff). In (iii), we use all subject-round observations from rounds 2-9 where the subject chose 100 (to allow a comparison of earnings before and after the choice of 100).

Online Appendix II.6 Supplemental experiment

Online Appendix II.6.1 Introduction to the supplemental experiment

We collected supplemental experimental data from 141 subjects. Building on state-of-the-art methodology for studying forward-looking behavior in repeated games (Dal Bó and Fréchette, 2011, 2018), in the supplemental experiment we varied the incentive to engage in forward-looking behavior by varying across treatments the probability that the game continues to the next round. At the same time, we designed the supplemental experiment to mimic as closely as possible the ten-round repeated p-beauty contest experiment of Gill and Prowse (2016) that we study in this paper (indeed, the two settings share the same equilibrium properties).

If forward-looking behavior matters, then behavior should change with the probability that the game continues to the next round. We find no such evidence, which helps to support our conclusion in Online Appendix II.1 that forward-looking behavior is unlikely to be an important driver of behavior in our repeated beauty contest setting.

Online Appendix II.6.2 Design of the supplemental experiment

We collected supplemental experimental data from 141 student subjects at the Vernon Smith Experimental Economics Laboratory (VSEEL) at Purdue University during November 2021 (this in-person experiment was reviewed by the Purdue University IRB and followed all "Protect Purdue" Covid protocols).[33] The design of this supplemental experiment closely followed the design of the experiment in Gill and Prowse (2016) described in Section 2, with fixed groups of three subjects repeatedly playing the p-beauty contest with feedback at the end of each round, no rematching, $p = 0.7$, and a prize of \$6 per round (Online Appendix I.2 provides the experimental instructions).[34]

Instead of playing ten rounds of the p-beauty contest for sure as in Gill and Prowse (2016), in this supplemental experiment subjects played an uncertain number of rounds. In particular: (i) subjects played five rounds for sure; (ii) starting from the fifth round, the game continued to the next round with probability $q < 1$; and (iii) if the game reached the tenth round, the game ended for sure at the end of that tenth round. This information was carefully described to the subjects (with reminders about the continuation probability as the game progressed). We ran three between-subject treatments corresponding to three values of the continuation probability $q \in \{0.5, 0.75, 0.9\}$. In particular, we ran twelve sessions with an average of 11.75 subjects per session, with four sessions for each value of q. In total, we collected 999 subject-round observations.[35]

[33] IRB-2021-1558; November 9, 2021.

[34] We drew subjects from the VSEEL subject pool, which is administered using ORSEE (Greiner, 2015), and the experiment was programmed in oTree (Chen et al., 2016). We randomized seating positions. Experimental instructions were provided on the subject's computer screen. Here we did not measure cognitive ability before the repeated p-beauty contest, and therefore we did not match subjects by cognitive ability. After the repeated p-beauty contest, we measured basic demographics and cognitive ability (we plan to analyze the cognitive ability data in future work).

[35] Following, e.g., Fudenberg et al. (2012)'s indefinitely repeated prisoner's dilemma experiment, we drew the random game lengths in advance. We drew one game length for each of the twelve sessions. To balance the order of the treatments, we ran the sessions in blocks of three, with one iteration of each treatment within each block, and with rotation across blocks of the within-block treatment order.

Online Appendix II.6.3 Discussion of the design of the supplemental experiment

We designed the supplemental experiment to mimic as closely as possible the ten-round repeated p-beauty contest experiment of Gill and Prowse (2016) described in Section 2, while varying the incentive to engage in forward-looking behavior by varying the continuation probability across treatments. The experimental literature on the indefinitely repeated prisoner's dilemma shows that subjects cooperate more when the continuation probability is higher (Dal Bó and Fréchette, 2011, 2018), thus providing evidence for forward-looking behavior in that setting. Building on this state-of-the-art methodology for studying forward-looking behavior in repeated games, in the supplemental experiment we vary the continuation probability $q \in \{0.5, 0.75, 0.9\}$.[36] In the treatment with $q = 0.5$, starting from the fifth round, there is only a 50 percent probability that the game continues to the next round; by contrast, in the treatment with $q = 0.9$ this probability increases to 90 percent. If forward-looking behavior matters, then behavior should change with the probability that the game continues to the next round: in the next section we test this hypothesis.

Even though the repeated game that we study in the supplemental experiment lasts for an uncertain number of rounds, our design ensures that the game shares the same equilibrium properties with the ten-round repeated p-beauty contest game of Gill and Prowse (2016), and so the theoretical considerations discussed in Online Appendix II.3 continue to apply.[37] Furthermore: (i) conditional on reaching a particular round, the subjects share the same amount of experience across the two settings; and (ii) conditional on reaching the tenth round, the game ends for sure at the end of that final round in both settings. Note also that we let subjects in the supplemental experiment gain experience by having them play five rounds for sure.

Online Appendix II.6.4 Results from the supplemental experiment

As discussed in Online Appendix II.6.3, we look for evidence of forward-looking behavior by analyzing whether behavior in the supplemental experiment varies with the probability that the game continues to the next round. We find no such evidence, which helps to support our conclusion in Online Appendix II.1 that forward-looking behavior is unlikely to be an important driver of behavior in our repeated beauty contest setting.

Specifically, we run OLS regressions of five measures of behavior on the probability that the game continues to the next round, with round fixed effects and demographic controls.[38] In all five regressions, the coefficient on the probability that the game continues to the next round is not statistically significantly different from zero at the ten-percent level. We check robustness by regressing the measures of behavior on the probability that the game continues for at least two more rounds: we also find

[36] In the context of the indefinitely repeated prisoner's dilemma, Dal Bó and Fréchette (2011) use $q \in \{0.5, 0.75\}$, while Dal Bó and Fréchette (2019)'s main experiment further adds $q = 0.9$.

[37] As we note in Online Appendix II.3, in the ten-round repeated p-beauty contest game of Gill and Prowse (2016): (i) the stage-game is a constant-sum game with a unique Nash equilibrium in which all three players choose zero; and (ii) the unique subgame-perfect Nash equilibrium of the ten-round repeated game is for all players to choose zero in every round. Even though the repeated game that we study in the supplemental experiment lasts for an uncertain number of rounds, it shares these same properties, noting that backward induction starts from the tenth (final) round, conditional on the game proceeding that far.

[38] Recall that the game lasted five rounds for sure, and if the game reached the tenth round, then the game ended for sure at the end of that tenth round; thus, in rounds 1-4 the probability that the game continues to the next round equals 1, in rounds 5-9 this probability equals $q \in \{0.5, 0.75, 0.9\}$ (with between-subject variation in q), and in round 10 this probability equals 0. We measured the following basic demographics (for which we include indicators): gender; age group; whether native English speaker; and if not age group at which started learning English. p-values are calculated based on heteroskedasticity-consistent standard errors with clustering at the group level (two-sided tests).

that the coefficient on this probability is not statistically significant in any of the five regressions.[39]

In particular, we consider the following five measures of behavior:

1. Choice in round r.
2. Response time in round r.
3. Indicator for choice of 100 in round r.
4. Indicator for whether subject's choice increased by 25 percent or more from round $r-1$ to round r and then fell back in round $r+1$ to the level from round $r-1$ or lower.
5. Indicator for whether subject's choice increased by 50 percent or more from round $r-1$ to round r and then fell back in round $r+1$ to the level from round $r-1$ or lower.

We include the third to fifth measures because in the repeated beauty contest a player could try to manipulate her opponents by choosing a very high number in a particular round in an attempt to induce her opponents to believe that she will continue to choose a high number in the next round (although we explain in Online Appendix II.3 why this manipulative strategy is unattractive from a theoretical perspective). As we describe in Online Appendix II.5, Ho et al. (1996, 1998) provide evidence that high choices in the repeated beauty contest are unlikely to be driven by forward-looking manipulation; Ho et al. (1996, 1998)'s evidence is based on analysis of choices of 100 (the highest available number, which they call 'spoilers'), and so we include choices of 100 as one of our measures of behavior. We further include two measures based on subjects substantially increasing their choice from one round to the next and then decreasing their choice in the following round.

[39] In rounds 1-3 the probability that the game continues for at least two more rounds equals 1; in round 4 this probability equals $q \in \{0.5, 0.75, 0.9\}$; in rounds 5-8 this probability equals q^2; and in rounds 9-10 this probability equals 0.

Online Appendix III More on categorizing situations

Online Appendix III.1 Introduction

In Section 4 we categorized situations according to characteristics of play in the previous round. In this appendix, we provide different types of evidence that support our categorization methodology.

- First, we provide three types of evidence that support our methodology that categorizes situations using only characteristics of play in the previous round (see bullet points 1-3 below for a summary).
- Second, we provide evidence that supports the specific methodology that we use for categorizing situations according to characteristics of play in the previous round (see bullet points 4-5 below for a summary). As part of this evidence, we show that our results are robust when we expand the set of situations (from 11 to 101) by further categorizing situations using the average choice of the three group members in the previous round.

In Section 4 we categorized situations using only characteristics of play in the previous round. We provide three types of evidence that support this methodology:

1. In Online Appendix III.2 we use lasso regressions to show that choices in a particular round depend strongly on characteristics of the group of three subjects' play in the previous round, but depend little on characteristics of play in earlier rounds (after controlling for the characteristics of play in the previous round).
2. Our methodology that categorizes situations independently of the number of remaining rounds finds support from the evidence in Online Appendix II that forward-looking behavior is unlikely to be an important driver of behavior in our finitely repeated beauty contest setting.
3. Recalling that the feedback screen at the end of each round reported only characteristics of play in that round, evidence from the literature about limited memory and attention, which we summarize in Online Appendix III.3, also supports categorizing situations using only characteristics of play in the previous round.

In Section 4 we categorized situations according to: (i) the subject's earnings in the previous round; (ii) the rank-order of the choices of the three group members in the previous round; and (iii) whether the group played the Nash equilibrium in the previous round. This approach gave us the 11 situations described in Table 4.

4. In Online Appendix III.4 we provide evidence that our results are robust to expanding the set of situations: in particular, we show that our results are robust when we further categorize situations using the average choice of the three group members in the previous round, giving 101 situations (of which we observe 71 in our sample).
5. Evidence from other settings that people focus on rank-order, which we summarize in Online Appendix III.3, supports our use of the rank-order of choices in the previous round as one of the criteria to categorize situations.

Online Appendix III.2 Backward-looking behavior

In Section 4 we categorized situations using only characteristics of play in the previous round. In this section we provide support for this methodology by showing that choices in a particular round depend strongly on characteristics of the group of three subjects' play in the previous round, but depend little on characteristics of play in earlier rounds (after controlling for the characteristics of play in the previous round).

We address this problem using lasso regressions (Tibshirani, 1996; Hastie et al., 2019), and we also conduct a robustness check using OLS regressions. Lasso selects the best-fitting set of regressors while avoiding overfitting.[40] We base the set of potential regressors on the characteristics of play that were reported on the feedback screen at the end of each round. Recall, the feedback screen at the end of round r reported the following characteristics of play in round r: the numbers chosen by the subject, the subject's first opponent and the subject's second opponent; the amounts won by the subject, the subject's first opponent and the subject's second opponent; the average of the three chosen numbers; and 70% of the average of the three chosen numbers. Let Z_{r-s} denote the set of characteristics that were reported to subjects at the end of round $r-s$, and define $W_{r-s} \equiv \{j \in Z_{r-s};\ j \times j \text{ for } j \in Z_{r-s};$ and $j \times k$ for $j \in Z_{r-s}$, $k \in Z_{r-s}$ and $j \neq k\}$ (i.e., each of the characteristics of play in round $r-s$, together with the square of each characteristic and all cross products). We estimate two lasso regression models of choices in round r. In the first model, the potential regressors are W_{r-1} plus a round fixed effect, while in the second model, the potential regressors are W_{r-1}, W_{r-2} and W_{r-3} plus a round fixed effect.

We assess the predictive power of each model using the out-of-sample R-squared.[41] Our results show that the first lasso model predicts 21.74% of the variation in out-of-sample choices. The second lasso model predicts 22.10% of the variation in out-of-sample choices, which is only 0.36 percentage points more than the first model. In summary, regressors based on the characteristics of the group's play in rounds $r-2$ and $r-3$ have close-to-zero incremental explanatory power in a model of round r choices that includes regressors based on the characteristics of the group's play in round $r-1$.

As a robustness check, we conduct a similar exercise using OLS regressions. We proceed in the same way as when using lasso regressions except that: (i) we replace the lasso regressions with OLS regressions; and (ii) to avoid overfitting we replace W_{r-s} with Z_{r-s}. The results from the OLS regressions are similar to those from the lasso regressions.[42]

[40] Lasso avoids overfitting by estimating coefficients subject to a constraint on the sum of the absolute values of the coefficients. The STATA lasso package selects the constraint using k-fold cross-validation (we set $k = 5$ and cluster at the group level).

[41] Each time we estimate one of the models, we use a randomly selected 50% of the groups from the full sample of 260 groups of three, and we calculate the R-squared value in the hold-out sample containing the remaining 50% of groups. We estimate each model 25 times and calculate the mean of the out-of-sample R-squared values across the 25 estimates. When estimating the second model and calculating the R-squared value out-of-sample, we cannot use choices from rounds $r \leq 3$; to ensure a fair comparison between models we therefore also exclude these choices in the case of the first model.

[42] The first OLS regression model predicts 20.29% of the variation in out-of-sample choices, while the second OLS regression model predicts 21.58% of the variation. Because the OLS regressions allow a less flexible relationship between choices in round r and the characteristics of play in round $r-1$, we observe a slightly larger incremental explanatory power of the regressors based on the characteristics of play in rounds $r-2$ and $r-3$ as compared to the lasso regressions.

Online Appendix III.3 Memory, attention, and rank-order

In Section 4 we categorized situations using only characteristics of play in the previous round. Recalling that the feedback screen at the end of each round reported only characteristics of play in that round, evidence from the literature about limited memory and attention help to support our methodology:

- First, evidence shows that memory is limited. According to Miller (1956), the "finite span of immediate memory" imposes "severe limitations on the amount of information that we are able to receive, process, and remember," while Cowan (2001) provides a summary of the evidence on the limited capacity of short-term memory. Furthermore, subjects in games generally choose low memory strategies: for example, Dal Bó and Fréchette (2018)'s meta-analysis finds that simple memory-1 or memory-0 strategies (which require players to condition only on the history of play in the previous round) account for the vast majority of behavior in the indefinitely repeated prisoner's dilemma (see Table 10 together with Figure 5 in Dal Bó and Fréchette, 2018; see also Gill and Rosokha, 2020, who show that subjects often fail to choose lenient memory-2 strategies when they are optimal given their beliefs).
- Second, evidence also shows that attention is limited. Kahneman (1973) provides an early summary of research in psychology on selective attention and discusses research that highlights how limits on attention are driven by capacity constraints, while Handel and Schwartzstein (2018) survey more recent evidence from economics on limited attention driven by rational inattention and 'mental gaps'. For example, Handel and Schwartzstein (2018) discuss the findings of Grubb and Osborne (2015) whereby cell phone customers are inattentive to their past usage.

Furthermore, we used the rank-order of the group of three subjects' choices in the previous round as one of the criteria to categorize situations. This criterion is supported by evidence from other settings that people focus on rank-order:

- First, evidence from psychology and marketing suggests that consumers focus on the rank-order of prices (or other attributes) of goods within lists, while paying too little attention to the actual prices. For example, the compromise effect implies that a good becomes more attractive when a more expensive comparator good is added to the choice set (e.g., Kivetz et al., 2004), while the attraction effect implies that a good becomes more attractive when it ranks above an unchosen decoy good (e.g., Sivakumar, 2016).
- Second, recent evidence from economics shows that students focus on the rank-order of their ability within their classroom. In particular, using quasi-random variation in rank due to peer composition effects, Elsner and Isphording (2017) and Murphy and Weinhardt (2020) show that rank-order causally affects confidence, expectations, test scores, subject choice, and later educational outcomes.

Online Appendix III.4 Robustness to expanding the set of situations

In this section we provide evidence that our results are robust to expanding the set of situations.

In Section 4 we categorized situations according to: (i) the subject's earnings in the previous round; (ii) the rank-order of the choices of the three group members in the previous round; and (iii) whether the group played the Nash equilibrium in the previous round. This approach gave us the 11 situations described in Table 4, with a total of 7,012 subject-round observations across the 11 situations.

Here, we further categorize situations according to the average choice of the three group members in the previous round. Note that the situation in which a subject's group played the Nash equilibrium in the previous round (the 'equilibrium' situation) is already defined by an average choice of 0 (because all the choices were 0). We split each of the other 10 situations (the 'non-equilibrium' situations) into 10 new situations: (i) we bin average group choices in the previous round into ten bins, with bin $1 = (0, 10]$, bin $2 = (10, 20]$, and so forth; and (ii) we then split each of the 10 'non-equilibrium' situations described in Table 4 into ten new situations according to the bin that the average group choice in the previous round falls into.

This approach gives 101 situations in total: 10 new situations for each of the original 10 'non-equilibrium' situations, plus the 'equilibrium' situation. We observe 71 of these 101 situations in our sample of 7,012 subject-round observations that we use in Section 4. We then repeat the analyses from Section 4 and Section 6 using this expanded set of 71 situations and the same 7,012 subject-round observations as before. As we describe below, the results from Section 4 and Section 6 are robust to expanding the set of situations in this way.

First, the variation in strategic complexity across the expanded set of 71 situations continues to be statistically significant: the null hypothesis that strategic complexity is constant across situations is strongly rejected (an F-test returns $p = 0.000$); and this result holds whether or not we include the 'equilibrium' situation (in which the group played the Nash equilibrium in the previous round).

Second, the measures of strategic complexity for the expanded set of 71 situations are consistent with those from Table 4 for the original set of 11 situations. Specifically, Table OA.4 shows that when we calculate the average of the strategic complexities of the new situations within each of the original 10 'non-equilibrium' situations, we obtain average strategic complexities that are similar to the complexities reported in Table 4, while the complexity of the 'equilibrium' situation is also stable.

Third, Table OA.5 and Table OA.6 show that our results on the heterogeneous effect of strategic complexity on response time from Table 5 and Table 6 are robust when we use the expanded set of 71 situations.

Finally, Table OA.7 and Table OA.8 show that our results on the effect of thinking for longer than normal from Table 9 and Table 10 are robust when we use the expanded set of 71 situations. Recall that the estimates in Table 9 and Table 10 are identified from within-subject variation in response times across rounds in which the subject faced the same situation. When we expand the set of situations, we therefore lose some of the subject-round observations that we used in the analyses reported in Table 9 and Table 10. Despite the smaller sample, the coefficients in Table OA.7 and Table OA.8 all share the same signs as their equivalents in Table 9 and Table 10 (in fact, the estimates increase somewhat in magnitude, while unsurprisingly the standard errors also increase).

Original Situations from Table 4	Subject-round observations	Strategic Complexity (Average response time in seconds)
$n = \underline{n}^o = \overline{n}^o = 0$ ($\therefore$ subject earned \$2)	318	15.674 ‡
$\underline{n}^o < \overline{n}^o < n$ ($\therefore$ subject earned \$0)	1,739	25.170 †
$\underline{n}^o = \overline{n}^o < n$ ($\therefore$ subject earned \$0)	273	25.491 †
$n < \underline{n}^o < \overline{n}^o$ & subject earned \$0	633	25.984 †
$\underline{n}^o < n < \overline{n}^o$ & subject earned \$0	1,102	26.248 †
$n = \underline{n}^o = \overline{n}^o > 0$ ($\therefore$ subject earned \$2)	123	26.718 †
$\underline{n}^o < n = \overline{n}^o$ ($\therefore$ subject earned \$0)	362	26.744 †
$n < \underline{n}^o = \overline{n}^o$ ($\therefore$ subject earned \$6)	181	27.020 †
$n < \underline{n}^o < \overline{n}^o$ & subject earned \$6	1,102	27.795 †
$n = \underline{n}^o < \overline{n}^o$ ($\therefore$ subject earned \$3)	546	28.526 †
$\underline{n}^o < n < \overline{n}^o$ & subject earned \$6	633	30.228 †

‡ The 'equilibrium' situation remains a situation in the expanded set.
† Average strategic complexity of the new situations within the original situation.
Notes: See the notes to Table 4. First, using the methodology explained in Section 4.2, we measure the strategic complexity of each of the 71 situations described in the third and fourth paragraphs of this section. For the purposes of this table, we then calculate the average of the strategic complexities of the new situations within an original situation (weighting the strategic complexity of each new situation by the empirical frequency of that new situation within the original situation in our sample of 7,012 subject-round observations).

Table OA.4: Robustness of the results in Table 4 to expanding the set of situations

	Type 1	Type 2
β	1.380***	0.182***
	(0.134)	(0.044)
σ	22.915***	5.835***
	(0.410)	(0.315)
π (type probability)	0.633***	0.367***
	(0.022)	(0.022)

Notes: Parameter estimates were obtained by applying Maximum Likelihood estimation using the expanded set of 71 situations and sample of 7,012 subject-round observations described in the third and fourth paragraphs of this section. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests for β and one-sided tests elsewhere).

Table OA.5: Robustness of the results in Table 5 to expanding the set of situations

	Type 1	Type 2	Type 3
β	1.484***	0.243***	0.052*
	(0.152)	(0.065)	0.031
σ	23.647***	7.631***	2.131***
	(0.391)	(0.411)	(0.192)
π (type probability)	0.582***	0.312***	0.106***
	(0.023)	(0.019)	(0.014)

Notes: Parameter estimates were obtained by applying Maximum Likelihood estimation using the expanded set of 71 situations and sample of 7,012 subject-round observations described in the third and fourth paragraphs of this section. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests for β and one-sided tests elsewhere).

Table OA.6: Robustness of the results in Table 6 to expanding the set of situations

	Winner of round	Earnings in round	Log earnings in round
Effect of thinking longer	-0.105**	-42.289*	-0.223**
than normal (minutes)	(0.041)	(21.634)	(0.098)
Subject-round observations	1,925	1,925	1,925

Notes: See the notes to Table 9. Starting from the expanded set of 71 situations and sample of 7,012 subject-round observations described in the third and fourth paragraphs of this section, we use the subject-round observations for which the subject is observed in the same situation in at least one other round (this ensures that subject-round observations are not fully absorbed by the fixed effects). This gives us 1,925 subject-round observations and 860 subject-situation fixed effects. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table OA.7: Robustness of the results in Table 9 to expanding the set of situations

	Choice in round	Choose zero in round	Increased choice in round
Effect of thinking longer	4.176**	-0.067***	0.095***
than normal (minutes)	(1.722)	(0.022)	(0.033)
Subject-round observations	1,925	1,925	1,925

Notes: See the notes to Table 10. Starting from the expanded set of 71 situations and sample of 7,012 subject-round observations described in the third and fourth paragraphs of this section, we use the subject-round observations for which the subject is observed in the same situation in at least one other round (this ensures that subject-round observations are not fully absorbed by the fixed effects). This gives us 1,925 subject-round observations and 860 subject-situation fixed effects. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table OA.8: Robustness of the results in Table 10 to expanding the set of situations

Online Appendix IV More on descriptive statistics

In Section 3 we provided descriptive statistics. In this appendix, we expand upon these descriptive statistics.

Online Appendix IV.1 Response times across rounds: Related literature

Figure 1(a) in Section 3 shows that the average response time varies little across the ten rounds of our beauty contest experiment. Studies that report the trend in response times in repeated games are surprisingly rare. Overall the evidence from these studies is mixed, which suggests that details of the experimental setting (e.g., the type of game that is repeated, the number of players and repetitions, the speed of convergence toward equilibrium, and the format in which feedback is provided) determine the trend in response times across rounds in ways that we do not yet understand well. The number of relevant studies is too small to come to any firm conclusions, but the data suggest that response times fall sharply with repetition in two-by-two matrix games, while response times are much more stable in games with larger action spaces.

Evans et al. (2015) study a repeated public goods game (36 rounds in groups of 4 with 21 possible stage-game actions) and find that response times decreased only very modestly across rounds. In particular, log-transformed response times decreased by 0.10 over the 36 rounds (the mean log response time was 0.59); in a separate regression, the decrease in mean response times across rounds was not statistically significant at the 5% level. By contrast, Proto et al. (2019) find a more substantial decline in response times across repetitions of the prisoner's dilemma, the battle of the sexes, and the stag hunt games (in all cases two-by-two stage games with random termination).

In contrast to the games studied by Evans et al. (2015) and Proto et al. (2019), the beauty contest that we study is a repeated competitive constant-sum game. Hardly any studies report the trend in response times in this type of setting. As noted by Spiliopoulos (2018): "To the best of my knowledge there are no other published studies that examine RT in a strictly competitive repeated game where social preferences are irrelevant. The closest work is a working paper by Gill and Prowse..." Spiliopoulos (2018) studies a repeated constant-sum two-by-two matrix game and finds that response times declined over rounds for almost all of the 31 subjects. The setting is quite different from ours: (i) the game is played against a computer rather than a human opponent; (ii) the stage-game action space is small; (iii) the game is played for hundreds of rounds and average response times are short (< 4s); and (iv) the unique Nash equilibrium is in mixed strategies.

Finally, Kocher and Sutter (2006) report response times in a repeated modified beauty contest in which there is no winner, and instead each subject's stage-game payoff depends linearly on the distance between that subject's guess and the target. In each of three phases, subjects repeat the game eight times in groups of four, with the three phases varying in how the target is calculated. In one treatment, the decision-making time limit was short (15s) and in the other long (120s). Some of the results provide minimal support for declining response times over rounds: (i) in the treatment with a 15s time limit, the mean response time was "rather stable at about 10s"; (ii) in the treatment with a 120s time limit, the mean response time in phase 1 was only slightly higher than in phase 2 (42s vs. 39s); and (iii) in the treatment with a 120s time limit, response times were not monotonic in the round number (mean response time in round 2 was higher than in round 1, significant at the 5% level). Having said this, in the treatment with a 120s time limit, there was a tendency for response times to fall over rounds within a phase; however, this might have followed from the fact that choices converged toward equilibrium much faster than in our setting.

Online Appendix IV.2 More on choices and response times

Figure 1(a) in Section 3 presents average response times across rounds. Here, Figures OA.1, OA.2 and OA.3 further present average choices across rounds as well as the variance of choices and response times across rounds.

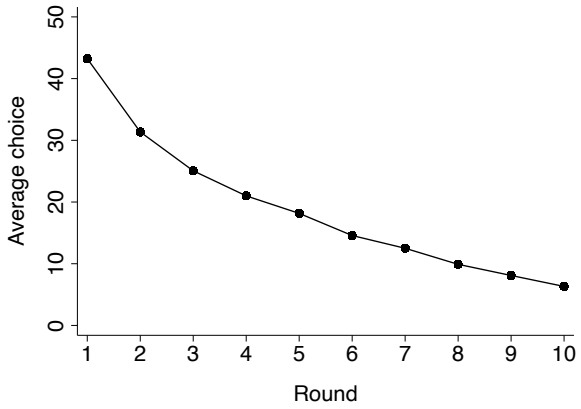


Figure OA.1: Average choice in rounds 1–10

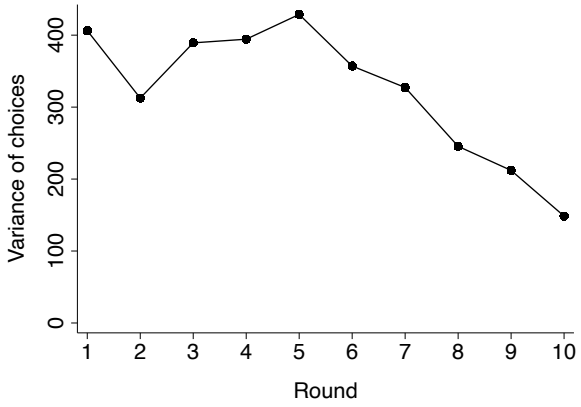


Figure OA.2: Variance of choices in rounds 1–10

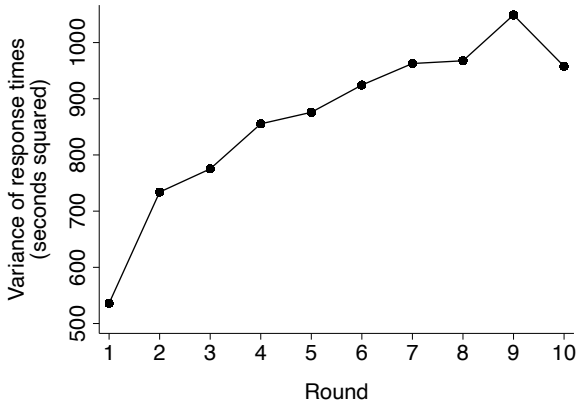


Figure OA.3: Variance of responses times in rounds 1–10

Online Appendix IV.3 Table 1 and Table 2: Robustness

In Table 1 in Section 3 we run a between-subject analysis in which we regress measures of success and choices in the experiment on the subject-level average response time. Because this analysis uses averages of success, choices and response times across all ten rounds, we cannot include a round number in these regressions. Instead, to understand whether the relationships studied in Table 1 change across rounds, here Table OA.9 replicates Table 1 using data only from the first five rounds, and Table OA.10 replicates Table 1 using data only from the last five rounds. Comparing Tables OA.9 and OA.10 to Table 1, we conclude that the relationships in Table 1 hold in both the first and second halves of the experiment. There is one exception: the relationship between response time and choices is no longer statistically significant in rounds 6-10, but of course choices compress toward zero in the last five rounds (see Figure OA.1 in Online Appendix IV.2).

Furthermore, the null results from Table 2 in Section 3 extend when we consider only the first five rounds (Table OA.11) or only the last five rounds (Table OA.12).

In summary, the results in Table 1 and Table 2 show stability from the first half to the second half of the experiment. We further note that subjects converge toward Nash equilibrium, but slowly: (i) the equilibrium choice of zero is played at a rate of 0.5% in rounds 1-5 and 11.7% in rounds 6-10; and (ii) as reported in Online Appendix VI.1, the 'equilibrium' situation, that is the situation in which a subject's group played the Nash equilibrium in the previous round, occurs in only 318 of the 7,800 subject-round observations.

	Success and choices in rounds 1–5			
	Fraction of rounds won	Earnings per round (cents)	Log earnings per round (cents)	Average choice
Average response time in rounds 1–5 (minutes)	0.054**	30.424**	0.134**	-3.340***
	(0.025)	(14.095)	(0.062)	(1.131)
Intercept	0.340***	186.164***	4.745***	29.280***
	(0.012)	(6.430)	(0.028)	(0.774)
Subjects	780	780	780	780

Notes: We run the same regressions as reported in Table 1 except that all averages are calculated using data from rounds 1–5 only. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table OA.9: Average response time, success and choices in rounds 1–5

	Success and choices in rounds 6–10			
	Fraction of rounds won	Earnings per round (cents)	Log earnings per round (cents)	Average choice
Average response time in rounds 6–10 (minutes)	0.042*	42.118***	0.150***	-0.309
	(0.024)	(10.208)	(0.049)	(0.984)
Intercept	0.474***	180.959***	4.892***	10.420***
	(0.018)	(4.676)	(0.028)	(0.687)
Subjects	780	780	780	780

Notes: We run the same regressions as reported in Table 1 except that all averages are calculated using data from rounds 6–10 only. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table OA.10: Average response time, success and choices in rounds 6–10

	Average response time in rounds 1–5 (minutes)		
Raven test score (cognitive ability)	-0.005		0.006
	(0.012)		(0.019)
Personality factor 1 (conscientiousness, grit and future orientation)		-0.026	-0.025
		(0.022)	(0.022)
Personality factor 2 (agreeableness and emotional stability)		-0.002	-0.003
		(0.018)	(0.018)
Personality factor 3 (openness, extraversion and future orientation)		-0.018	-0.018
		(0.021)	(0.021)
Intercept	0.455***	0.465***	0.465***
	(0.013)	(0.020)	(0.020)
Subjects	780	270	270

Notes: We run the same regressions as reported in Table 2 except that average response times are calculated using data from rounds 1–5 only. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table OA.11: Raven test score, personality, and average response time in rounds 1–5

	Average response time in rounds 6–10 (minutes)		
Raven test score	-0.008		0.018
(cognitive ability)	(0.016)		(0.023)
Personality factor 1		-0.029	-0.026
(conscientiousness, grit and future orientation)		(0.026)	(0.026)
Personality factor 2		0.015	0.014
(agreeableness and emotional stability)		(0.022)	(0.023)
Personality factor 3		-0.011	-0.011
(openness, extraversion and future orientation)		(0.024)	(0.024)
Intercept	0.452***	0.420***	0.420***
	(0.016)	(0.024)	(0.024)
Subjects	780	270	270

Notes: We run the same regressions as reported in Table 2 except that average response times are calculated using data from rounds 6–10 only. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table OA.12: Raven test score, personality, and average response time in rounds 6–10

Online Appendix V More on level-k

Online Appendix V.1 Level-k model

In a setting where subjects play the beauty contest game a single time, Alós-Ferrer and Buckenmaier (2021) allocate each subject to a level-k type for $k \in \{0, 1, ..., 6\}$ by calculating the quadratic distances between the subject's choice and each of the level-k choices, and then choosing k to minimize the quadratic distance. We extend Alós-Ferrer and Buckenmaier (2021)'s quadratic distance methodology to our repeated beauty contest setting by allowing level-k types to learn over time.

In particular, following Gill and Prowse (2016), the level-k choice in round $r > 1$ is given by a fraction $(0.7)^k$ of the mean of the choices in the subject's group g in the previous round, $\overline{\text{Choice}}_{g,r-1}$. As explained by Gill and Prowse (2016), a subject who selects the level-0 choice learns in a strategically unsophisticated adaptive manner by "following the crowd" and copying their group's average choice from the previous round, while subjects who select the level-$k > 0$ choice anticipate how lower-level agents learn across rounds.[43]

Specifically, we allocate subject i to a level-k type by choosing k to minimize the sum over rounds 2 to 10 of the quadratic distances between subject i's choice in round r, $\text{Choice}_{i,r}$, and the level-k choice described in the previous paragraph. That is, we choose $k \in \{0, 1, ..., 6\}$ to minimize:

$$f_i(k) = \sum_{r=2}^{10} (\text{Choice}_{i,r} - (0.7)^k \overline{\text{Choice}}_{g,r-1})^2. \tag{7}$$

In this level-k model with learning, level-k choices depend on behavior in the previous round, and so $f_i(k)$ sums quadratic distances starting from round 2; as in Gill and Prowse (2016), choices in round 1 seed the model by determining the level-k choices in round 2.

Occasionally, subjects choose high numbers far from any level-k choice: to avoid such choices influencing the type classification, when calculating $f_i(k)$ we exclude rounds in which subject i's choice was more than double the level-0 choice (we are still able to classify all 780 subjects). Finally, we note that for all subjects, a unique k minimizes $f_i(k)$.

[43] Note that a subject who selects the level-$k > 0$ choice does not take into account the effect of her own choice on the target. Gill and Prowse (2016) find that an alternative specification of their model in which subjects account for the effect of their own choice on the target fits less well.

Online Appendix V.2 Tables with level-k, cognitive ability and personality

	Average response time (minutes)
Raven test score (cognitive ability)	-0.016 (0.020)
Level≥2	0.057** (0.029)
Level≥2 × Raven test score	0.003 (0.027)
Personality factor 1 (conscientiousness, grit and future orientation)	-0.024 (0.024)
Personality factor 2 (agreeableness and emotional stability)	0.005 (0.019)
Personality factor 3 (openness, extraversion and future orientation)	-0.014 (0.022)
Intercept	0.405*** (0.025)
Subjects	780

Notes: Here we add our three personality factors to the regression reported in the final column of Table 3 (see the notes to that table). Personality factors 1-3 have been standardized to have means of zero and standard deviations of one (Gill and Prowse, 2016, describe the construction of the personality factors). Personality was measured for 270 of our 780 subjects: here missing values for personality are arbitrarily coded as zero and we include an indicator for missing personality. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table OA.13: Level-k type, Raven test score and average response time: Robustness to controlling for personality

	Type 1	Type 2
Panel A: Heterogeneous effect of strategic complexity on response time		
β	1.309***	0.166***
	(0.179)	(0.057)
σ	23.092***	5.842***
	(0.419)	(0.325)
Average π	0.634***	0.366***
(average type probability)	(0.022)	(0.022)
Panel B: Average marginal effects on type probabilities		
Raven test score	-0.011	0.011
(cognitive ability)	(0.019)	(0.019)
Level-k type	0.031**	-0.031**
	(0.014)	(0.014)
Personality factor 1	-0.036	0.036
(conscientiousness, grit and future orientation)	(0.032)	(0.032)
Personality factor 2	0.029	-0.029
(agreeableness and emotional stability)	(0.032)	(0.032)
Personality factor 3	-0.004	0.004
(openness, extraversion and future orientation)	(0.033)	(0.033)

Notes: Here we add our three personality factors to Model 3 from Table 7 (see the notes to that table). Personality factors 1-3 have been standardized to have means of zero and standard deviations of one (Gill and Prowse, 2016, describe the construction of the personality factors). Personality was measured for 270 of our 780 subjects: here missing values for personality are arbitrarily coded as zero and we include an indicator for missing personality. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests for β and the average marginal effects, and one-sided tests elsewhere).

Table OA.14: Level-k type, Raven test score, and the heterogeneous effect of complexity on response time: Robustness to controlling for personality

Online Appendix VI Further discussion of Table 4

Online Appendix VI.1 Response times across rounds: Robustness

Table 4 in Section 4.2 shows that the average response time in the situation in which a subject's group played the Nash equilibrium in the previous round (the 'equilibrium' situation) is substantially lower than in the other ten situations (the 'non-equilibrium' situations). Furthermore, the equilibrium situation tends to occur in the later rounds of the experiment. Despite this, Figure 1(a) in Section 3 shows that the average response time varies little across the ten rounds (as reported in footnote 23, the coefficient on the linear round trend is not statistically significantly different from zero).

To understand this better, Figure OA.4 presents the average response time across rounds for: (i) the full sample (reproducing the data from Figure 1(a)); and (ii) excluding the equilibrium situation (that is, excluding the 318 subject-round observations where the subject's group played the equilibrium in the previous round). Figure OA.4 shows that the time trends are similar: this is because the equilibrium situation occurs in only 318 of the 7,800 subject-round observations.[44]

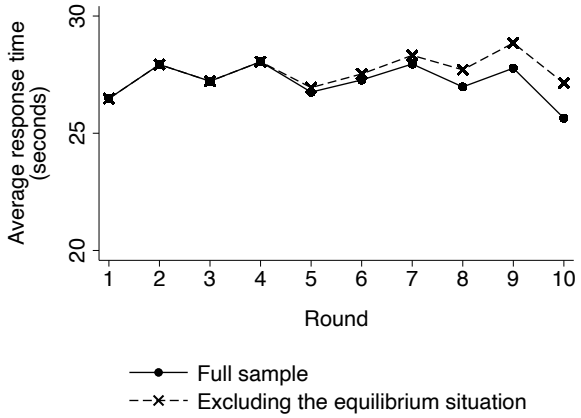


Figure OA.4: Robustness of Figure 1(a) to excluding the equilibrium situation

Online Appendix VI.2 Response times preceding the equilibrium situation

Table 4 in Section 4.2 shows that the average response time in the situation in which a subject's group played the Nash equilibrium in the previous round (the 'equilibrium' situation) is substantially lower than in the other ten situations (the 'non-equilibrium' situations). The equilibrium is 'sticky': when a group plays the equilibrium in a particular round, 90 percent of the time that group continues to play the equilibrium in the next round. We interpret this evidence as showing that once a group has reached the equilibrium, the equilibrium logic becomes obvious and choices are quick.

This raises an interesting question: do response times tend to be lower when a group approaches the equilibrium or when the group plays the equilibrium for the first time? In short, the answer is no. The average response time in the round in which a group first plays the equilibrium (27s) and in the round before that (29s) fall within the range of the average response times in the non-equilibrium situations reported in Table 4.

[44] Note that equilibrium choices in round 10 do not create equilibrium situations in the following round since the 10th round is the final round.

Online Appendix VI.3 Further discussion of differences in strategic complexity

Table 4 in Section 4.2 reveals that strategic complexity, measured by average thinking time, varies substantially across the eleven situations.

As we say in Section 4.2: "The most complex situation is where the subject won the entire prize of six dollars in the previous round with a choice that was between that of her opponents; in this situation subjects think for an average of thirty seconds. The least complex situation is where the group was at the Nash equilibrium in the previous round (that is, all three group members chose zero); in this situation subjects think for an average of fifteen seconds."

It is intuitive that choices are quick when the group was at the Nash equilibrium in the previous round because the equilibrium logic likely becomes obvious. We discuss this in Online Appendix VI.2, where we say: "Table 4 in Section 4.2 shows that the average response time in the situation in which a subject's group played the Nash equilibrium in the previous round (the 'equilibrium' situation) is substantially lower than in the other ten situations (the 'non-equilibrium' situations). The equilibrium is 'sticky': when a group plays the equilibrium in a particular round, 90 percent of the time that group continues to play the equilibrium in the next round. We interpret this evidence as showing that once a group has reached the equilibrium, the equilibrium logic becomes obvious and choices are quick."

It is also intuitive that choices are slow when the subject won the entire prize of six dollars in the previous round with a choice that was between that of her opponents. When engaging in iterative reasoning, subjects in games tend to conceive of their opponents as behaving similarly to each other: "players tend to treat opponents as a single aggregate player" (Ho et al., 2021). Subjects who think in this way and understand the payoff structure of the p-beauty contest decide what choice they expect from their opponents and undercut that choice by an appropriate amount; when the subject then wins unexpectedly with the middle choice, she needs to reconsider how she thinks about the game and her opponents.

Turning to the intermediate situations between the most and least complex, Table 4 shows that situations in which the subject chose a number lower than those of her opponents in the previous round are uniformly more complex than situations in which the subject chose a number higher than those of her opponents. After choosing the lowest number, the subject usually wins in that round: to stay one step ahead of her opponents, it is intuitive that the subject tends to think carefully about how much she expects her opponents to reduce their choices. On the other hand, choosing the highest number guarantees that the subject loses in that round: likely the subject will quickly see that she needs to move her choice down, but she may not have good insights about how much she should lower her choice.

Online Appendix VII Discussion of the matching procedure

As noted in Section 2, in the experiment subjects were matched into groups of three to play the p-beauty contest according to their Raven test score. The matching procedure is described in detail in Gill and Prowse (2016). Here we briefly discuss some aspects of the design that relate to the matching procedure.

We did not tell the subjects that we would use their Raven test score to match them into groups to play the p-beauty contest, and nor did we pay subjects for their performance in the Raven test (subjects were not made aware of what would follow the Raven test until the test was completed). We made these choices in order to keep the psychometric evaluation procedure as similar as possible to that used in the psychometric literature, where subjects complete the Raven test without any consideration of incentives.

Although we matched subjects according to whether their Raven test score was in the top half (H) or the bottom half (L) of the test scores in the session, our design created all possible combinations (groups of three H, groups of three L, mixed groups with two H and one L, mixed groups with one H and two L). We created somewhat more groups with three H subjects or three L subjects than would have occurred if we had randomly matched the subjects. However, on average, H subjects still had a substantial likelihood (42%) of being matched with one or more L subjects, and vice-versa. Furthermore, before playing the p-beauty contest, we told each subject whether their own test score was in the top or bottom half of the test scores in the session, and whether each of the subject's opponents' scores was in the top or bottom half.

We feel that our design strikes a good balance between simplicity, external validity with respect to the psychometric evaluation, and provision of relevant information to the subjects. We emphasize that at no point did we provide false information to our subjects, and the subjects discovered whether their own and their opponents' test scores were above or below average before starting to play the game.

Examples of papers in top economics journals that match subjects by cognitive ability include Proto et al. (2019) and Proto et al. (forthcoming). Like in our setting, subjects first complete a Raven test, and are then matched according to their Raven test score before playing a strategic game. Proto et al. (2019) and Proto et al. (forthcoming) never mention anything about the matching to the subjects: unlike in our setting, they never disclose anything to the subjects about their own test score or that of their opponents. Furthermore, in Proto et al. (2019) almost all subjects are matched with opponents with test scores from the same half of the test score distribution: unlike in our setting, H subjects are almost always matched with another H subject, and vice-versa.

Wilson (2014) discusses literature in economics where, instead of matching using a personal characteristic like cognitive ability, subjects are matched based on previous game choices without revealing anything about this matching procedure to the subjects. For example, in a paper by Vernon L. Smith and co-authors in the *Economic Journal* (Rigdon et al., 2007), subjects play a trust game with rematching, and in one treatment subjects are matched each round according to their previous levels of trust (with more trusting first movers matched with more trustworthy second movers).

Online Appendix VIII Round fixed effects

Table OA.15 shows the estimates of the round fixed effects in equation (1). These round fixed effects capture any trends in response times over the experiment after accounting for the effect of the strategic complexity of the situations faced by the subjects in each round. The estimates show that the round fixed effects in equation (1) do not follow a systematic trend across rounds.

Table OA.16 shows the estimates of the round fixed effects in the two-type mixture model reported in Table 5 and the three-type mixture model reported in Table 6. In both models, the round fixed effects appear to first decline over the first few rounds and then stabilize. We report these estimates for completeness, but we interpret them with caution because these mixture models impose a specific functional form on the distribution of the error term (in particular normality), which entails that round-individual specific outliers in response times can have substantial effects on the estimates of the round fixed effects.

Table OA.17 shows the estimates of the round fixed effects in the regressions reported in Table 9. Leaving aside round 2, the round fixed effects for "winner of round" increase in the final rounds: this reflects that a subject is defined to be "winner of round" if she won all or part of the prize, and prizes are split more frequently in later rounds as choices compress toward zero. Again leaving aside round 2, the round fixed effects for "earnings in round" and "log earnings in round" do not follow a systematic trend across rounds (compared to the round 5 reference category, the estimates after round 2 are all far from statistical significance). Finally, the round 2 fixed effects are all statistically significantly negative (compared to the round 5 reference category): this reflects lower earnings in round 2 at the subject-situation level compared to when the subject faces the same situation later; recall however that Table A.4 in the Appendix shows that the results in Table 9 are robust to including controls for the subject's experience of her current situation.

Table OA.18 shows the estimates of the round fixed effects in the regressions reported in Table 10. Consistent with the gradual convergence of choices toward equilibrium in the data, the round fixed effects for "choice in round" decline over rounds while the round fixed effects for "choose zero in round" increase over rounds. Finally, the round fixed effects for "increased choice in round" do not follow a systematic trend across rounds (the estimates are similar in rounds 2, 6 and 10). This result is consistent with Online Appendix II where we provide evidence that forward-looking behavior is unlikely to be an important driver of behavior in our finitely repeated beauty contest setting: if subjects tried to manipulate their opponents by choosing a high number in an attempt to induce their opponents to believe that they will continue to choose a high number in later rounds, then we would expect a declining trend in the round fixed effects for "increased choice in round" as the number of future rounds falls.

Round 2	1.201
	(1.090)
Round 3	0.419
	(0.898)
Round 4	1.205
	(0.953)
Round 5 (reference category)	–
Round 6	0.594
	(0.810)
Round 7	1.626*
	(0.952)
Round 8	0.834
	(1.106)
Round 9	1.919*
	(1.127)
Round 10	0.455
	(1.127)

Notes: There is no round 1 fixed effect because we categorize situations according to the characteristics of play in the previous round, and so equation (1) is defined from round 2 onward. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table OA.15: Round fixed effects in equation (1)

	Two-type model	Three-type model
Round 2	1.704***	1.245***
	(0.381)	(0.301)
Round 3	0.724*	0.767***
	(0.383)	(0.251)
Round 4	0.654*	0.147
	(0.383)	(0.249)
Round 5 (reference category)	–	–
Round 6	-0.599*	-0.772***
	(0.336)	(0.170)
Round 7	-0.351	-0.296
	(0.308)	(0.188)
Round 8	-0.429	-0.206
	(0.391)	(0.251)
Round 9	-0.483	-0.109
	(0.365)	(0.295)
Round 10	-0.636*	-0.513**
	(0.326)	(0.207)

Notes: There is no round 1 fixed effect because we categorize situations according to the characteristics of play in the previous round, and so equation (3) is defined from round 2 onward. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table OA.16: Round fixed effects in the models reported in Table 5 and Table 6

	Winner of round	Earnings in round	Log earnings in round
Round 2	-0.067***	-22.400**	-0.134***
	(0.023)	(10.520)	(0.049)
Round 3	-0.030	-0.352	-0.037
	(0.023)	(10.455)	(0.048)
Round 4	-0.025	-5.346	-0.044
	(0.022)	(10.428)	(0.047)
Round 5 (reference category)	–	–	–
Round 6	-0.000	-1.118	-0.000
	(0.024)	(11.046)	(0.051)
Round 7	0.028	-6.773	0.021
	(0.024)	(9.981)	(0.048)
Round 8	0.036	-15.156	0.011
	(0.027)	(11.271)	(0.053)
Round 9	0.050*	-12.827	0.031
	(0.026)	(11.382)	(0.053)
Round 10	0.060**	-13.143	0.044
	(0.025)	(10.796)	(0.051)

Notes: There is no round 1 fixed effect because we categorize situations according to the characteristics of play in the previous round, and so equation (6) is defined from round 2 onward. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table OA.17: Round fixed effects in the regressions reported in Table 9

	Choice in round	Choose zero in round	Increased choice in round
Round 2	13.132***	-0.032**	-0.065**
	(1.401)	(0.013)	(0.027)
Round 3	6.244***	-0.026*	-0.012
	(1.303)	(0.015)	(0.028)
Round 4	2.773**	-0.020	0.015
	(1.333)	(0.013)	(0.026)
Round 5 (reference category)	–	–	–
Round 6	-4.853***	0.003	-0.064**
	(1.237)	(0.016)	(0.025)
Round 7	-6.792***	0.003	-0.030
	(1.302)	(0.014)	(0.028)
Round 8	-9.000***	0.018	-0.071***
	(1.316)	(0.018)	(0.027)
Round 9	-11.265***	0.033*	-0.097***
	(1.428)	(0.017)	(0.026)
Round 10	-13.204***	0.062***	-0.076***
	(1.238)	(0.018)	(0.026)

Notes: There is no round 1 fixed effect because we categorize situations according to the characteristics of play in the previous round, and so equation (6) is defined from round 2 onward. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table OA.18: Round fixed effects in the regressions reported in Table 10

Online Appendix IX Quantifying the importance of response times

In this appendix, we quantify the importance of response times in two ways: (i) we show that the variation that we find in average response times across situations is substantial in the context of related literature that we cite in the paper; and (ii) we show that the absolute magnitudes of the effects of response times are broadly comparable to the magnitudes of the effects of cognitive ability, which is considered an important predictor in the beauty contest game.

Table 4 in Section 4.2 reveals that strategic complexity, measured by average response time, varies by around five seconds between the ten 'non-equilibrium' situations (the average response time in the 'equilibrium' situation, in which a subject's group played the Nash equilibrium in the previous round, is substantially lower). In the context of related literature that we cite in the main body of the paper, this five-second difference in response times is substantial:

- In variants of the 11-20 game, Alós-Ferrer and Buckenmaier (2021) find that doubling incentives (by doubling the bonus) changes response times by around 2 seconds (relative to an average response time of around 10 seconds).[45]
- In Frydman and Krajbich (forthcoming)'s experiment on information cascades, a five-second increase in a subject's response time increases the probability that the next subject in the sequence chooses to follow their own signal by 14.8 percentage points.[46]
- As we describe in footnote 16, related literature in economics includes treatments that constrain total response time to just fifteen seconds.
- As noted by Rubinstein (2016), much of the psychology literature that studies response times attaches importance to response times measured in fractions of seconds.

Table 1 in Section 3 and Table 9 in Section 6 describe the relationship between response times and success: using between-subject variation, Table 1 shows that subjects who think for longer on average are more successful; and using within-subject variation, Table 9 shows that when a subject thinks for longer than normal in a particular situation, she is less successful. To quantify the importance of response times, we show that the absolute magnitudes of the effects in Tables 1 and 9 are broadly comparable to the magnitudes of the effects of cognitive ability, which is considered an important predictor in the beauty contest game (e.g., Burnham et al., 2009, Brañas-Garza et al., 2012, Fehr and Huck, 2016, Gill and Prowse, 2016, Alós-Ferrer and Buckenmaier, 2021; see also Fe et al., forthcoming, in the 11-20 game). Tables OA.19 and OA.20 report the comparison: to make the effect sizes comparable, we standardize both response times and cognitive ability, so that the effect sizes can be interpreted as the effect of a one-standard-deviation increase in the relevant variable (thus the effect sizes in Panel A of Table OA.19 (Table OA.20) are rescaled versions of the effect sizes in Table 1 (Table 9)).

[45] We use the original data from Alós-Ferrer and Buckenmaier (2021) to calculate the average response times for the samples used in Alós-Ferrer and Buckenmaier (2021)'s Tables 5 and 6.

[46] Column (1) of Table 1 from Frydman and Krajbich (forthcoming) shows that, when a subject's private signal does not match their predecessor's move, the predecessor's response time has a statistically significant positive effect on the probability that the subject chooses to follow their own signal: the effect size that we report here is the average marginal effect of a five-second increase in response time, starting at the median response time, based on the model in Column (1) of Table 1. To calculate this effect size we used the original data from Frydman and Krajbich (forthcoming)'s experiment.

	Fraction of rounds won	Earnings per round (cents)	Log earnings per round (cents)	Average choice
Panel A:				
Average response time	0.022***	13.992***	0.059***	-0.735**
	(0.007)	(3.313)	(0.015)	(0.320)
Panel B:				
Raven test score (cognitive ability)	0.033***	6.251**	0.054***	-1.594***
	(0.007)	(3.109)	(0.015)	(0.380)
Subjects (Panels A & B)	780	780	780	780

Notes: In Panel A we modify the regressions reported in Table 1 by replacing average response time in minutes with a standardized average response time variable. In particular, we standardize average response time by dividing this variable by the standard deviation of average response time. In Panel B we regress each outcome on the standardized Raven test score. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table OA.19: Comparison to the effect of cognitive ability for Table 1 (all effect sizes are standardized)

	Winner of round	Earnings in round	Log earnings in round
Panel A:			
Effect of thinking longer than normal	-0.029**	-13.448*	-0.065**
	(0.014)	(7.401)	(0.033)
Panel B:			
Raven test score (cognitive ability)	0.037***	6.357	0.058***
	(0.010)	(3.942)	(0.019)
Subject-round observations (Panels A & B)	4,774	4,774	4,774

Notes: In Panel A we modify the regressions reported in Table 9 by replacing response time in minutes with a standardized response time variable. In particular, we standardize response time by dividing this variable by the standard deviation of the response times for the 4,774 subject-round observations that we use in Table 9. In Panel B we use the same subject-round observations as in Panel A, and regress each outcome on the standardized Raven test score. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table OA.20: Comparison to the effect of cognitive ability for Table 9 (all effect sizes are standardized)

Online Appendix X Additional tables

	Model 1		Model 2		Model 3	
	Type 1	Type 2	Type 1	Type 2	Type 1	Type 2
Panel A: Heterogeneous effect of strategic complexity on response time						
β	1.310***	0.166***	1.313***	0.165***	1.311***	0.166***
	(0.179)	(0.057)	(0.179)	(0.057)	(0.179)	(0.057)
σ	23.088***	5.839***	23.105***	5.855***	23.090***	5.840***
	(0.422)	(0.330)	(0.419)	(0.327)	(0.422)	(0.331)
Average π	0.634***	0.366***	0.633***	0.367***	0.634***	0.366***
(average type probability)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)
Panel B: Average marginal effects on type probabilities						
Fraction of rounds won	0.059	-0.059				
	(0.087)	(0.087)				
Earnings per round			-0.014	0.014		
(standardized)			(0.020)	(0.020)		
Log earnings per round					0.008	-0.008
(standardized)					(0.019)	(0.019)

Notes: Here we extend the two-type mixture regression model from Table 5 by allowing the type probability to depend on metrics of the subject's success according to a logistic distribution function. All success metrics are calculated using data from rounds 2–10 (this matches the estimation of the mixture model that uses response times only from round 2 onward because situations, and thus complexity, are defined only from round 2 onward). When calculating the fraction of rounds won, a subject is considered to be a winner if she won all or part of the prize. Earnings per round and log earnings per round are subject-level averages. When taking the log of earnings in round r, we add fifty cents to earnings (the show-up fee of five dollars divided by the number of rounds) to avoid taking the log of zero. The average type probability was obtained by computing the type probability for each subject, conditional on the relevant metric of the subject's success, and then averaging over the subjects. The average marginal effects are averages of the individual-level marginal effects. As in Table 5, parameter estimates were obtained by applying Maximum Likelihood estimation to the sample of 7,012 subject-round observations described in Table 4. Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests for β and the average marginal effects, and one-sided tests elsewhere).

Table OA.21: Success and the heterogeneous effect of strategic complexity on response time

	Choice in round	Choose zero in round	Increased choice in round
Effect of thinking longer than normal (minutes)	2.440*	-0.060***	0.068*
	(1.422)	(0.022)	(0.039)
Subject-round observations	1,793	1,793	1,793

Notes: We run the same regressions as reported in Table 10 except that we use only subject-round observations from situations where the subject won all or part of the prize in the previous round (see Table 4 for the list of situations). Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table OA.22: Effect of thinking for longer than normal on strategic behavior: including only situations where the subject won in the previous round

	Choice in round	Choose zero in round	Increased choice in round
Effect of thinking longer than normal (minutes)	0.258	-0.013	0.053**
	(1.272)	(0.015)	(0.021)
Subject-round observations	2,981	2,981	2,981

Notes: We run the same regressions as reported in Table 10 except that we use only subject-round observations from situations where the subject lost in the previous round and therefore earned $0 (see Table 4 for the list of situations). Heteroskedasticity-consistent standard errors with clustering at the group level are shown in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% levels (two-sided tests).

Table OA.23: Effect of thinking for longer than normal on strategic behavior: including only situations where the subject lost in the previous round