Cognitive Ability and Strategic Sophistication*

Jeffrey Carpenter† Michael Graham‡ Jesse Wolf§

October 31, 2012

Abstract

In three experiments we examine the extent to which strategic sophistication (i.e., inductive reasoning, iterative dominance and level-k thinking) is determined by broader cognitive skills. In the first experiment we replicate previous results showing strong associations between cognitive ability and sophistication in a game of iterative dominance. We then extend the results to a game requiring induction to win. In the second two experiments we extend the literature in new directions. In Experiment 2 we modify the games to better capture participants' ability to reason inductively and predict the sophistication of others and, again, find strong associations between cognitive ability, measured using a common IQ test, and sophistication. In Experiment 3 we examine more closely the causal nature of the relationship between cognitive ability and sophistication. We use a standard tool from cognitive psychology to randomly lower the cognitive ability of participants and show that this significantly affects game performance.

JEL Codes: C72, C92, D03, D83

Keywords: Cognitive ability, induction, dominance, level-k, bounded rationality.

*We thank the editors, the referees, Casey Rothschild and Aldo Rustichini for valuable comments. David Cha, Andrew DeLoach, Jamie Herther and Brian Swartz contributed excellent research assistance. Middlebury College provided financial support.

†Corresponding author. Department of Economics, Middlebury College.

‡Department of Economics, Middlebury College.

§Department of Economics, Middlebury College.

1 Introduction

The ability to reason deeply is key to choosing wisely in economic situations and yet little is known about the traits that determine strategic sophistication. We examine whether economic reasoning is a reflection of broader cognitive abilities, the kind measured in college board exams, IQ tests or other related instruments. To better identify the causal relationship we also examine whether temporarily diminishing one's cognitive ability has an effect on sophistication.

To be more precise, we examine three aspects of strategic sophistication. Considering the first two, in a wide variety of economic settings agents are asked to reason iteratively and think inductively. They must reason iteratively in situations in which one risky choice stochastically dominates another, for example, and they must use induction to make a number of critical financial choices like how much to save for retirement or how much to invest in education. Of course things become messier when other agents are involved. Here, to properly assess their roles in even common games agents must iterate to determine, for instance, how best to compete in markets and they have to induct to do well in settings like contract negotiations or other situations that require bargaining. While interacting in some games, successful agents also need a third skill - the ability to assess the strategic sophistication of others. Seeing deeply into a situation is clearly important but to do well one often needs to properly assess how deeply others see too. Proving how clever one is can actually be costly when what really matters is how clever everyone else is.

A number of recent studies have found strong associations between broader cognitive skills and related aspects of economic choice. For example, Frederick (2005), Benjamin et al., (2006), Burks et al., (2009), and Dohmen et al., (2010) all find significant links between cognitive ability measured variously as scores from college board, IQ, or cognitive reflection tests and preferences for risk and time. Here, the emerging concensus appears to be that people with higher cognitive ability tend to be more patient and closer to risk neutral.

In terms of the experimental games that we employ to measure strategic sophistication, our results dovetail nicely with a number of previous studies. Gneezy et al., (2010), Dufwenberg et al., (2010), and Levitt et al., (2009) all employ versions of the game that we use to examine inductive reasoning and document a lot of variation in the ability of participants to play optimally. Both Gneezy et al. and Levitt et al. find that players are very likely to make errors at the beginning of the game: Only 41% of players make the right move and go on to play optimally in the rest of the game in Gneezy et al. and the fraction is only slightly higher in Levitt et al. In addition, Dufwenberg et al. demonstrate, in accordance with cognitive ability being important, that the rate at which players solve the game depends on how complicated the game is.

By now there are lots of studies of the second game that we use, the p-Beauty contest, many of which are surveyed in Camerer (2003). However, there are still few papers that unpack the determinants of how well participants in this game make use of iterative dominance. Burnham et al., (2009) use an IQ test and find that participants with higher measured IQ pick lower numbers in the p-Beauty Contest, a result that hints at sophistication. Gill and Prowse (2012) find similar results in their implementation of the beauty contest and also extend the earlier results to show that high ability types appear better at assessing and learning the strategic landscape in which they interact. Rydval et al., (2009) and Georganas et al., (2010) use beauty contest games that are dominance solvable to identify level-k aptitude and test the extent to which different measures of cognitive ability predict sophistication. While Rydval et al., find that working memory predicts reasoning, Georganas et al, show that neither Frederick's (2005) cognitive reflection test (CRT) nor an IQ test reliably predicts reasoning up to level 3.

We attempt to replicate many of the key findings of the related literature but also extend these earlier results in a number of important directions. Our contributions come from two distinct experiments and a total of three treatments. The first experiment was designed to gather and compare richer sets of both cognitive ability measures and strategic behavior in a relatively large sample. In the first treatment of the second experiment we modify the games and employ a more standard measure of cognitive ability to better unpack the relationships documented in the previous literature. In the second treatment of experiment two, we continue to use the games and measures developed for the first treatment but borrow a technique from cognitive psychology that, in comparing the two treatments, allows us to better identify the causal nature of the relationship between cognitive ability and sophistication. Because the contributions of the three treatments are separable to a great extent, we organize the paper along these three themes (replication in a large sample, modification and extension, and identifying causation) and, for simplicity, refer to the three treatments as different experiments.

In Experiment 1, we use an online survey to gather a large sample of responses. We use "Hit 15" to measure the ability of participants to think inductively. In our version of this zero-sum game of perfect information (originally discussed in Gneezy et al., 2010 and Dufwenberg et al., 2010), two players are each given three "chips" per round and told that when it is their turn they can place any number of these chips into a fictitious basket. They are then told that the winner of the game is the person who places the 15^{th} chip in the basket. To win the game, players need to think inductively. Clearly, whoever places the 11^{th} chip in the basket can win because the best that the next person can do is position 12, 13 or

14. Induction suggests that the same logic applies to being the first one to get to position 7 (in which case the next person can only achieve positions 8, 9, or 10) and position 3. While induction is necessary to solve Hit 15, as Gneezy et al., (2010) point out, because the game is zero-sum social preferences play no role and given the game is sequential with a dominant strategy one also need not worry about the strategic sophistication of the other player. In other words, the game is a relatively clean measure of inductive ability.

In our Experiment 1 implementation of the p-Beauty Contest (originating in Nagel, 1995), used to measure the ability of our participants to reason iteratively, participants are asked to pick any number between 0 and 100 and then are told that the winner will be the person who picks the number that is closest to $\frac{1}{2}$ the average of all the guesses (i.e., $p = \frac{1}{2}$). Starting from the naïve assumption that everyone can see deeply into the structure of this game, winners should realize that all the numbers above 50 are clearly dominated by picking 50 because even if everyone else chooses 100, half the average would be only 50. If one understands this, then a second iteration makes it clear that 25 dominates everything between 25 and 50, a third iteration gets one to 12.5, and n iterations ends at 0. Given this logic, everyone should choose 0 and the winner should then be determined at random. In terms of the ability to think iteratively, getting to 50 is thought of as 0 steps of iterative dominance, 25 is 1 step, 12.5 is 2 steps, 6.25 is 3 steps, and getting all the way to 0 is n steps.

To expand on the previous literature in Experiment 1, we use two measures of cognitive ability: Fredrick's (2005) CRT and a more standard measure, college entrance exam scores from both the Scholastic Achievement Test (SAT) and the American College Test (ACT). Contrary to standard tests like those determining college entrance decisions which focus on learned rules and concentration, we use the CRT to capture one's ability to recognize correct answers and patterns instantly and effortlessly.

The results of Experiment 1 show that both measures of cognitive ability are strong predictors of strategic sophistication. Participants with higher measured cognitive ability (i) formulate better strategies in the Hit 15 game and are thus better at inductive reasoning; (ii) choose guesses in the p-Beauty Contest game that are significantly closer to the winning guess and are therefore better are iterative dominance and level-k thinking and (iii) these relationships are strong regardless of the measure of cognitive ability used.

In Experiment 2 we attempt to amplify the previous results by digging deeper into the two games. Using a lab experiment, we expand the Hit 15 game used in the survey and have participants also play "Hit 5" and "Hit 10". Further, we modify the standard p-Beauty Contest in Experiment 2 by eliciting the expectations of our participants. The problem with inferring sophistication in the standard beauty contest is that without observing player expectations of what the distribution of guesses by the $n-1$ other participants will be, it is hard to know whether the player has reasonably accurate expectations of what others will guess and whether the player's guess is a best response to those beliefs or not. Using our new method to capture those missing beliefs we can examine the amount of sophistication players attribute to each other and test the extent to which players best respond.

In Experiment 2 we also explore using the Raven's standard progressive matrices (Raven et al., 2003) to measure cognitive ability. Raven's matrices are now a standard non-verbal assessment of reasoning or general intelligence. In each of sixty puzzles, participants are asked to pick one item to complete patterns that become progressively more complicated.

In this experiment we first replicate our survey results by showing that high ability participants do significantly better (i.e., are more sophisticated) in both games: they win more Hit games and they come closer to the winning guess in the p-Beauty Contest. We then extend the literature by showing that the beliefs of high ability players attribute more sophistication to other players than do the beliefs of low ability players, that higher ability participant are better at estimating the guesses of the other participants and, most importantly, that high ability players come much closer to best responding to their stated beliefs than do low ability players. In other words, high ability players understand the interative structure of the game much better.

Experiment 3 was designed to push harder on identifying the causal relationship between cognitive ability and strategic sophistication. The persistent problem with much of the related literature is that it remains unclear whether cognitive ability causes sophistication or is just associated with it. At the same time cognitive psychologists have developed various methods to temporarily (and randomly) "shock" one's cognitive ability. The most common of these methods, and the one that we employ, is to place a "load" on a player's working memory. In our third experiment we ask participants to play all the same games that were played in Experiment 2, but before they start, we ask them to memorize a random 7-digit number and tell them that we will pay them $5 if they can recall the number at the end of the experiment. In other words, we randomly reduce the cognitive ability of half our lab participants and test whether this exogenous shock affects sophistication. Indeed, it does: players whose cognitive abilities are reduced win fewer Hit games, expect other participants in the p-Beauty Contest to be less sophisticated and do worse at best responding to their beliefs. In addition, because we gather Raven's scores for these participants before imposing the cognitive load, we can also assess any differential impact of the load on high and low ability participants. Here we find that the cognitive load falls heaviest on high ability types in the Hit games (i.e., the reduction in performance for high ability types is larger than it is for low ability types), it has more of an affect on the beliefs of low ability types in the Beauty Contest and it impairs the ability to best respond of the two types roughly equally in the beauty contest.

We proceed by describing each experiment, in turn. For each trial we explain both how (and with whom) the experiments were conducted and how the cognitive ability data were gathered. We then summarize the data and explore the extent to which our measures of strategic sophistication can be predicted by cognitive ability.

2 Experiment 1 - Replication in a large sample

To gather a large sample, the protocol was implemented on the campus of Middlebury College as an incentivized web survey. Participants were promised \$10 to fill out the survey which took less than 15 minutes to complete and told they could earn an additional \$25 for winning the p-Beauty Contest. At the beginning of the survey participants were asked to enter their campus mailbox numbers so that final payments could be sent anonymously to mailboxes. Solicitation emails were sent to a list of 750 student addresses. The list was a random sample provided by the College Registrar. In total, there were 422 responses.

2.1 Methods

The survey began by asking the participants for a number of demographics. They were asked for their gender, ethnicity, year in college, and household income. After the Hit 15 game was explained (see the appendix for details), it was carried out in strategic form as the following two survey questions: (i) *If you go first, how many points will you place in the basket?* (ii) *If you go second and the other player has already put 2 points in the basket on her first turn, how many would you put in?* For our purposes the number of correct responses is used as a measure of one's ability to think inductively. The p-Beauty Contest came at the very end of the survey. Here the rules were explained, including that the total pool of responses would be split into groups of 20 at random so that there would be more than one winner (of the \$25 prize) and the participants knew exactly how many other guessers they would be competing against. Participants then responded by inputting a number between 0 and 100. Guesses in this original version of the p-Beauty Contest measure a combination of iterative reasoning and level-k aptitude.

As part of the demographics, the participants were first asked which college entrance exam they took and then they were asked to self-report their scores. To minimize errors due to not remembering exactly what one's score was, for each exam the possible range of scores was broken into seven equal-sized bins. Consistent with the selectivity of the College, only 8 of the 337 people who reported a SAT score disclosed having a score below the fourth bin (all 8 being in the fifth bin) and none of the 85 people who reported an ACT score indicated a bin below the fourth. To create a common measure of cognitive ability from these two types of exam score responses, we created four indicators. If your score was in the highest bin for either test you were put in the "Highest" category, if your score was in the second highest bin for either exam, you were put in the "Second" category, and so on with the exception that the 8 people with reported SAT scores below the fourth bin were also put in the "Lowest" ability category together with all the people from the fourth bins.

Frederick's (2005) cognitive reflection test is short; it is composed of the following three questions: (i) *A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?* (ii) *If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?* (iii) *In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half the lake?* We recorded the total number of correct responses (i.e., 5 cents, 5 minutes, and 47 days) as our second measure of cognitive ability.

2.2 Participants and their behavior

Our participants in Experiment 1 were representative of the student population from which they were sampled. Slightly more than half (i.e., 54%) of our respondents were women, 82% were born in the United States, and although 14% classified themselves as Asian, only 2% were African-American, and only 3% had Hispanic heritage. The respondents came from relatively wealthy families: 35% reported coming from families that earn more than $150k per year and only 11% reported coming from families that earn less than $50k per year. Lastly, the distribution of years of college experience is approximately uniform: 25% reported being seniors, slightly fewer were juniors (16%) which is explained by the fact that many juniors spend time abroad, 30% were sophomores and the remaining 29% were freshmen.

The data on cognitive ability are summarized in Table 1. In the bottom row of the table one can see that that most of the participants reported an exam score in the second highest category, although there is some variation. One-quarter reported being in the highest bin and approximately one-tenth reported having a score in the lowest bin. There is also considerable variation in the CRT scores as one can see in the last column of Table 1. One-third answered all three questions correctly, about one-quarter got one or two right, and more than one-tenth did not get any of them right. Compared to Frederick (2005), we see that the current sample did quite well on the CRT. In fact, the mean score of 1.79 correct answers for the Middlebury sample fits neatly between the Princeton mean of 1.63 and the MIT mean of 2.18. At first blush, one might be tempted to explain the higher Middlebury mean by the fact that there may have been less time pressure during a web survey; however, Frederick

(2005) reports the mean of a different web-based study also being considerably lower (i.e., 1.10).

The last thing to take away from Table 1 is that the two measures of cognitive ability correlate to some extent, especially at the upper ends of the distributions. This is also consistent with Frederick (2005) who reports similar, if not slightly stronger, correlations between CRT scores and SAT/ACT scores. Although college board exams are perhaps better suited to capture what Stanovich and West (2000) call "System 2" processes which require sustained concentration and the CRT is designed to capture the more spontaneous "System 1" processes, the two measures are still highly significantly correlated in the Middlebury sample (Kendall's $\tau = 0.23$, $p < 0.01$).

In panel (a) of Figure 1 one can see the distribution of correct answers to the Hit 15 game. Almost half of the participants answered one of the two questions correct and the remaining half seem almost equally distributed between getting neither question right and getting them both right. It also seems that the first question was a bit easier than the second - 59% of the participants got the first one right but only 37% got the second one. In panel (b) we present the distribution of guesses in the p-Beauty Contest. As is typical (e.g., Camerer 2003), there is a lot of variation in responses, including a number of clearly dominated choices larger than 50. As is also typical, there are prominent peaks at 0, 1, 2, and 3 steps of iterative dominance and the overall mode is to complete two steps of iterative dominance. Like the measures of cognitive ability, behavior in these two strategic situations is correlated. Participants who are better at backward induction in the Hit 15 game tend to choose lower numbers in the p-Beauty Contest ($\rho = -0.15$, $p < 0.01$).

Although not perfect, like other related experiments we first use one's p-Beauty guess as a measure of iterative reasoning; the lower one's guess the more iteratively sophisticated one is. A slightly better measure that we also examine is the extent to which one clearly iterates 0, 1, 2,..., n steps. Guesses are not the perfect measure, however, because the choice of 0 almost never wins. To win the p-Beauty Contest, one needs to also be strong in the third aspect of sophistication mentioned above - the ability to properly estimate the sophistication of others (a la the level-k reasoning of Costa-Gomes et al., 2001). In the p-Beauty Contest one needs to best respond to (i.e., think one iteration deeper than) the expected distribution of guesses. In this sense, the deviation of one's choice from the winning choice is a measure of both iterative reasoning and the ability to judge the sophistication of others.

2.3 Does cognitive ability predict sophistication in a large survey?

As the previous section suggests, the two measures of cognitive ability are correlated, as are choices in the two games. The question that remains is whether the measures of cognitive ability are correlated with strategic choices. Figure 2 summarizes these correlations. In panel (a) we present the relationship between strategic choices and college board exam scores. The correlations appear strong, especially in the tails of the distributions. In the Hit 15 game, those participants who scored highest on the college entrance exam responded correctly an average of 53% of the time while those who did worst on the exam responded correctly only 37% of the time. Although there is not a large difference for the two intermediate levels of the exam, both are clearly between the extremes and the relationship is monotonic. Monotonicity also appears to hold for the relationship between exam scores and the deviation from the winning number in the p-Beauty Contest. While those participants who reported scoring relatively poorly on the college entrance exam guess a number that is on average 24.97 away from the winning number (which tended to be about 10), those who did the best on the exam guessed numbers that only differed by an average of 8.77.

To make the analysis more rigorous, In Table 2 we regress the number of correct responses in the Hit 15 game and the number chosen in the p-Beauty Contest on the test score indicators and other controls including gender, ethnicity, year in college and household income. Because the responses are bound in both games we use the Tobit regressor. In the first column of Table 2 one can see that the relationship between exam scores and correct responses in the Hit 15 game is indeed monotonic. Those who did best on the college entrance exam tend to make another 0.717 ($p < 0.05$) correct responses than those who did the worst, those who did second best respond correctly 0.464 ($p < 0.10$) more times and those in the third category respond correctly 0.420 more.[1]

Similar, if not stronger, results for the p-Beauty Contest appear in the second column of Table 2. High scorers on the exam tend to pick numbers that are 18.41 lower ($p < 0.01$) than those who report being among the low exam performers. In addition, those who report being among the second highest exam scorers pick numbers that are 12.24 lower ($p < 0.01$), and even those in the third highest category pick numbers that are 9.83 lower ($p < 0.01$), on average.[2] Lastly, the estimates also suggest that women in our sample tended to pick higher numbers than men ($p < 0.01$).

In panel (b) of Figure 2 one can see that using the CRT instead of entrance exam scores results in similar correlations between cognitive ability and strategic sophistication. However, looking more closely, it seems that with the CRT we get slightly more separation between the intermediate ability categories. In the Hit 15 game, those who get all three questions

[1] Comparing these point estimates to each other reveals more evidence of monotonicity; however, the levels of significance are more modest (i.e., in the neighborhood of $p = 0.15$).

[2] In this case the evidence for monotonicity is particularly strong - all point estimate comparisons are highly significant (i.e., $p \leq 0.01$) except for the comparison of the estimates for the second and third highest scorers.

right respond correctly 58% of the time while those who do not get any of the CRT questions right only respond correctly in Hit 15 37% of the time. There is also a 7 percentage point difference between those who get 1 CRT question right and those who get 2 right.

In addition, panel (b) of Figure 2 suggests that doing better on the CRT correlates with being closer to the winning guess in the p-Beauty Contest. Those participants who do not answer any of the CRT questions correctly deviate from the winning guess by an average of 22.40, those who get one right deviate by 18.69, those who get two right deviate by 12.89, and those who get all the questions in the CRT right deviate by only 7.55, on average.

Table 3 substitutes CRT scores for exam scores but largely replicates what we saw in Table 2.[3] Those who have a CRT score of three make 0.855 ($p < 0.01$) more correct responses in the Hit 15 game than those who score zero, the size of the effect is close to half ($\beta = 0.410$, $p < 0.10$) for those scoring only two correct, and it is more than halved again ($\beta = 0.150$, $p > 0.10$) for those scoring only one on the CRT. The second column of Table 3 shows that the strong association between guesses in the p-Beauty Contest and cognitive ability also remains when we use the CRT. Compared to those who get none of the CRT questions right, the guesses of those who get them all right falls by almost 20 ($\beta = -17.198$, $p < 0.01$), guesses fall by 9.454 ($p < 0.01$) for those who score two and by 3.798 ($p > 0.10$) for those who get one right on the CRT.[4]

So far we have seen that there appear to be strong correlations between cognitive ability and strategic sophistication; however, as mentioned above, we might be able to better differentiate between the ability to think iteratively and one's knowledge of the strategic sophistication of the other players in the p-Beauty Contest by examining the extent to which participants make discrete steps along the perfectly rational equilibrium path and by looking at deviations from the winning guess.

Tables 4 and 5 offer some evidence that cognitive ability predicts iterative reasoning. In both tables we create indicators for making at least 1, 2, 3 or n iterations and the only difference between the tables is the measure of cognitive ability used. We coded players as having made at least n iterations if they guessed a number less than 1, those who were coded as having made at least three iterations included the players who guessed numbers between

[3] As a robustness check, we also considered an alternative specification of the dependent variable for the Hit 15 game, one that better accounts for the fact that people might have gotten one or two of the questions correct by chance. We recoded the responses: 0 for those who got both the questions wrong, 1 for those who just got the first (easier) one correct, 2 for those who only got the second correct and 3 for those who got both questions right and might, therefore, have a much deeper understanding of the game. Further, to push a bit harder on the cardinality assumption made in our original regressions, we used ordered probit. The results were identical, however, to the original, OLS, results.

[4] Even more so than with the estimates in Table 2, in Table 3, the point estimates tend all to be significantly different from each other.

6 and 7 and the 10% who made at least n iterations.[5] At a minimum, we would like to know if cognitive ability at least rules out all the obviously dominated numbers between 50 and 100. In the first columns of ables 4 and 5 we find that doing better on the exams does predict being progressively more likely to guess 50 or less. Further, although there are a few significant coefficients in Tables 4 and 5, the overall pattern of the coefficients is what is interesting. Using either measure of cognitive ability, we see that those in the highest ability category are significantly more likely to complete at least one step of iterative dominance (and they are more likely to complete any number of steps if just the CRT results are considered). Many of the participants in the second highest ability category also understand this. However, increasingly fewer comprehend this logic in the bottom two ability categories. As one can see, the broader pattern is that those of higher ability tend to be more likely to make each step of iterative reasoning. That said, while the pattern is strong, the differences tend to be modest. In other words, there is some evidence of cognitive ability influencing iterative reasoning, but it is not conclusive. At the same time, the results in Table 6 are more substantial. Here we switch from examining participant guesses to deviations from the winning guess to get a better sense of the participants' ability to correctly assess the sophistication of the other players. As one can see, the results are very similar to those in Tables 2 and 3 which suggests that cognitive ability might be more strongly associated with level-k reasoning.

3 Experiment 2 - Modificaton and extension

For the second experiment, we return to the lab to replicate our results and extend them in ways that would be hard, if not impossible, to do in a survey. First, we add two additional "Hit" games (Hit 10 and Hit 5) to broaden our measure of inductive reasoning and we incentivize all three games. Second, we add a stage to the p-Beauty Contest in which participants are asked for their expectations of what others will guess. Not only does this last feature allow us to examine these expectations, we can determine the extent to which participants best respond to this distribution of expected guesses. Third, instead of relying on self-reported college entrance exams and the short CRT to measure cognitive ability, in Experiment 2 we employ a common instrument from the related literature, the standard progressive matrices of Raven et al. (2003). A total of 64 Middlebury College students participated in this trial and earned an average of $15.94 in 8 sessions that lasted less than an hour.

[5] The at least two iterations group included the n players, the 1 players and the players who guessed numbers between 12 and 13. Finally, the at least one iteration group included the ns, the 3s, the 2s and the players who guesed numbers between 24 and 25.

3.1 Methods

In the second experiment, participants were asked to complete three tasks. The first task was to correctly solve as many of Raven's standard progressive matrices as possible in 12 minutes. There were a total of 60 puzzles and in each one, participants were asked to pick the best of six or eight options, the one that best completed the pattern currently appearing on the computer screen.

The second task was to play three versions of the Hit game against the computer.[6] The participant always made the first move and the computer was programed to best respond. As a result, and unlike the survey version from the first experiment, participants in these versions of the Hit game had to play the full game and play it perfectly from the start to win. The first game was easy, Hit 5. Here the participant only had to figure out that if she placed just one chip in the basket to start, the computer could not get past four. The second game, Hit 10, required two steps of backward induction to realize that to guarantee a victory by placing the 6th chip in the basket, one needed to put two in on the first round. The final Hit 15 game required three steps of induction to realize that one had to place the 3rd, 7th and 11th chips in the basket. Participants were paid $5 for each game won.

To make it easy for participants to think about and register their expectations in the p-Beauty Contest, the game was altered in two ways for Experiment 2 and played using paper and pen. First, participants were asked to create a simple histogram of their expectations. The participants were reminded that there would be 19 other people per contest and then they were asked to make 19 marks on a pre-printed graph, one for each of the other guessers. When they were finished, they had a simple visual representation of their expectations. Second, the strategy space was limited to guesses between 0 and 20 so that it was not too difficult and not too time-consuming to record one's expectations and think about the average of these expected guesses. As before, the prize was $25 for each 20-person group.[7] In addition, participants could also win if they were best at predicting what the distribution of the other participants' guesses would be. The person who most accurately predicted the other guesses also won $25.[8]

3.2 Participants and their behavior

Forty-one percent of the participants in Experiment 2 were women. Considering ethnicity, 3% listed African-American heritage, 28% classified themselves as Asian or Asian-American, 2% were Latino or Hispanic, 9% claimed other or mixed heritage and the residual 57%

[6] Both the Raven task and the Hit games were programmed in Z-tree (Fischbacher, 2007).
[7] The groupings were randomly determined after all the sessions were run just like in Experiment 1.
[8] Instructions for both games appear in the appendix.

categorized themselves as White/Caucasian. In this experiment more freshman (40%) and seniors (33%) participated than did sophomores (14%) and juniors (13%).

Overall, participants in Experiment 2 attempted 51 Raven's matrices and got 46 right, on average. The resulting cognitive ability scores (i.e., the number answered correctly) ranging from 30 to 57 are illustrated in Figure 3. Despite having only 12 minutes to work on the matrices, the results of our implementation are very similar to those gathered in another recent study of college students (Branas-Garza and Rustichini, 2011).

The behavior of the participants in Experiment 2 is summarized in Figure 4. In panel (a) one can see the distribution of Hit games won. Relatively few people (17%) won none of the games and the average number of games won was 1.56. Also of interest is the fact that, as intended, Hit 5 is easier than the other two games but Hit 10 seems about as hard as Hit 15. We find that while 83% of players win Hit 5, only 38% win Hit 10 and 36% win Hit 15.

As for the modified p-Beauty Contest, as one can see in panel (b), compared to Experiment 1 there are many fewer observations of clearly dominated choices - only one person guesses a number higher than 10. There is still plenty of variation, however, and many people (15%) pick numbers close to one step of iterative dominance (i.e., choices of between 5 and 6) and even more pick numbers close to two or three steps. That said, sophistication can only be properly assessed once we control for player expectations. Participants in Experiment 2 expected the $n-1$ other participants to do approximately two levels of iterative dominance; they expect them to guess 5.14, on average.

3.3 Does cognitive ability predict sophistication in the lab?

We build our analysis of the lab data slowly by first dividing our participants into two ability groups. Those with Raven's scores below 46, the median, are classified as low ability and those with scores of 46 or more are classified as high ability. In Figure 5 we examine the extent to which strategic sophistication varies with Raven's scores. In panel (a) we see that like the survey data, cognitive ability predicts success in the Hit games. Those participants of high ability solve an average of 0.85 games more than those of low ability and the difference is highly significant ($t = 3.63$, $p < 0.01$). Interestingly, we also find that the "ability gap" does not seem to widen as the games get harder. Starting with Hit 5, 97% of high ability participants and 70% of low ability participants solve the game correctly resulting in a gap of 27 percentage points. The gap grows only slightly to 28 percentage points in Hit 10 (comparing 52% and 24%) and just a bit more to 30 percentage points in Hit 15 (51% versus 21%).

Although less significant ($t = 1.90$, $p = 0.06$), it appears from Figure 5(b) that high ability participants also attribute more strategic sophistication to the rest of the participants in the beauty contest game than low ability participants. The average expected guess of the low group is 5.81 and that mean falls to 4.42 for the high ability group. In terms of accuracy, the high ability group is better at estimating what others will guess. The realized average guess in Experiment 2 was 3.64 which is significantly lower than the average guessed by the low ability group ($t = 3.99$, $p < 0.01$) but not significanly different from the average guess of the high ability group ($t = 1.62$, $p = 0.11$). Most importantly in the context of the beauty contest, however, is that high ability participants are much better at best responding to their stated expectations. In panel (c) of Figure 5 one can see that low ability players tend to deviate from their best response by close to two units while the high ability group is much better at choosing a guess that is in concert with their beliefs - their deviation is just 0.59 units, on average. This difference is also highly significant ($t = 3.86$, $p < 0.01$). Based on this "first blush" analysis, it seems clear that not only have we replicated the survey results in an enviroment in which all the games are incentivized, we have also strengthened our results on the beauty contest by showing that participants with high cognitive ability don't simply choose lower guesses, they are better at guessing what others will do and responding to this estimation.

To press harder on our results, in Table 7 we regress game performance on the Raven's score along with controls for gender, ethnicity and year in college. In column (1) we find that each additional Raven's puzzle that one solves is associated with a tenth of a won game increase in Hit performance ($p < 0.01$) and that women tend to win 0.776 fewer hit games ($p < 0.01$).[9] In the second column we turn attention to the beauty contest. Here we see that one more solved puzzle is associated with a reduction in one's expectation of what the other participants will guess, on average, of 0.138 units ($p < 0.05$) and that women tend to have higher expectations ($p < 0.05$). That is, high ability participants expect the other participants to submit lower guesses. The dependent variable in column (3) is a measure of the accuracy of one's beliefs, in this case the entire distribution of beliefs. Specifically, for each participant we calculated the Kolmogorov-Smirnov statistic comparing the cumulative distribution of their expectations to the cdf of all the actual guesses (excluding their own). For each participant this statistic measures the largest difference at any step in the resulting two cdfs. As one can see in column (3), participants who solve more Raven's puzzles have smaller differences between their beliefs and the actual distribution of guesses ($p < 0.01$) - they are better at anticipating what others will guess. In column (4) we find that higher cognitive ability is also significantly associated with better responses to one's beliefs. Solving

[9] Circling back to the issue of game difficulty, when we analyze the probability of winning each game separately, consistent with the constant ability gap discussed above, we find a stable effect of cognitive ability. For each added Raven's puzzle solved, participants are 3% more likely to solve Hit 5, 4% more likely to solve Hit 10 and 5% more likely to solve Hit 15.

another matrix puzzle is associated with reducing the absolute distance between one's guess and one's best response by 0.123 units ($p < 0.01$). We also see that there is no gender effect in best anticipating guesses or in best responding - women are just as good as men at formulating and translating estimates into appropriate guesses. Finally, to connect our results to those in the survey experiment, we find, again, that higher cognitive ability is associated with making guesses that are closer to the winning guess. The absolute deviation between one's guess and the winning guess falls by 0.181 units for each Raven's puzzle solved ($p < 0.01$).

To sum, in the context of Experiment 2 in which we have incentivized performance in each of the games and allowed players to express their expectations in the beauty contest, we both replicate and expand on the results of Experiment 1. We still find strong associations between cognitive ability, this time measured by the Raven's progressive matricies, and performance in the games; however, we also find that part of the reason why high ability types do better in the beauty contest is because they are both better at predicting what others will do and they are better at best responding to those beliefs.

4 Experiment 3 - Identifying causation

Our third experiment focuses on identifying the causal nature of the relationship between cognitive ability and strategic sophistication. To corroborate the associations that we (and others) have repeatedly found, we sought a method that would allow us to randomly shock the cognitive functioning of participants to see if these exogenous interruptions would reverberate through and affect game play. In this experiment 54 Middlebury College students participated in one of six sessions and earned and average of $17.31.

4.1 Methods

The protocol for Experiment 3 was identical to that of Experiment 2 except for one change - the participants in Experiment 3 played games under a cognitive load. Our target for this load was the deliberate processing capability of our participants (Berkowitz, 1993). In terms of the type of load (Kalyuga, 2011), we chose to implement an "extraneous load," which emerges from stimuli unrelated to the assigned task and detracts from learning and performance. Holding between two and seven digits in one's working memory has been previously shown to affect controlled cognitive processes (Conway et al., 2003). Further, working memory appears to be essential to performing reasoning and problem solving and there also seems to be a strong empirical link between broader cognitive ability and working memory (Conway et al., 2003; Light et al., 2010). With this in mind, we asked participants to memorize a random 7-digit number after they had done the Raven's matrices but before they played the games. In addition, they were paid an additional \$5 at the end of the experiment if they could correctly recall the number (83% did so correctly).

4.2 Participants and their behavior

The participants in Experiment 3 were distributed as follows. They were 43% female, 2% African-American, 15% Asian or Asian-American, 6% Latino or Hispanic, 72% White or Caucasian and the remaining 5% listed mixed heritage. The distribution of years in college was 17% freshman, 22% sophomores, 35% juniors and 26% seniors. In terms of cognitive ability, the average Raven's score for these participants was 45.96 which is not significantly different from the "un-loaded" participants in Experiment 2 ($t = 0.19$, $p = 0.85$).[10]

Approximately 65% of the participants in Experiment 3 won Hit 5, 22% won Hit 10, 20% won Hit 15 and the overall average number of Hit games won was 1.07. In the p-Beauty Contest these participants submitted average guesses of 3.95. These guesses tended to be 1.92 units away, on average, from the best responses to their stated expectations. The expectations of the guesses that would be submitted by the other participants averaged 6.03.

4.3 Identifying the causal effect of cognitive ability

Even the simple summary statistics illustrated in Figure 6 indicate a strong causal relationship between cognitive ability and strategic sophistications. In panel (a) we see that the imposition of the load accounts for approximately half a game reduction in Hit performance ($t = 2.83$, $p < 0.01$).[11] Considering the beauty contest, panel (b) indicates a smaller effect of the load on one's expectation of how strategically sophisticated other participants will be. Those bearing the load expect others to guess numbers that are approximately 1 unit higher than those without a load ($t = 1.75$, $p = 0.08$). As for their responses to these expectations, in panel (c) we see that "loaded" participants are significantly worse at best responding ($t = 2.85$, $p < 0.01$). The most common mistake for the loaded participants was to guess a number closer to the average of their expected guesses instead of half the average.[12] In other words, the load appears to have diminished their ability to see as deeply into the game.

[10] Considering the other obserables, the balance between Experiment 2 and 3 suggests randomization to treatment was achieved. At the 5% level the only significant differences were that fewer freshman and more sophomores participated in Experiment 3.

[11] Consistent with our findings in Section 3.3, we see that the "load gap" in performance is also relatively constant across Hit games. Unloaded players are 18% more likley to solve Hit 5, 15% more likley to solve Hit 10 and 15% more likely to solve Hit 15.

[12] Creating an indicator for being closer to the average expected guess than the best response and running a probit regression (with all the controls) indicates that loaded participants were 23% more likely to be closer to the average ($p = 0.02$).

In Table 8 we take a closer look at the effect of the cognitive load on game performance. Consistent with random assignment to treatment, the effect measured in Figure 6(a) is very similar to what we find in the first column of Table 8: loaded participants solve 0.514 fewer games, on average, than unloaded players ($p < 0.01$). From the second column we see that the point estimate on how much less sophistication loaded players attribute to their peers is a bit stronger than Figure 6(b) suggests; the difference is 1.129 ($p = 0.03$). Interestingly, in column (3) we see that attributing less sophistication to their peers might have the unintended, but helpful, consequence of accounting for the fact that loaded participants play with less sophistication. In other words, if the load reduces one's ability to see deeply into the game it could easily cause one to both do fewer steps of dominance one's self and attribute fewer to one's peers. In this case, the effects seem to balance out and be reflected in no discernible difference in the accuracy of loaded player beliefs. That said, in column (4) we confirm that loaded players are worse at best responding to their expectations; on average they tend to make guesses that are 0.919 units further away from their estimated best response ($p < 0.01$). The effect of the lack of sophistication about best responding on one's deviation from the winning guess is attenuated to some extent by the fact that loaded players' beliefs aren't, perhaps inadvertently, that inaccurate. In column (5) we see that loaded players do deviate more from the winning guess but the point estimate is not precise. Overall, however, our results from Experiment 3 confirm that the link between cognitive ability and strategic sophistication is causal.

4.4 Who bears the (cognitive) load?

Because we collected Raven's scores for the participants of Experiment 3 before the load was imposed, we can ask whether the load had a larger effect on low or high ability players. That is, we can ask if the load was "heavier" for one type of player than another. Figure 7 presents what are, in effect, crude difference-in-difference estimates. Considering Hit games won in panel (a), it appears that imposing the cognitive load has a larger deleterious effect on high ability players than it has on low ability players. The reduction in performance for the high ability players averages 0.74 games while it is only a loss of 0.26 games for the low ability players. Panel (b) suggests that in the beauty contest the load might fall disproportionately on the low ability players, at least when it comes to predicting what others will guess. Here the low ability loaded players predict average guesses that are 1.55 units higher than their unloaded counterparts compared to a difference that is only 0.282 among the high ability players. Lastly, the load seems to hit both low and high ability players equally when it comes to best responding: among both low and high ability participants, the reduction in the ability to best respond averages about 0.70 units.

To sharpen our difference-in-difference estimates, in Table 9 we regress game outcomes on Raven's score, an indicator for the cognitive load and the interaction of the two (along with the other controls). Because the Raven's score is now treated as a continuous variable, the marginal effects are harder to interpret and so, like in Figure 7, we also report the estimated marginal effects of the load for the high and low ability groups (i.e., $\partial Sophistication/\partial Load = \beta_{Load} + \beta_{Raven's \times Load}(Raven's_{mean})$). The mean Raven's score for the high ability group is 51 and for the low ability group the average is 42.

Comparing the first two columns of Table 9 it is clear that the load falls disproportionately on the high ability types in the Hit games. Not only is the interaction term negative and significant at the 5% level in column (2), the estimated marginal effect of the load is -0.24 ($p = 0.18$) for the low ability group and -0.73 ($p < 0.01$) for the high ability group. Further, the comparison of these point estimates (i.e., the diff-in-diff) is also significant ($p = 0.02$). In columns (3) and (4) we test whether the load disproportionately affected the capacity for low ability players to estimate what others would guess. While the interaction term in column (4) is negative, it is only significant at the 11% level. Calculating the estimates at the mean levels of ability for the two groups indicates a similar difference. For the low ability participants, the estimated effect of the load is an increase in the average expected guess of 1.45 units ($p < 0.01$), while for the high ability participants the increase is only estimated to be 0.56 units ($p = 0.28$). The difference-in-difference in this case, while not small, remains marginally significant ($p = 0.11$). The last two columns of Table 9 indicate there is no difference in the effect of the load on the ability of participants to best respond. While the effect of the load is highly significant in column (5), the interaction is far from significant in column (6). Calculating the marginal effects at the means of the two groups also indicates no significant difference in the load's effect: The low ability group deviates by an estimated 0.94 units more and the high ability group deviates by 0.75 units more.

Taken together, our results suggest that imposing a cognitive load on high ability players has more of an affect on their performance in the Hit games than it does for low ability players (although they both do significantly worse). In the p-Beauty Contest there is some suggestive evidence that low ability types are disproportionately worse at predicting what others will guess when the load is imposed but no evidence that there is a difference in the load's effect on the ability to best respond in this game.

5 Concluding Remarks

Given the amount of reasoning needed to make even the commonest of economic choices, it is important for economists to spend considerable effort to uncover the attributes that determine individual strategic sophistication, especially if greater sophistication is sought. One obvious place to start is to examine the hypothesis that economic and strategic reasoning is an expression of a more general set of cognitive skills. The major contributions of our experiments are to replicate existing results in a large sample, to dig considerably deeper by extending the current literature to new strategic settings and to identify the causal relationship between cognitive skills and strategic sophistication. In this sense our study is the first comprehensive treatment of this topic.

We begin, in Experiment 1, by gathering a large sample which allows us to first replicate previous results in the p-Beauty contest and to then expand the existing set of results. Here we show that similar associations arise in the Hit 15 game and when we use two common measures of cognitive ability. We then extend the literature in two directions. In Experiment 2 we modify the games to better capture the ability of participants to reason inductively, perform iterative dominance and predict the sophistication of others. Again, we find strong associations between cognitive ability, this time measured with Raven's standard progressive matrices, and sophistication. Those with higher ability win more Hit games, are better at predicting the average sophistication of others and come closest to best responding to their stated expectations. In Experiment 3 we extend the related literature in a second direction. Here we examine more closely the causal nature of the relationship between cognitive ability and strategic sophistication. We use a standard tool from cognitive psychology to randomly lower the cognitive ability of our participants. Specifically, we impose a cognitive load on the working memory of our players and show that this significantly affects game performance. Those asked to bear the load win fewer Hit games, are worse at predicting what others will guess in the beauty contest and submit guesses further from their best responses. In other words, we establish that cognitive ability is an important causal determinant of sophistication.

Although there are now a growing number of interesting results in the nascent literature on strategic sophistication, one of the strongest appears to be that general cognitive skills (both the more systematic and taxing "System 2" processes and the more spontaneous "System 1" processes) seem to play an important role. An interesting next step might be to examine whether non-cognitive skills like stable personality traits also affect sophistication.

6 Appendix

6.1 Instructions for Experiment 1

6.1.1 Hit 15

Consider the following two-person game: There is a "basket" in which people place "points". The two players take turns placing 1, 2, or 3 points in the basket. The person who places the 15th point in the basket wins a prize. Say you are playing and want to win the prize.

If you go first, how many points will you place in the basket? Please pick one of the answers below (1, 2, or 3).

If you go second and the other player has already put 2 points in the basket on her first turn, how many would you put in? Please pick one of the answers below (1, 2, or 3).

6.1.2 p-Beauty Contest

The last thing you will do is play a game with all the other people that are taking this survey. We expect a total of 200 people to participate and we will randomly sort the participants into groups of 20 for this game. To play you will guess a number between zero and one hundred. The person who wins the game is the person who picks the number that ends up being closest to one-half of the average of all the guesses in their session of 20 players. After the survey is finished, we will find the average of all the guesses, take half the average, and determine the winners for each session. The winners will each receive a prize of $25.
Input your guess (i.e. a number between 0 and 100) to the right.

6.2 Instructions for Experiments 2 & 3

6.2.1 Hit 5, 10, 15

Consider the following two-person game: There is a basket filled with pennies from which players alternate removing 1, 2, or 3 pennies per turn. The person who takes the final penny wins the game. You will play against the computer until one of you takes the final penny. We will play three rounds with differing numbers of initial pennies in the basket. You will be paid $5 for each round that you win. Please do not speak with other participants until the experiment is finished and everyone has been paid. Remember, in order to win you must be the one to take the LAST penny.

6.2.2 p-Beauty Contest

To play, you will guess a number between 0 and 20 up to two decimal places. The person who wins the game is the person who picks the number that ends up being closest to one-half the average of the guesses from all the 19 other participants. This winner will receive $25 in addition to his or her other earnings. There is also a second way to win. Using the chart below, the person who most accurately predicts the distribution of guesses will win another $25 (to be collected in his or her mailbox, as the distribution accuracy takes time to calculate).

To participate, first please place 19 X's in the chart below indicating what you think the other the 19 participants will guess (round their guesses to integers) so that each X corresponds to a different participant's anticipated guess. Second write your guess below in the space labeled "Guess."

7 References

Benjamin, D., Brown, S. and Shapiro, J., (2006). Who is behavioral? Cognitive ability and anomalous preferences, Harvard University Working Paper.

Berkowitz, L., (1993). Towards a general theory of anger and emotional aggression. In R. Wyer and T. Srul (Eds.), Advances in social cognition. (pp. 1-46). Hillsdale, NJ: Erlbaum.

Branas-Garza, P. and Rustichini, A., (2011). Organizing effects of testosterone and economic behavior: Not just risk taking. Plos One, 6(12), e29842.

Burks, S. V., Carpenter, J. P., Goette, L. and Rustichini, A., (2009). Cognitive skills affect economic preferences, strategic behavior, and job attachment. Proceedings of the National Academy of Sciences, 106(19), 7745-7750.

Burnham, T., 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.

Camerer, C., (2003). Behavioral game theory: Experiments on strategic interaction. Princeton: Princeton University Press.

Conway, A., Kane, M. and Engle, R., (2003). Working memory capacity and its relations to general intelligence. Trends in Cognitive Sciences, 7(12), 547-552.

Costa-Gomes, M., Crawford, V. and Broseta, B., (2001). Cognition and behavior in normal-form games: An experimental study. Econometrica, 69(5), 1193-1235.

Dohmen, T., Falk, A., Huffman, D. and Sunde, U., (2010). Are risk aversion and impatience related to cognitive ability? American Economic Review, 100(3), 1238-1260.

Dufwenberg, M., Sundaram, R. and Butler, D., (2010). Epiphany in the game of 21. Journal of Economic Behavior & Organization, 75(2), 132-143.

Fischbacher, U., (2007). Z-tree: Zurich toolbox for ready-made economic experiments. Experimental Economics, 10(2), 171-178.

Frederick, S., (2005). Cognitive reflection and decision making. Journal of Economic Perspectives, 19(4), 25-42.

Georganas, S., Healy, P. and Weber, R., (2010). On the persistence of strategic sophistication, Department of Economics, Ohio State Working Paper.

Gill, D. and Prowse, V., (2012). Cognitive ability and learning to play equilibrium: A level-k analysis.

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

Kalyuga, S., (2011). Cognitive load theory: How many types of load does it really need? Educational Psychology Review, 23(1), 1-19.

Levitt, S., List, J. and Sadoff, S., (2009). Checkmate: Exploring backward induction among chess players. American Economic Review, forthcoming.

Light, K., Kolata, S., Wass, C., Denman-Brice, A., Zagalsky, R. et al., (2010). Working memory training promotes cognitive abilities in genetically heterogeneous mice. Current Biology, 20(8), 777-782.

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

Raven, J., Raven, J. C. and Court, J. H., (2003). Manual for Raven's progressive matrices and vocabulary scales. San Antonio, TX: Harcourt Assessment.

Rydval, O., Ortmann, A. and Ostatnicky, M., (2009). Three very simple games and what it takes to solve them. Journal of Economic Behavior & Organization, 72(1), 589-601.

Stanovich, K. and West, R., (2000). Individual differences in reasoning: Implications for the rationality debate. Behavioral and Brain Sciences, 22(5), 645-726.

8 Tables and Figures

Table 1: The Distribution of Cognitive Ability (N=422)

	College Board Exam Level				
CRT Score	Lowest	Third	Second	Highest	total
0 Correct	0.02	0.04	0.06	0.02	0.14
1 Correct	0.04	0.09	0.11	0.03	0.27
2 Correct	0.02	0.05	0.11	0.08	0.26
3 Correct	0.01	0.06	0.14	0.12	0.33
total	0.09	0.24	0.42	0.25	1.00

Notes: Fractions of the total sample reported.
Kendall's tau: 0.23, $p<0.01$.

Table 2: Entrance Exam Scores and Strategic Sophistication

	(1) Hit 15	(2) p-Beauty
College Board Highest	0.717**	-18.413***
	(0.279)	(3.726)
College Board Second	0.464*	-12.241***
	(0.262)	(3.491)
College Board Third	0.420	-9.834***
	(0.276)	(3.679)
Female (I)	-0.046	8.003***
	(0.142)	(1.908)
Constant	0.911**	32.591***
	(0.376)	(5.018)
Controls included?	Yes	Yes
Observations	422	420

Notes: Tobit estimates; (standard errors); Controls include ethnicity, college year and household income; * p<0.10, ** p<0.05, *** p<0.01.

Table 3: Cognitive Reflection and Strategic Sophistication

	(1) Hit 15	(2) p-Beauty
CRT Score Three	0.855***	-17.198***
	(0.231)	(3.073)
CRT Score Two	0.410*	-9.454***
	(0.231)	(3.103)
CRT Score One	0.150	-3.798
	(0.229)	(3.075)
Female (I)	0.127	4.659**
	(0.145)	(1.938)
Constant	0.893**	31.280***
	(0.358)	(4.788)
Controls Included?	Yes	Yes
Observations	422	420

Notes: Tobit estimates; (standard errors); Controls include ethnicity, college year and household income; * p<0.10, ** p<0.05, *** p<0.01.

Table 4: Cognitive Ability and Iterative Reasoning

	(1) Guess <51	(2) 1 Step	(3) 2 Steps	(4) 3 Steps	(5) n Steps
College Board Highest	0.080***	0.177*	0.150	0.151	0.120
	(0.024)	(0.100)	(0.101)	(0.103)	(0.107)
College Board Second	0.060*	0.135	0.094	0.141*	0.121
	(0.032)	(0.093)	(0.090)	(0.085)	(0.082)
College Board Third	0.054**	0.086	-0.010	0.074	0.097
	(0.026)	(0.101)	(0.095)	(0.097)	(0.103)
Female (I)	-0.049**	-0.068	0.115**	0.116***	-0.088***
	(0.024)	(0.049)	(0.047)	(0.038)	(0.029)
Controls Included?	Yes	Yes	Yes	Yes	Yes
Observations	420	420	420	420	420

Notes: Mfx after probit estimates; (standard errors); Controls include ethnicity, college year and household income; * p<0.10, ** p<0.05, *** p<0.01.

Table 5: Cognitive Reflection and Iterative Reasoning

	(1) Guess <51	(2) 1 Step	(3) 2 Steps	(4) 3 Steps	(5) n Steps
CRT Score Three	0.104***	0.210**	0.253***	0.157**	0.174*
	(0.026)	(0.082)	(0.086)	(0.078)	(0.093)
CRT Score Two	0.055***	0.025	0.098	0.009	0.068
	(0.021)	(0.084)	(0.088)	(0.070)	(0.084)
CRT Score One	0.032	-0.015	0.071	-0.006	0.109
	(0.023)	(0.083)	(0.087)	(0.068)	(0.092)
Female (I)	-0.017	-0.020	-0.060	-0.078**	-0.064**
	(0.023)	(0.052)	(0.048)	(0.038)	(0.028)
Controls Included?	Yes	Yes	Yes	Yes	Yes
Observations	420	420	420	420	420

Notes: Mfx after probit estimates; (standard errors); Controls include ethnicity, college year and household income; * p<0.10, ** p<0.05, *** p<0.01.

Table 6: Cognitive Ability and p-Beauty Deviations

	(1) Deviation	(2) Deviation
College Board Highest	-14.876***	
	(3.216)	
College Board Second	-9.856***	
	(3.016)	
College Board Third	-7.848**	
	(3.178)	
CRT Score Three		-13.473***
		(2.665)
CRT Score Two		-8.885***
		(2.698)
CRT Score One		-3.616
		(2.673)
Female (I)	5.038***	2.500
	(1.643)	(1.676)
Constant	26.268***	25.588***
	(4.331)	(4.153)
Observations	420	420

Notes: The dependent variable is |guess-winning guess|; OLS estimates; (standard errors); Controls include ethnicity, college year and household income; * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table 7: Raven's Scores and Strategic Sophistication

	(1) Hit Games Won	(2) E[Mean Guess]	(3) Dev. Guess	(4) Dev. B.R.	(5) Dev. Win
Raven's Score	0.100***	-0.138**	-0.012***	-0.123***	-0.181***
	(0.017)	(0.058)	(0.004)	(0.027)	(0.044)
Female (I)	-0.776***	1.755**	0.049	0.237	0.868
	(0.211)	(0.812)	(0.055)	(0.400)	(0.625)
Constant	-2.816***	10.441***	0.962***	6.612***	9.921***
	(0.861)	(2.930)	(0.208)	(1.287)	(2.147)
Observations	64	64	64	64	64
Adj. R^2	0.385	0.162	0.109	0.267	0.288

Notes: OLS estimates; (standard errors); Controls include ethnicity and college year; * p<0.10, ** p<0.05, *** p<0.01.

Table 8: Cognitive Load and Strategic Sophistication

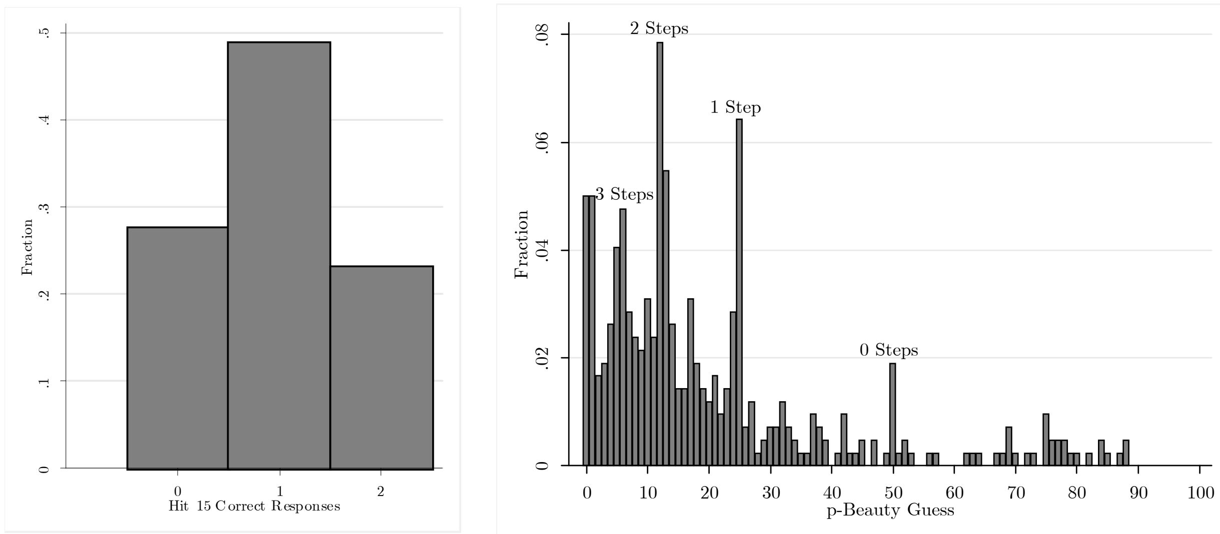
	(1) Hit Games Won	(2) E[Mean Guess]	(3) Dev. Guess	(4) Dev. B.R.	(5) Dev. Win
Cognitive Load	-0.514***	1.129**	-0.028	0.919***	0.292
	(0.187)	(0.512)	(0.035)	(0.224)	(0.416)
Female (I)	-0.608***	1.478**	0.093**	0.302	1.022**
	(0.180)	(0.566)	(0.040)	(0.292)	(0.487)
Constant	1.766***	4.463***	0.445***	0.976***	1.606***
	(0.200)	(0.574)	(0.041)	(0.216)	(0.348)
Observations	118	118	118	118	118
Adj. R^2	0.116	0.097	0.037	0.112	0.041

Notes: OLS estimates; (standard errors); Controls include ethnicity and college year; * p<0.10, ** p<0.05, *** p<0.01.

Table 9: Is there a differential effect of the cognitive load?

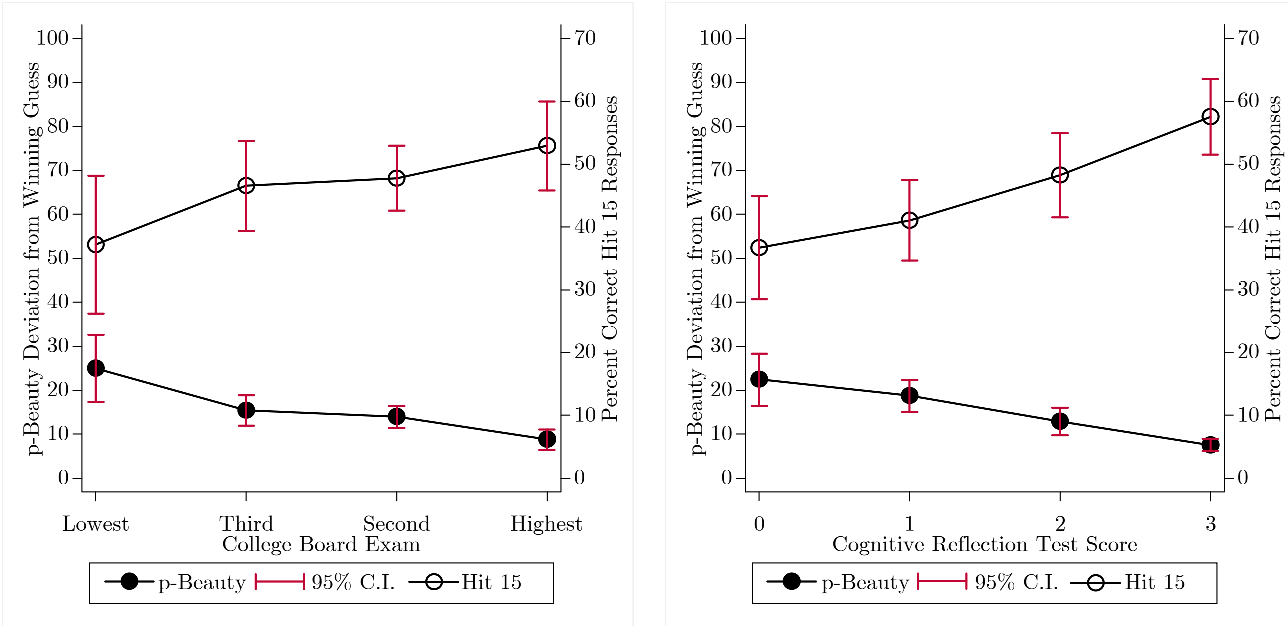
	(1) Hit Won	(2) Hit Won	(3) E[Guess]	(4) E[Guess]	(5) Deviation	(6) Deviation
Raven's Score	0.073***	0.099***	-0.206***	-0.158***	-0.124***	-0.114***
	(0.012)	(0.015)	(0.035)	(0.048)	(0.017)	(0.026)
Cognitive Load	-0.475***	2.006*	1.019**	5.559*	0.853***	1.817
	(0.167)	(1.045)	(0.434)	(2.901)	(0.168)	(1.446)
Raven's×Load		-0.054**		-0.099		-0.021
		(0.023)		(0.062)		(0.029)
Female (I)	-0.562***	-0.555***	1.349***	1.362***	0.225	0.227
	(0.164)	(0.164)	(0.499)	(0.497)	(0.245)	(0.247)
Constant	-1.579**	-2.815***	13.893***	11.633***	6.646***	6.167***
	(0.615)	(0.722)	(1.710)	(2.341)	(0.842)	(1.246)
Observations	118	118	118	118	118	118
Adj. R^2	0.311	0.333	0.281	0.286	0.407	0.404

Notes: OLS estimates; (standard errors); Controls include ethnicity and college year; * $p<0.10$, ** $p<0.05$, *** $p<0.01$.



panel (a) panel (b)

Figure 1: The distribution of strategic choices in experiment 1.



panel (a) panel (b)

Figure 2: Cognitive ability predicts strategic choices in experiment 1.

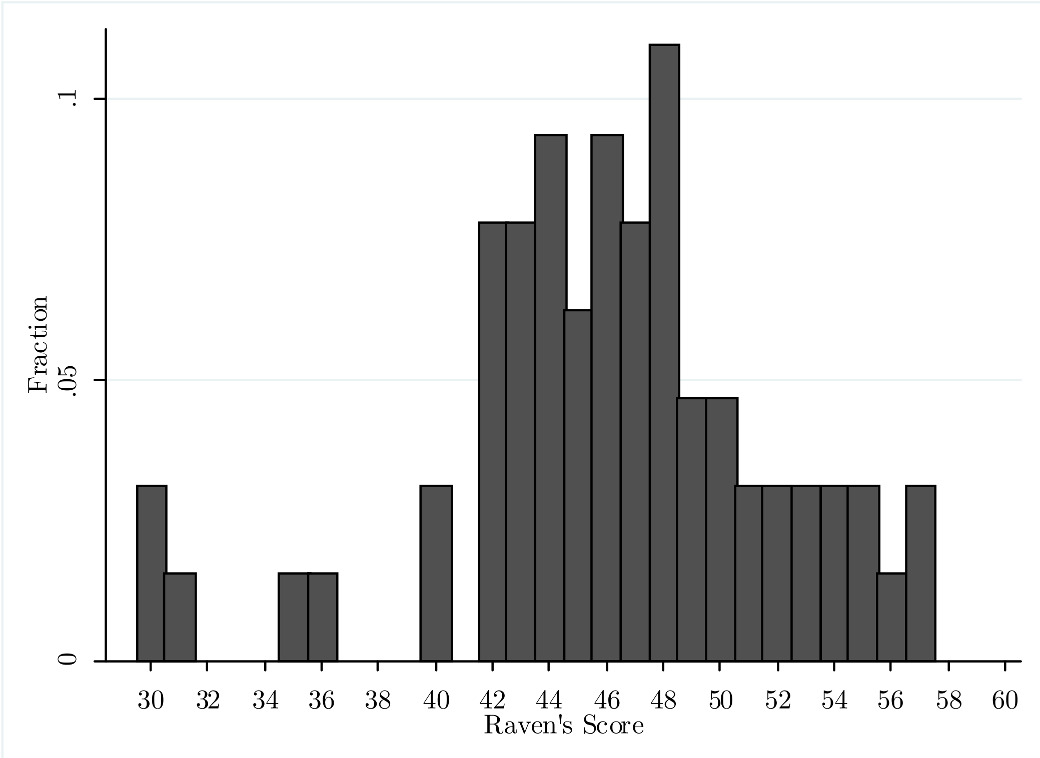
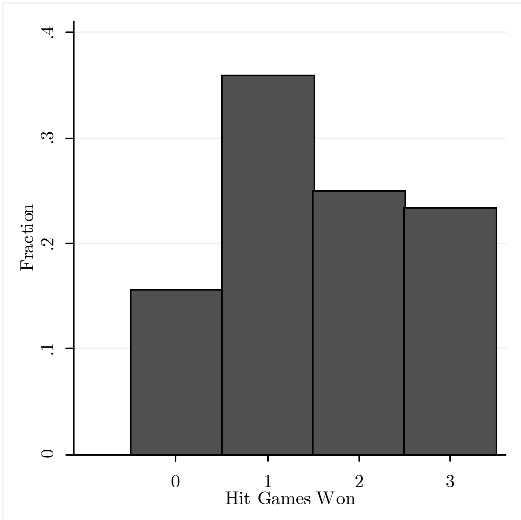
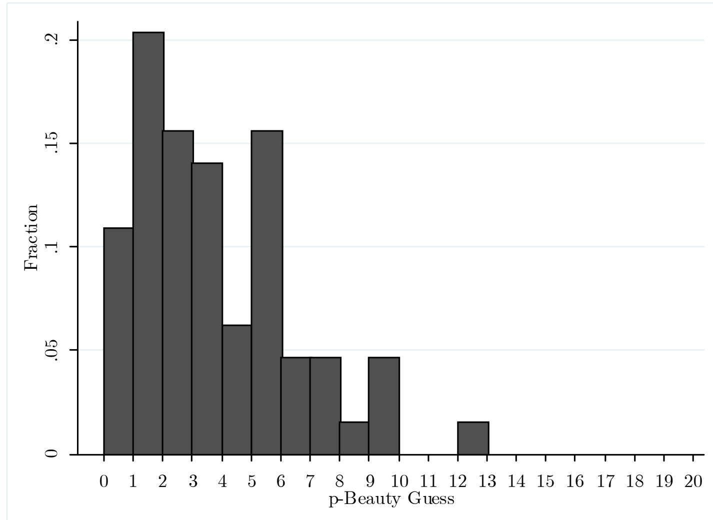


Figure 3: The distribution of cognitive ability in experiment 2.

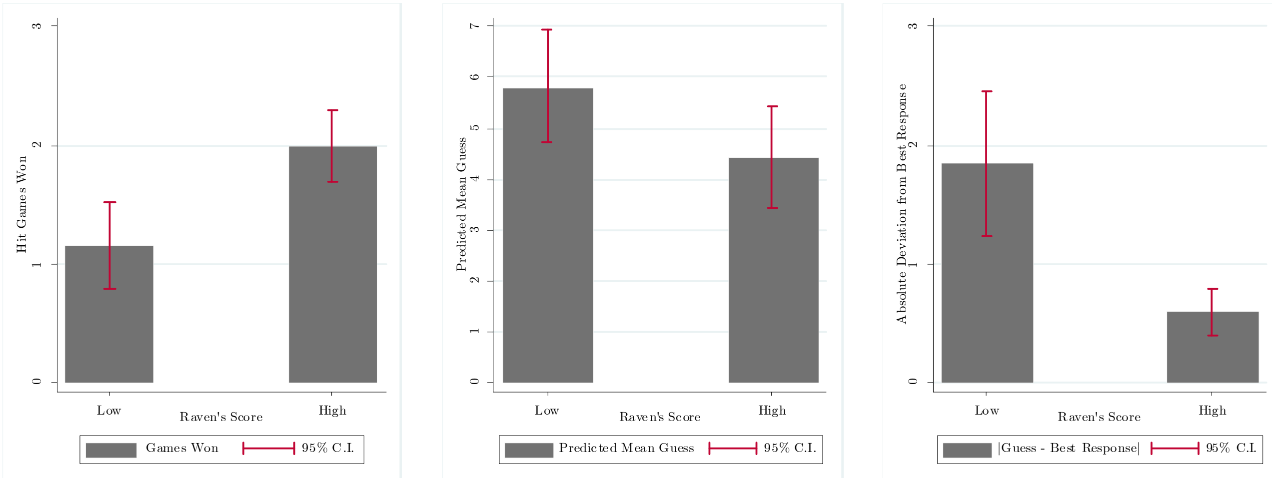


panel (a)



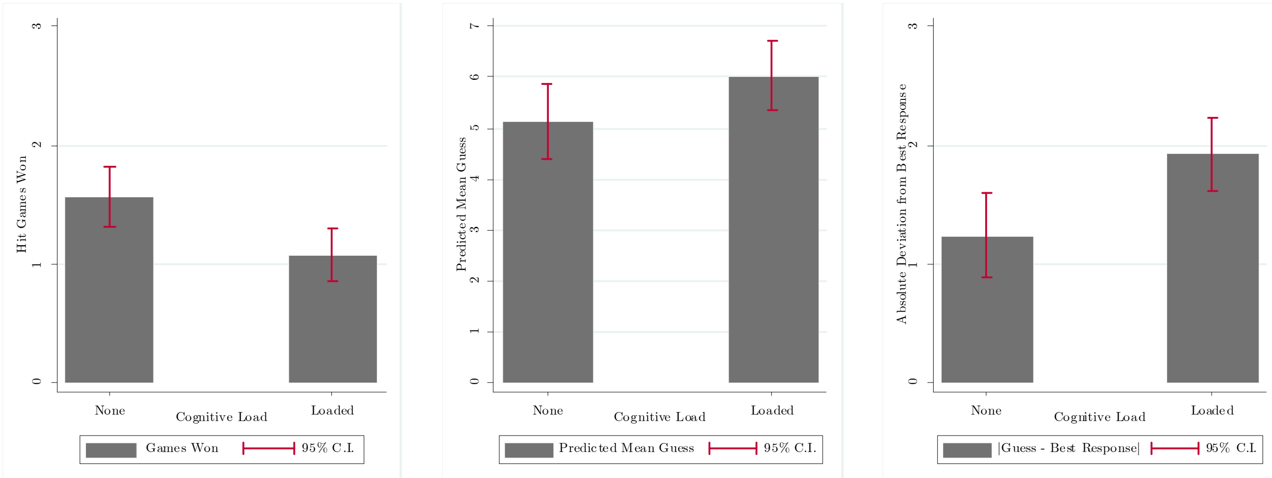
panel (b)

Figure 4: The distribution of strategic choices in experiment 2.



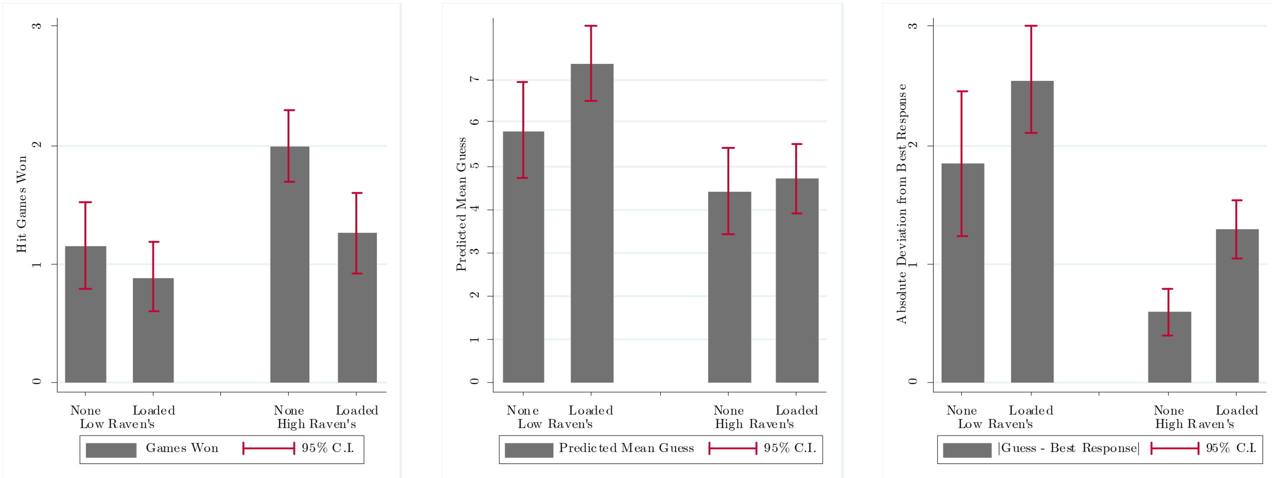
panel (a) panel (b) panel (c)

Figure 5: Cognitive ability predicts player performance.



panel (a) panel (b) panel (c)

Figure 6: The causal affect of cognitive ability.



panel (a) panel (b) panel (c)

Figure 7: Where is the cognitive load borne?