Effort is not a monotonic function of skills: results from a global mobile experiment*

Konrad Grabiszewski† Alex Horenstein ‡

2020

Abstract

At the core of economic theory is the monotonicity hypothesis: an agent's effort, as a function of their skills, is either non-decreasing or non-increasing, but not both. To test this hypothesis, we use data from *Blues and Reds*, a mobile app designed to conduct economic experiments that consists of a series of interactive puzzles. The sample includes 6,463 subjects from 141 countries. We measure subjects' skills and effort levels using their response times. We replicate the same test 22 times. Surprisingly, each time we find that the optimal effort is not a monotonic function of skills but rather has a U-shape contradicting the monotonicity hypothesis.

JEL Codes: D01, C72, C80, C90.

Key words: monotonicity of choice, effort, skills, experimental game theory, mobile experiment.

*We are grateful to Andrew Leone, who as the Vice Dean for Faculty Development and Research at the University of Miami Business School, helped us secure financing to conduct this project. We thank the associate editor and two anonymous referees for their constructive comments, which helped us to improve the paper.

†Mohammad bin Salman College, Saudi Arabia; konrad.grabiszewski@gmail.com

‡Department of Economics, University of Miami, FL, USA; horenstein@bus.miami.edu

1 Introduction

The monotonicity of choice is one of the key tenants of economic theory. It states that the agent's effort is a monotonic function of their skills; agents with different skills should (according to the theory) choose different levels of effort. If effort and skills are complements (e.g., Dewatripont et al. (1999a), Dewatripont et al. (1999b), Bonatti and Hörner (2017), and Fudenberg and Rayo (2019)), then effort is non-decreasing in skills. However, if effort and skills are substitutes (e.g., Holmström (1999) and Cisternas (2018)), then effort is non-increasing in skills.[1] In this article, we test the hypothesis of monotonicity

We design an experiment to test the monotonicity hypothesis and repeat the same test in 22 different scenarios. We reject the hypothesis of monotonicity in all of them. Rather, in each scenario, we find that the relationship between the optimal effort and skills is U-shaped. Whether effort and skills are complements or substitutes depends on the decision-maker's skill levels: they are substitutes for low levels of skills and complements for high ones. Figure 1 captures the main result of this article.[2]

[Figure 1 here.]

When testing the monotonicity hypothesis, the main challenge is to measure subjects' skills and effort levels. Our experiment was designed to ensure that both variables, skills and effort, are objectively measured.[3]

[1] From a modelling perspective, monotonicity is about second-order derivatives. Consider the standard optimization problem $\max_a B(s,a) - C(a)$, where a denotes effort, s stands for skills, B is benefit from effort (non-decreasing and concave in a), and C is cost of effort (increasing and convex in a). The optimal level of effort is a^*, and the sign of derivative $\frac{da^*}{ds}$ is the same as the sign of the cross-derivative B_{sa}. I.e., a^* is in a monotonic relation with s if the sign of B_{sa} does not change from negative to positive, or vice versa.

[2] While this article focuses on monotonicity in the context of effort and skills, starting with the Spence-Mirrless single crossing property (Mirrlees (1971) and Spence (1973)), the field of monotone comparative statics (e.g., Milgrom and Shannon (1994), Athey (2002), and Quah and Strulovici (2009)) has been of major importance in various areas of economics literature. The basic idea is that agents with different characteristics will make different choices. Monotonicity is crucial in establishing the existence of equilibria in various games (e.g., Milgrom and Roberts (1990), Vives (1990), Kreps and Sobel (1994), and Athey (2001)).

[3] As indicated in Charness et al. (2018), when it comes to measuring effort, "there are two major methodological paradigms in this literature: stated effort and real effort." The same can be said about measuring

To collect the data, we recruited a team of software developers to create *Blues and Reds*, a mobile app that has been available for iOS and Android devices since August 2017 in every corner of the world. Everyone who installs and plays *Blues and Reds* becomes a subject of the experiment. The motivation behind conducting an experiment on the subjects' smartphones and tablets is the desire to explore and exploit what the global availability of mobile technology and devices has to offer for empirical research.[4]

The data in this article comes from 22 interactive puzzles in *Blues and Reds*. Figure 2 in Section 2 depicts a screenshot from the app with an example of a puzzle that subjects play. Each puzzle is a zero-sum, no-tie, finite dynamic game with perfect and complete information in which a human subject plays against Artificial Intelligence (AI). The objective is to win against AI, which is programmed to win against the subject. In each puzzle, it is possible for a subject to win, but this requires following the unique path as prescribed by backward induction; any deviation from that path results in the subject losing.

Each puzzle should be treated as a separate experiment in which we test the hypothesis of monotonicity of subjects' effort levels. In each puzzle, we measure subjects' skills and effort levels using their response times which are recorded at each round of a multi-round puzzle.[5] Next, we implement standard semi-parametric methods to estimate the functional form between effort and skills. In each of the 22 puzzles, that functional form has a U-shape.

skills: either stated by the subject or real (i.e., assessed by observing subject's behavior). In our experiment, both skills and effort are real.

[4]We encourage the readers to download and play *Blues and Reds*. The app can be downloaded from Google Play or the App store. Relevant links and additional information are available on `http://www.bluesandreds.com/`

[5]Response time (RT) is a relatively novel addition to the economists' toolbox and was almost unseen in economics articles just 15 years ago. The change in economists' attitudes towards RT can be attributed to the 2004 Presidential Address to the Econometric Society by Ariel Rubinstein (Rubinstein (2006)). Since then, the economic literature (both empirical and theoretical) using RT has very quickly grown (e.g., Ofek et al. (2007), Rubinstein (2007), Milosavljevic et al. (2010), Fehr and Rangel (2011), Arad and Rubinstein (2012), Rubinstein (2013), Eliaz and Rubinstein (2014), Krajbich et al. (2014), Woodford (2014), Agranov et al. (2015), Caplin and Martin (2016), Geng (2016), Rubinstein (2016), Echenique and Saito (2017), Gill and Prowse (2017), Lohse et al. (2017), Masiliunas (2017), Recalde et al. (2018), and Enke and Zimmermann (2019)). For comprehensive reviews of the relevant literature, see Clithero (2016) and Spiliopoulos and Ortmann (2017). A unique feature of our data is that we record RT at each round of a dynamic game.

In light of the replication crisis that has become a major issue in empirical studies in social sciences (e.g., Maniadis et al. (2014), Open Science Collaboration (2014), Open Science Collaboration (2015), Camerer et al. (2016), Munafò et al. (2017), and Camerer et al. (2018)), it is important to emphasize the fact that the same qualitative result has been replicated 22 times. This replication success in tandem with the fact that our experiment is based on a large and varied sample pool (6,463 subjects from 141 countries) confirms the reliability of our results.

The rest of the article is organized as follows. Section 2 describes the experiment. Section 3 discusses the measurements of effort and skills. Section 4 includes the main empirical analysis. Section 5 concludes.

2 Experimental Design

This article relies on data from the 22 interactive puzzles from *Blues and Reds*. Figure 2 presents a typical puzzle. Each puzzle is a zero-sum finite dynamic game with perfect and complete information played against Artificial Intelligence (AI). The goal is to win against AI whose objective is to win against a human subject; hence, there are no ties.

[Figure 2 here.]

The rules are the same in each puzzle. The puzzle starts with a subject moving the RoboToken (a spherical object placed in the center) by choosing which blue bridge the RoboToken crosses. Then, AI selects a red bridge for the RoboToken to cross. And so on. The puzzle ends when the RoboToken lands on a blue node (subject wins) or a red node (subject loses).

Each puzzle is a classical game-theoretic tree and has a symmetric structure $N_1.N_2.N_3.N_4.N_5.N_6$ where N_i denotes the number of bridges (branches) at each node at round i. The puzzle in Figure 2 is denoted as 4.2.2 (for clarity, puzzle labels do not contain zeros).

We use data from 3-, 4-, 5-, and 6-round puzzles. Table 1 lists all the 22 puzzles from the experiment grouped according to the number of rounds. Each subject has an opportunity to play the same 22 puzzles but the specific order of puzzles a subject plays is random. The average number of puzzles played by a subject is 5.78 (with a standard deviation of 6.10).

[Table 1 here.]

It is possible to win each puzzle but in order to do so a subject must follow an equilibrium path as prescribed by backward induction. Importantly, if, at any round, a subject deviates from that path, then she is guaranteed to lose as AI was programmed to exploit subjects' mistakes.

Once a person installs and opens *Blues and Reds*, she becomes a subject in the experiment, which begins with a mandatory tutorial that teaches the objectives of *Blues and Reds* and how to play it. Subjects can return to the tutorial puzzles at any time and at zero cost. The non-tutorial puzzles generate the relevant data for our article and can only be played once irrespective of whether the subject wins or loses.

We record whether the subject wins or loses for each puzzle and each subject. More importantly for the purposes of this article, we also collect the subject's response time (RT), measured in seconds, at each round. For example, in a 3-round puzzle, *Blues and Reds* records the subject's RT1 and RT3, response times at the first and third rounds, respectively. Response times depict the subject's reasoning processes and allow us to measure her effort and skills (see Section 3).

Data was collected from August 15, 2017 to February 6, 2018.[6] In each puzzle, observations with a total time[7] above the 95th percentile were removed. The final data consists of 35,827

[6]The design of *Blues and Reds* allows us to conduct several experiments in which we analyze various questions from dynamic game theory. In Grabiszewski and Horenstein (2019c), following in the spirit of Rubinstein (2016), we design a novel method of profiling players in dynamic games. In Grabiszewski and Horenstein (2019a), we develop an empirical measure of tree complexity. In Grabiszewski and Horenstein (2019b), we ask whether people perceive the reality in accordance with theory.

[7]Total time, denoted as TT, is the sum of round-based response times. For example, in a 3-round puzzle,

observations. The sample pool includes 6,463 subjects from 141 countries. *Blues and Reds* recorded IP addresses for 5,746 subjects; based on this sub-sample, Table 2 lists the 20 most popular countries in the sample.

[Table 2 here.]

3 Measuring Effort and Skills

Measuring effort in *Blues and Reds* is a relatively straightforward exercise. To solve a puzzle, a subject needs to reason strategically. Reasoning requires time, which is a scarce and costly resource. Rubinstein (2016) notes that "choices which require more cognitive activity will result in longer response times," a sentiment repeated in Enke and Zimmermann (2019) who note that response time is "a commonly used proxy for cognitive effort." Both in economics (see citations in footnote 5) and psychology (e.g., the area of mental chronometry (Jensen (2006)), it is standard to measure cognitive effort as total time TT, i.e., the sum of round-based response times (see footnote 7). In this article, we follow the path established in the literature and measure effort as TT. Higher TT implies more effort.

Measuring skills in *Blues and Reds* is a more challenging exercise. This is because the concept of "skills" is context-specific. For instance, while both solving mathematical problems and writing poems require exerting cognitive effort (measured as TT), the specific skills for these tasks are not the same and, consequently, should be measured in different ways. This is also the case of skills measured in *Blues and Reds*. Winning puzzles in *Blues and Reds* requires backward induction reasoning. Consequently, to have high skills means to reason in a way consistent with the backward induction algorithm.

To measure backward-inducting skills, we construct RRT1, the relative response time at the first round defined as $RRT1 = \frac{RT1}{TT}$. We posit that higher RRT1 is equivalent with higher

TT = RT1 + RT3.

skills. Before we elaborate on this equivalency, in Table 3, we present the summary statistics of our main empirical measures of interest for each puzzle: effort (TT) and skills (RRT1).

[Table 3 here.]

To establish the validity of RRT1 as a measure of backward-inducting skills we demonstrate the aforementioned equivalency: (1) if a subject is more likely to backward induct, then her RRT1 is higher, and (2) if a subject's RRT1 is higher, then she is more likely to behave consistently with the backward induction algorithm.

$$\textit{higher RRT1} \iff \textit{higher probability of backward induction} \tag{1}$$

To show that a subject who is more likely to backward induct is characterized with a higher RRT1, we begin with an observation: in terms of time allocation, the most important feature of the backward induction algorithm is that its implementation takes place at the very beginning of a dynamic game. Subjects who backward induct spend most of their total time on the first round, while those who incorrectly backward induct or do not backward induct at all allocate a non-negligible amount of total time to the later rounds. Since it is the percentage of total time spent at the first round that matters, subjects reasoning more consistently with backward induction will have higher RRT1.

To show that a higher RRT1 indicates a higher probability of a subject backward inducting, we rely on three empirical exercises. In the first exercise, for a given puzzle, we divide subjects into RRT1-terciles: Low (L), Medium (M), and High (H). The Low tercile corresponds to all subjects having RRT1 less than or equal to the percentile 33.3%. The Medium tercile corresponds to the subjects having RRT1 higher than the 33.3% percentile and less than or equal to the 66.6% percentile. Finally, the High RRT1 tercile corresponds to subjects with a TT higher than the 66.6% percentile.[8]

[8]If we divide the data into quintiles instead of terciles all results hold. These results are available from

If RRT1 measures skills, then for each puzzle, we expect the percentage of winners to be statistically non-decreasing starting with the 1st RRT1-tercile and ending with the 3rd RRT1-tercile. As Table 4 shows, this is true for every puzzle. RRT1 is an outstanding predictor of subjects backward inducting. This measure never fails in the sense that for each puzzle, higher RRT1 is associated with a higher probability of backward inducting.

[Table 4 here.]

Table 4 shows that the unconditional probability of backward inducting is non-decreasing in skills. However, from the theoretical perspective, we also expect that when skills increase but effort is kept fixed, the probability – this time conditional – of a subject backward inducting should not decrease either. This is confirmed in our second empirical exercise.

To keep effort fixed, in each puzzle, we start with dividing subjects into TT-quintiles; each TT-quintile represents a fixed value of effort.[9] Next, each TT-quintile is divided into RRT1-quintiles. These RRT1-quintiles represent various values of skills for a fixed value of effort. We expect that, for each puzzle and each TT-quintile, the percentage of winners is statistically non-decreasing starting with the 1st RRT1-quintile and ending with the 5th RRT1-quintile. As Table 5 shows, this is true in every puzzle. Again, we observe that RRT1 – this time, conditional on TT – is an outstanding predictor of subjects backward inducting.

[Table 5 here.]

In our final exercise, we estimate the following logit model for each of the 22 puzzles.

$$Logit(Y) = \alpha_0 + \alpha_1 RRT1 \tag{2}$$

the authors upon request.

[9] We corroborated that in each TT-quintile and each puzzle, the variable TT is relatively constant by calculating the metric $\frac{\text{standard deviation of TT}}{\text{average of TT}}$. We found that this metric is close to zero in each quintile and each puzzle, meaning that there is little variability in TT within each quintile. This supports the TT-quintile as representing a fixed value of effort.

Y is the dependent variable in the regression and captures whether the subject backward inducted ($Y_i = 1$) or did not backward induct ($Y_i = 0$), α_0 is the intercept, and RRT1 is a measure of skills. Table 6 shows the results of estimations. N denotes the number of observations. For α_1, we provide heteroskedastic robust standard errors (in parenthesis), and stars denote the significance level (*** 1%, ** 5%, and * 10%).

[Table 6 here.]

For each puzzle, the estimated coefficient of variable RRT1 is always positive and statistically significant at less than the 1% level of significance. This confirms that a higher RRT1 increases the likelihood of observing a subject winning a puzzle and provides additional support for RRT1 as a measure of backward-inducting skills.

The above discussion suggests that, at least for the puzzles in *Blues and Reds* and for the purposes of our paper, RRT1 is a very good predictor of a subject's ability to backward induct. The usefulness of RRT1 extends beyond the scope of this paper as it can be employed in any empirical study where there is a need to control for subjects skills. At the same time, it is important to discuss the limitations of RRT1 as a measure of skills.

First, consider a dynamic game in which the subject's choice at the first round is irrelevant. This would be the case of, for example, a tree in which there are N actions at the first round and each of the following N subgames is identical. If the subject realizes that the first-round choice is inconsequential, their response time at the first round would be very small – after all, it is not necessary to reason at the very beginning of a tree. In this case, RRT1 would not capture the skills.

Note, however, this is not an issue that is relevant in our experiment. As previously explained, making a mistake at any round, including the first round, results in a subject's loss. Consequently, in *Blues and Reds*, subjects cannot afford the luxury of not reasoning at the very beginning of a puzzle and their RRT1 correctly reflects their skills.

Second, suppose that, while playing a dynamic game, a subject realizes their opponent is not fully rational. It might be beneficial for the subject to exploit that irrationality. If this is the case, then a re-thinking of the follow-up actions is required[10] and the subject spends a significant amount of reasoning time at later rounds. Hence, RRT1 would not be a good proxy of skills.

However, due to the payoff structure, in all the puzzles in *Blues and Reds*, it is never beneficial to re-consider the initially designed strategy even if the subject believes AI to be irrational.[11] To explain why this is the case, consider for example the puzzle in Figure 2. Here, backward induction selects action "up" at the first node as it guarantees a win no matter how rational AI is. At the same time, every other of the remaining three actions at the initial node results with an uncertain win (and a certain loss if AI is rational). For example, if the subject picks "down," it is possible that even irrational AI chooses an action that forces the subject to pick a red node. Since a certain win is better than an uncertain win, there is nothing to be gained by trying to exploit an (incorrectly assumed) irrationality of AI. The same logic applies to all the puzzles in *Blues and Reds* and, as such, RRT1 is an appropriate measure of skills.

Impact of Experience. A reader might ponder about the impact of experience on the subject's effort and skills. To that end, we construct an independent variable Seq which corresponds to the order in which a puzzle appeared in the subject's sequence of puzzles. For instance, if an observation comes from the 4th puzzle that a subject plays, then $Seq = 4$.

[10] As it is documented in the literature (e.g., Palacios-Huerta and Volij (2009), Agranov et al. (2012), Alaoui and Penta (2016), Fehr and Huck (2016), and Gill and Prowse (2016)), belief about opponent's rationality affects the subject's behavior.

[11] In fact, in *Blues and Reds*, there is no reason to question AI's rationality. AI was programmed to impeccably implement the backward induction algorithm and to never make mistakes: there is no evidence to support the belief that AI is anything but fully rational.

For each of the 22 puzzles, we estimate the following two regression models.

$$RRT1 = a_0 + a_1 Seq \tag{3}$$

$$TT = b_0 + b_1 Seq \tag{4}$$

Table 7 below shows the results of the estimations and a detailed analyses follows. N denotes the number of observations. Adj R^2 is adjusted R^2. For a_1 and b_1, we provide heteroskedastic robust standard errors (in parentheses), and stars denote the significance level (*** 1%, ** 5%, and *10%). We restrict each regression to players who win the corresponding puzzle, as these players are the ones most likely to have obtained some value from learning.

[Table 7 here.]

In the case of skills (RRT1), as expected, the variable *Seq* is, in the majority of cases, positive. In nine cases it is positive and statistically significant. It is only negative and statistically significant in two cases. This shows a support for learning-by-doing (the further in the sequence that a specific puzzle appears, the higher the skills. When it comes to effort (TT), in most cases the coefficient of *Seq* is negative and statistically significant. It is only positive and statistically significant in two cases. This indicates that, in general, the further in the sequence that a specific puzzle appears, the lower effort a winning subject exerts.

4 Effort and Skills: Estimating the Functional Form

Equipped with the subjects' measures of skills (RRT1) and effort (TT), we study the relationship between these two variables. The fundamental hypothesis is that effort is a monotonic function of skills. We conduct the following empirical analysis for each puzzle in *Blues and Reds*. To estimate the functional form between RRT1 and TT, we use the semi-parametric

method developed in Robinson (1988).[12] More precisely, we estimate the equation,

$$y = \alpha + \beta x + f(z) + \varepsilon \tag{5}$$

where y is TT, x is a control variable, z is RRT1, and ε is the error term. The control variable is Seq, the order in which a given puzzle appears in the subject's sequence of puzzles (recall that the sequence of puzzles is random; hence, a given puzzle might be the 1st that a subject A plays but the 22nd that a subject B plays.)

At the end of this section, we present the estimation results. In each figure, the vertical axis represents effort (TT) while the horizontal axis is skills (RRT1). The figures show the fitted values of the semi-parametric estimations using a Gaussian kernel with an optimal bandwidth. The results are robust to different kernels and bandwidths.

In each figure, the blue line is the fitted value of $f(z)$ while the gray area is the 95% confidence interval for the kernel estimation. $f(z)$ is the object of interest as it captures the relationship between effort (TT) and skills (RRT1). The monotonicity hypothesis, a hallmark of economic theory, posits that $f(z)$ should be either non-decreasing or non-increasing. Each figure constitutes a separate test of the monotonicity hypothesis. To corroborate the robustness of our results, we replicate the same test 22 times.

In each figure, we find that the relationship between effort and skills is not monotonic but resembles a U-shaped curve. We observe in some figures that $f(z)$ is initially increasing for very low values of RRT1. This is simply a byproduct of having very sparse data for those values as illustrated by the wider estimated confidence intervals (gray area). Thus, the importance of the results for very low RRT1 can be discounted since the kernel's fit is not precise within that range. The parameter accompanying the control variable Seq is negative and statistically significant in 20 out of the 22 puzzles (and it is not statistically significant

[12] The estimations were conducted in STATA using the Semipar package developed by Verardi and Debarsy (2012).

in the remaining two puzzles). This, similarly to what we found in Section 3, indicates that as players gain experience, they apply less effort when solving a puzzle.

To summarize, we tested the monotonicity hypothesis in 22 different cases and rejected it 22 times. While our data does not allow to establish what causes this non-monotonicity, we discuss a plausible explanation below.

When solving a task – not just a puzzle in *Blues and Reds* – the choice of agent's effort will depend on their level of skills. Since different levels of skills would imply different optimal levels of effort, the optimization program implicitly requires that people correctly know their own skills. Yet this assumption is far from inconsequential.

As the psychology literature indicates, people "suffer, for lack of better terms, a meta-ignorance, remaining ignorant of the multitude ways they demonstrate gaps in knowledge" (Dunning (2011, p. 251)). This meta-ignorance depends on the skills as "incompetent individuals have more difficulty recognizing their true level of ability than do more competent individuals and that a lack of metacognitive skills may underlie this deficiency" (Kruger and Dunning (1999, p. 1122)). More importantly, "incompetent individuals, compared with their more competent peers, will dramatically overestimate their ability and performance relative to objective criteria" (Kruger and Dunning (1999, p. 1122)).

This phenomenon, called the *Dunning-Kruger effect* or the *Unskilled and unaware effect*, is well-established in the literature (e.g., Kruger and Dunning (1999), Dunning et al. (2003), Ehrlinger and Dunning (2003), Ehrlinger et al. (2008), Ferraro (2010), Dunning (2011), Schlösser et al. (2013), Staub and Kaynak (2014), Kim et al. (2015), and Sanchez and Dunning (2018)).[13] The mechanism behind the Dunning-Kruger effect is explained in Dunning et al (2003): "people fail to recognize their own incompetence because that incompetence

[13]In fact, the Dunning-Kruger effect attracted attention outside of academic circles: Encyclopedia Britannica (`https://www.britannica.com/science/Dunning-Kruger-effect`), Forbes (`{https://www.forbes.com/sites/markmurphy/2017/01/24/the-dunning-kruger-effect-shows-why-some-people-think-theyre-great-even-when-their-work-is-terrible/#1a6b40ed5d7c/}`), or Reader's Digest (`https://www.rd.com/advice/dunning-kruger/`).

carries with it a double curse. In many intellectual and social domains, the skills needed to produce correct responses are virtually identical to those needed to evaluate the accuracy of one's responses. The skills needed to produce logically sound arguments, for instance, are the same skills that are necessary to recognize when a logically sound argument has been made. Thus, if people lack the skills to produce correct answers, they are also cursed with an inability to know when their answers, or anyone else's, are right or wrong. They cannot recognize their responses as mistaken, or other people's responses as superior to their own. In short, incompetence means that people cannot successfully complete the task of metacognition, which, among its many meanings, refers to the ability to evaluate responses as correct or incorrect."

In the context of our experiment, the Dunning-Kruger effects posits that a subject's skills indicate not only how good they are at backward induction but also how self-aware they are of their own competencies when it comes to solving dynamic games. Lower skills imply larger overconfidence. In Figure 1, Chris with skills s_C overestimates his own skills more than Bob does, whose overconfidence is larger than that of Ann. If Chris believes that he is as good as Ann is, then he chooses the same level of effort as she does. This phenomenon of meta-ignorance decreasing with skills ultimately leads to a U-curve.

While the Dunning-Kruger effect is a well-documented and understood phenomenon and explains the results in the 22 puzzles studied in this paper, we emphasize that it is only a hypothetical explanation. Data limitations prevent accurate testing of this hypothesis. Then, we leave for future research the investigation of the Dunning-Kruger effect in the context of effort-skills non-monotonicity or the identification of additional channels that explain the U-curve relationship between effort and skills.

[Figure 3 – Figure 24 here.]

5 Conclusions

Since smartphones and tablets are in the hands of billions of people, it is only natural to engage the mobile technology for academic research. We hired a team of developers to create a mobile app, *Blues and Reds*, which allows us to conduct global experiments. In this article, we use the data from *Blues and Reds* to test one of the most fundamental hypotheses in economics; namely, that effort has a monotonic relationship with skills.

Data comes from 22 interactive puzzles, each played against Artificial Intelligence and treated as a separate experiment. In each puzzle and for each subject, we measure her skills and effort. The hypothesis of the monotonicity of choice posits that effort is either non-decreasing or non-increasing — but importantly not both — function of skills. We estimate the functional form between effort and skills. Repeating the same exercise 22 times, each time, we find that the relationship between effort and skills resembles a U-shaped curve; this rejects the hypothesis of the monotonicity of choice.

References

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

Agranov, M., E. Potamites, A. Schotter, and C. Tergiman (2012): "Beliefs and endogenous cognitive levels: An experimental study," *Games and Economic Behavior*, 75, 449–463.

Alaoui, L. and A. Penta (2016): "Endogenous Depth of Reasoning," *Review of Economic Studies*, 83, 1297–1333.

Arad, A. and A. Rubinstein (2012): "Multi-dimensional iterative reasoning in action:

The case of the Colonel Blotto game," *Journal of Economic Behavior & Organization*, 84, 571–585.

Athey, S. (2001): "Single Crossing Properties and the Existence of Pure Strategy Equilibria in Games of Incomplete Information," *Econometrica*, 69, 861–889.

——— (2002): "Monotone Comparative Statics under Uncertainty," *Quarterly Journal of Economics*, 117, 187–223.

Bonatti, A. and J. Hörner (2017): "Career Concerns with Exponential Learning," *Theoretical Economics*, 12, 425–475.

Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen, and H. Wu (2016): "Evaluating replicability of laboratory experiments in economics," *Science*, 351, 1433–1436.

Camerer, C. F., A. Dreber, F. Holzmeister, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. A. Nosek, T. Pfeiffer, A. Altmejd, N. Buttrick, T. Chan, Y. Chen, E. Forsell, A. Gampa, E. Heikensten, L. Hummer, T. Imai, S. Isaksson, D. Manfredi, J. Rose, E.-J. Wagenmakers, and H. Wu (2018): "Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015," *Nature Human Behaviour*, 2, 637–644.

Caplin, A. and D. Martin (2016): "The Dual-Process Drift Diffusion Model: Evidence from Response Times," *Economic Inquiry*, 54, 1274–1282.

Charness, G., U. Gneezy, and A. Henderson (2018): "Experimental methods: Measuring effort in economics experiments," *Journal of Economic Behavior & Organization*, 149, 74–87.

Cisternas, G. (2018): "Career Concerns and the Nature of Skills," *American Economic Journal: Microeconomics*, 20, 152–89.

Clithero, J. A. (2016): "Response Times in Economics: Looking Through the Lens of Sequential Sampling Models," *working paper*.

Dewatripont, M., I. Jewitt, and J. Tirole (1999a): "The Economics of Career Concerns, Part I: Comparing Information Structures," *Review of Economic Studies*, 66, 183–198.

——— (1999b): "The Economics of Career Concerns, Part II: Application to Missions and Accountability of Government Agencies," *Review of Economic Studies*, 66, 199–217.

Dunning, D. (2011): "The Dunning-Kruger Effect: On Being Ignorant of One's Own Ignorance," in *Advances in Experimental Social Psychology*, ed. by M. Zanna and J. Olson, Academic Press, vol. 44, 247–296.

Dunning, D., K. Johnson, J. Ehrlinger, and J. Kruger (2003): "Why People Fail to Recognize Their Own Incompetence," *Current Directions in Psychological Science*, 12, 83–87.

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

Ehrlinger, J. and D. Dunning (2003): "How Chronic Self-Views Influence (and Potentially Mislead) Estimates of Performance," *Journal of Personality and Social Psychology*, 84, 5–17.

Ehrlinger, J., K. Johnson, M. Banner, D. Dunning, and J. Kruger (2008): "Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent," *Organizational Behavior and Human Decision Processes*, 105, 98–121.

Eliaz, K. and A. Rubinstein (2014): "A model of boundedly rational "neuro" agents," *Transport Policy*, 57, 515–528.

Enke, B. and F. Zimmermann (2019): "Correlation Neglect in Belief Formation," *Review of Economic Studies*, 86, 313–332.

Fehr, D. and S. Huck (2016): "Who Knows It is a Game? On Strategic Awareness and Cognitive Ability," *Experimental Economics*, 19, 713–726.

Fehr, E. and A. Rangel (2011): "Neuroeconomic Foundations of Economic Choice–Recent Advances," *Journal of Economic Perspectives*, 25, 3–30.

Ferraro, P. J. (2010): "Know Thyself: Competence and Self-awareness," *Atlantic Economic Journal*, 38, 183–196.

Fudenberg, D. and L. Rayo (2019): "Training and Effort Dynamics in Apprenticeship," *American Economic Review*, 109, 3780–3812.

Geng, S. (2016): "Decision Time, Consideration Time, And Status Quo Bias," *Economic Inquiry*, 54, 433–449.

Gill, D. and V. Prowse (2016): "Cognitive Ability, Character Skills, and Learning to Play Equilibrium: A Level-k Analysis," *Journal of Political Economy*, 124, 1619–1676.

——— (2017): "Strategic Complexity and the Value of Thinking," *working paper*.

Grabiszewski, K. and A. Horenstein (2019a): "Measuring tree complexity with response times," *working paper*.

——— (2019b): "A Mobile Experiment on Tree Construction," *working paper*.

——— (2019c): "Profiling Players in Dynamic Games: A Mobile Experiment," *working paper*.

Holmström, B. (1999): "Managerial Incentive Problems: A Dynamic Perspective," *Review of Economic Studies*, 66, 169–182.

Jensen, A. R. (2006): *Clocking the Mind: Mental Chronometry and Individual Differences*, Amsterdam, the Netherlands: Elsevier.

Kim, Y.-H., C.-Y. Chiu, and J. Bregant (2015): "Unskilled and Don't Want to Be Aware of It: The Effect of Self-Relevance on the Unskilled and Unaware Phenomenon," *PLoS ONE*, 10, e0130309.

Krajbich, I., B. Oud, and E. Fehr (2014): "Benefits of Neuroeconomic Modeling: New Policy Interventions and Predictors of Preference," *American Economic Review*, 104, 501–506.

Kreps, D. and J. Sobel (1994): "Signalling," in *Handbook of Game Theory with Economic Applications, Volume 2*, ed. by R. Aumann and S. Hart, North Holland, 849–867.

Kruger, J. and D. Dunning (1999): "Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments," *Journal of Personality and Social Psychology*, 77, 1121–1134.

Lohse, J., T. Goeschl, and J. H. Diederich (2017): "Giving is a Question of Time: Response Times and Contributions to an Environmental Public Good," *Environmental and Resource Economics*, 67, 455–477.

Maniadis, Z., F. Tufano, and J. A. List (2014): "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects," *American Economic Review*, 104, 277–290.

Masiliunas, A. (2017): "Overcoming coordination failure in a critical mass game: Strategic motives and action disclosure," *Journal of Economic Behavior & Organization*, 139, 214–251.

Milgrom, P. and J. Roberts (1990): "Rationalizability, Learning, and Equilibrium in Games with Strategic Complementarities," *Econometrica*, 58, 1255–1277.

Milgrom, P. and C. Shannon (1994): "Monotone Comparative Statics," *Econometrica*, 62, 157–180.

Milosavljevic, M., J. Malmaud, A. Huth, C. Koch, and A. Rangel (2010): "The Drift Diffusion Model can account for the accuracy and reaction time of value-based choices under high and low time pressure," *Judgment and Decision Making*, 5, 437–449.

Mirrlees, J. A. (1971): "An Exploration in the Theory of Optimum Income Taxation," *Review of Economic Studies*, 28, 175–208.

Munafò, M. R., B. A. Nosek, D. V. M. Bishop, K. S. Button, C. D. Chambers, N. P. du Sert, U. Simonsohn, E.-J. Wagenmakers, J. J. Ware, and J. P. Ioannidis (2017): "A manifesto for reproducible science," *Nature Human Behaviour*, 1, 0021.

Ofek, E., M. Yildiz, and E. Haruvy (2007): "The Impact of Prior Decisions on Subsequent Valuations in a Costly Contemplation Model," *Management Science*, 53, 1217–1233.

Open Science Collaboration (2014): "The Reproducibility Project: A model of large-scale collaboration for empirical research on reproducibility," in *Implementing reproducible computational research*, ed. by V. Stodden, F. Leisch, and R. D. Peng, New York, NY: Taylor & Francis.

——— (2015): "Estimating the reproducibility of psychological science," *Science*, 349, aac4716.

Palacios-Huerta, I. and O. Volij (2009): "Field Centipedes," *American Economic Review*, 9, 1619–1635.

Quah, J. K.-H. and B. Strulovici (2009): "Aggregating the Single Crossing Property," *Econometrica*, 80, 2333–2348.

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

Robinson, P. M. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

Rubinstein, A. (2006): "Dilemmas of an Economic Theorist," *Econometrica*, 74, 865–883.

——— (2007): "Instinctive and Cognitive Reasoning: A Study of Response Times," *Economic Journal*, 117, 1243–1259.

——— (2013): "Response Time and Decision Making: An Experimental Study," *Judgment and Decision Making*, 8, 540–551.

——— (2016): "A Typology of Players: Between Instinctive and Contemplative," *Quarterly Journal of Economics*, 131, 859–890.

Sanchez, C. and D. Dunning (2018): "Overconfidence Among Beginners: Is a Little Learning a Dangerous Thing?" *Journal of Personality and Social Psychology*, 114, 10–28.

Schlösser, T., D. Dunning, K. L. Johnson, and J. Kruger (2013): "How unaware are the unskilled? Empirical tests of the "signal extraction" counterexplanation for the Dunning-Kruger effect in self-evaluation of performance," *Journal of Economic Psychology*, 39, 85–100.

Spence, M. (1973): "Job Market Signalling," *Quarterly Journal of Economics*, 87, 355–374.

Spiliopoulos, L. and A. Ortmann (2017): "The BCD of Response Time Analysis in Experimental Economics," *Experimental Economics*, 47, 1–55.

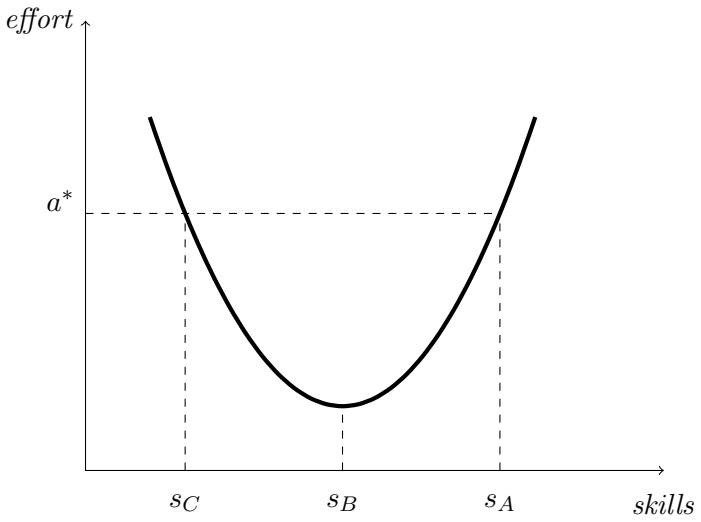
Staub, S. and R. Kaynak (2014): "Is an Unskilled Really Unaware of it?" *Procedia – Social and Behavioral Sciences*, 150, 899–907.

Verardi, V. and N. Debarsy (2012): "Robinson's square root of N consistent semiparametric regression estimator in Stata," *Stata Journal*, 12, 726–735.

Vives, X. (1990): "Nash Equilibrium with Strategic Complementarities," *Journal of Mathematical Economics*, 19, 305–321.

Woodford, M. (2014): "Stochastic Choice: An Optimizing Neuroeconomic Model," *American Economic Review*, 104, 495–500.

Figure 1: Main result.



Notes. The relationship between the optimal effort and skills has a U-shape. Although she has higher skills, Ann (s_A) chooses the same level of effort a^* as Chris (s_C). That is, people with different skills do not necessarily choose different levels of effort. In addition, higher skills can imply an increase in effort (from Bob with skills s_B to Ann with skills s_A) or a decrease in effort (from Chris with skills s_C to Bob with skills s_B).

Figure 2: A screenshot from *Blues and Reds* with an example of puzzle.

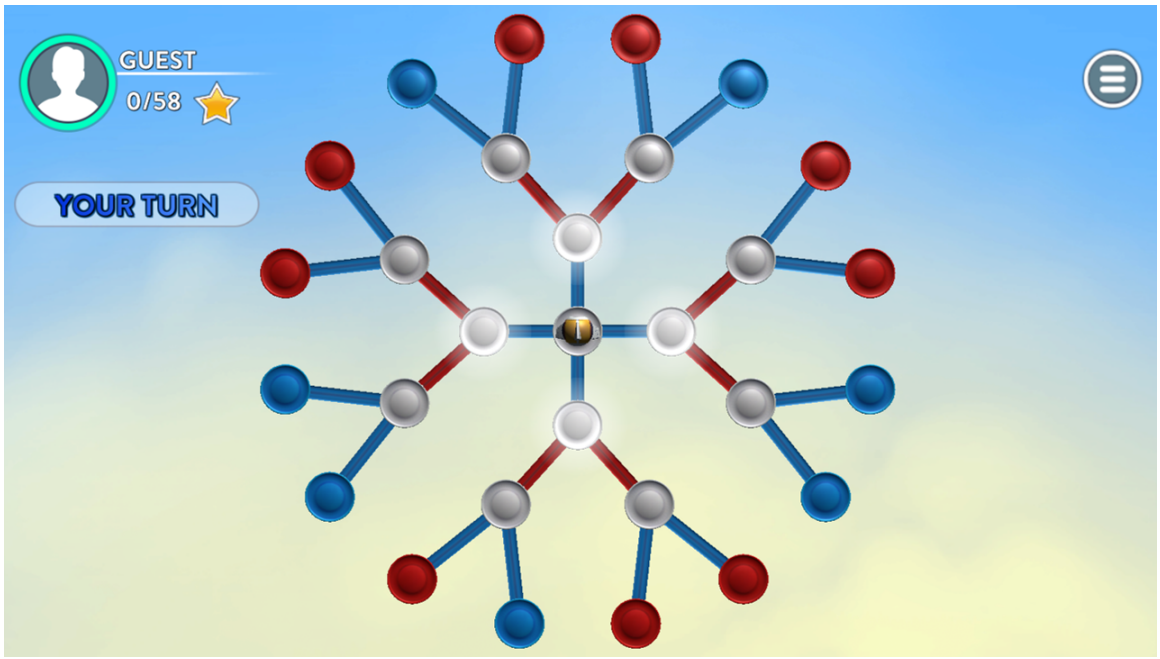


Figure 3: Estimation results for puzzle 2.2.2

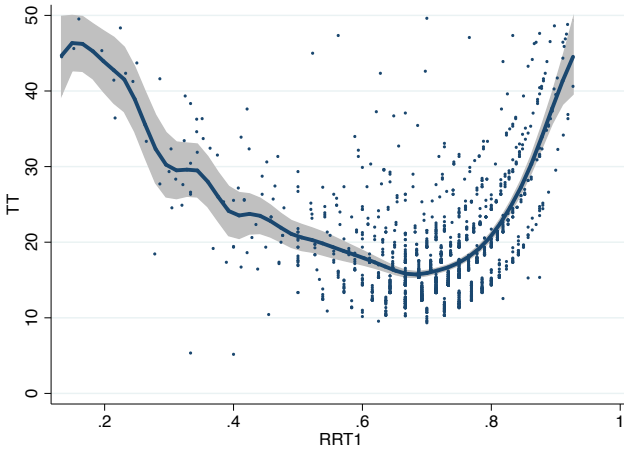


Figure 4: Estimation results for puzzle 2.2.3

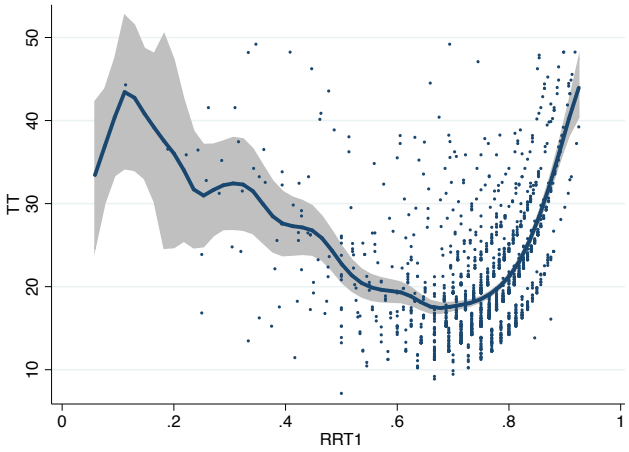


Figure 5: Estimation results for puzzle 2.3.2

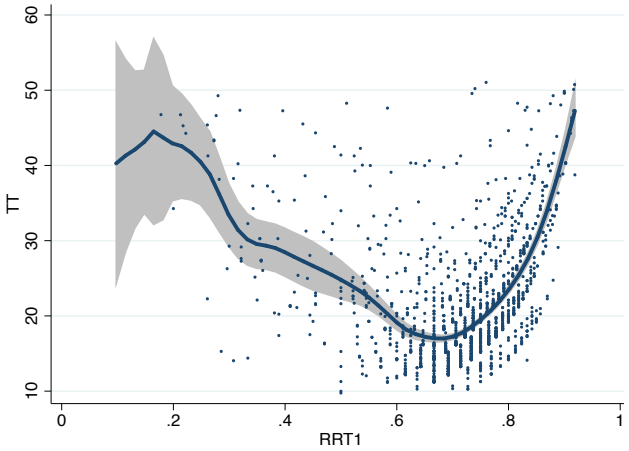


Figure 6: Estimation results for puzzle 2.3.3

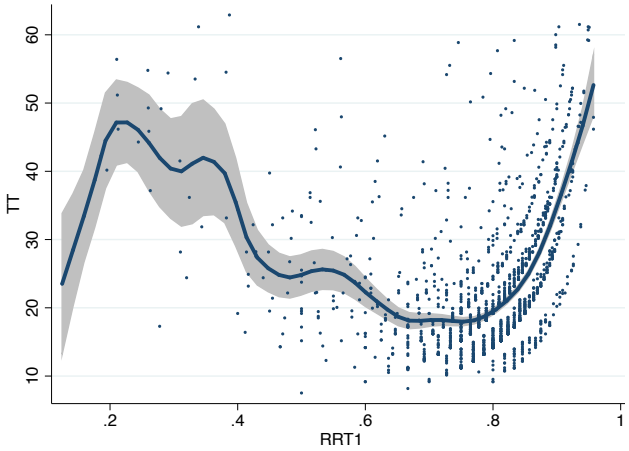


Figure 7: Estimation results for puzzle 3.2.2

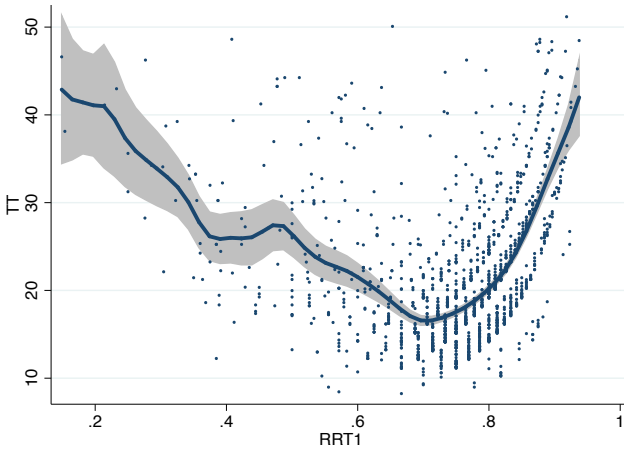


Figure 8: Estimation results for puzzle 3.2.3

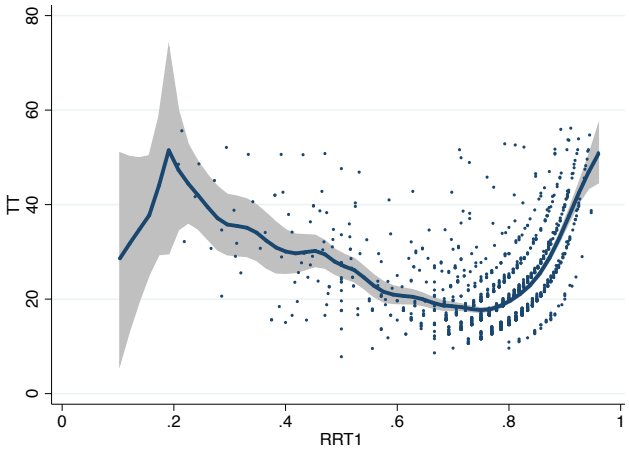


Figure 9: Estimation results for puzzle 3.3.2

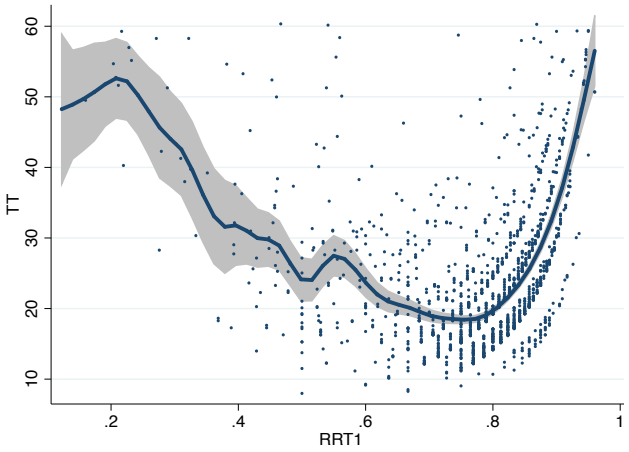


Figure 10: Estimation results for puzzle 3.3.3

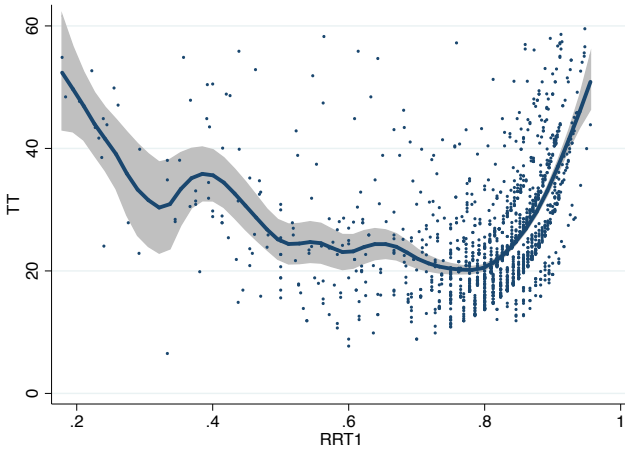


Figure 11: Estimation results for puzzle 4.2.2

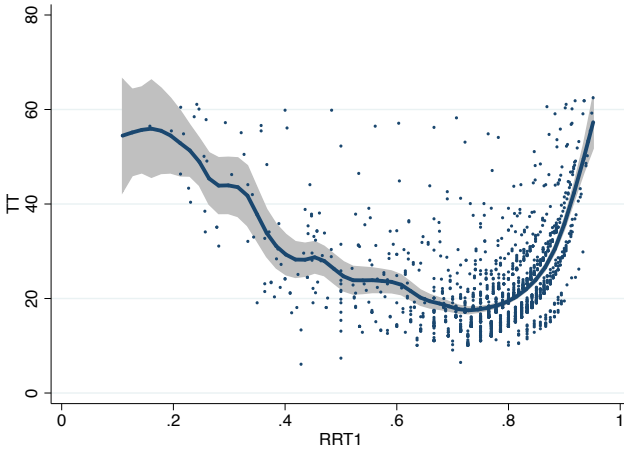


Figure 12: Estimation results for puzzle 2.2.2.2

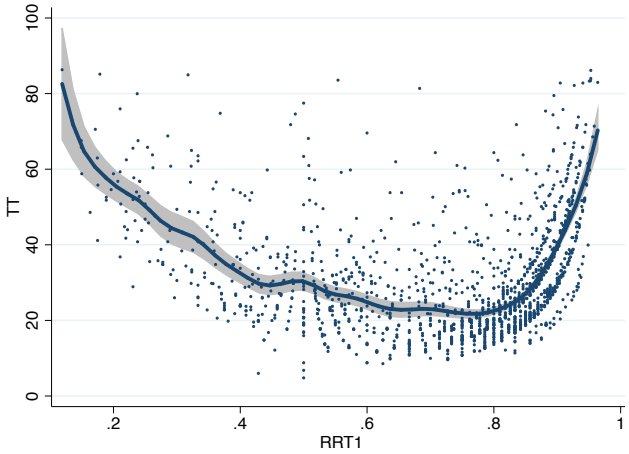


Figure 13: Estimation results for puzzle 3.2.2.2

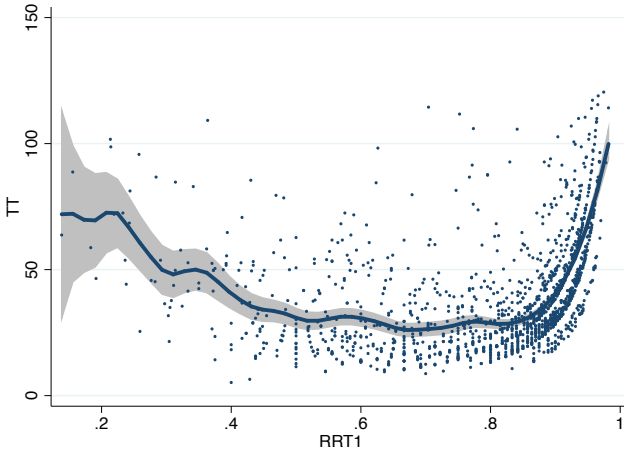


Figure 14: Estimation results for puzzle 4.2.2.2

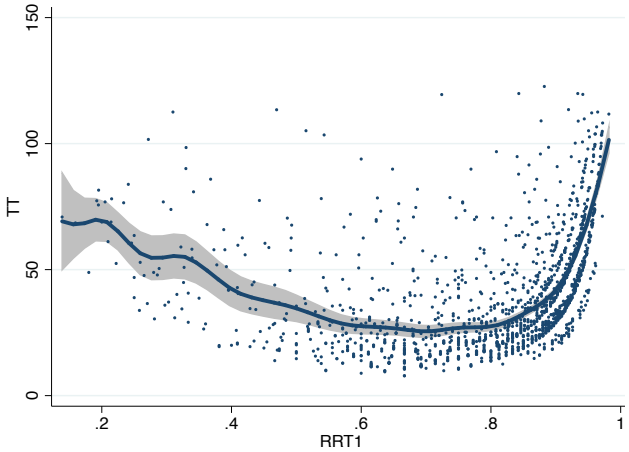


Figure 15: Estimation results for puzzle 2.3.2.2

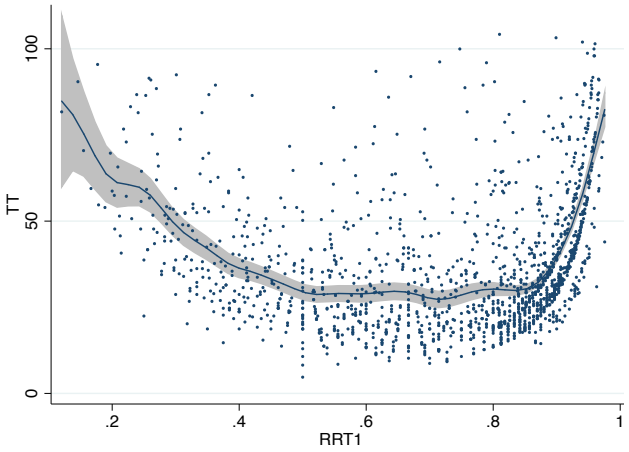


Figure 16: Estimation results for puzzle 2.4.2.2

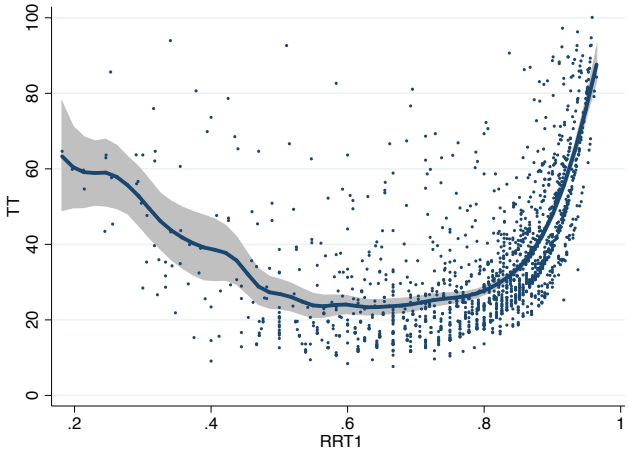


Figure 17: Estimation results for puzzle 2.2.3.2

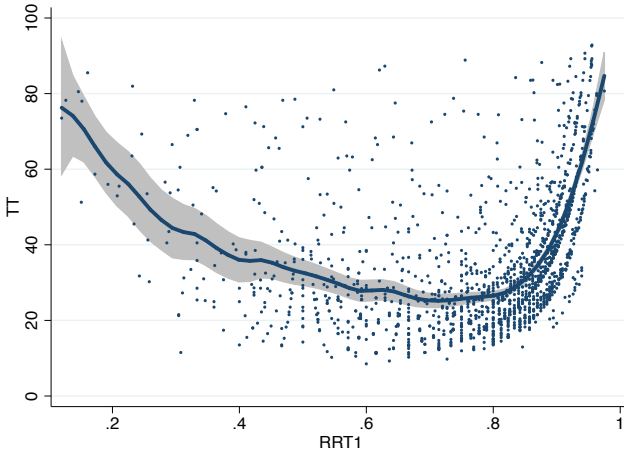


Figure 18: Estimation results for puzzle 2.2.4.2

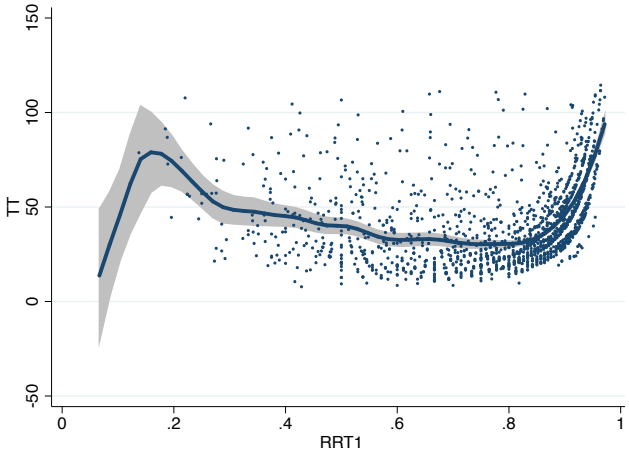


Figure 19: Estimation results for puzzle 2.2.2.3

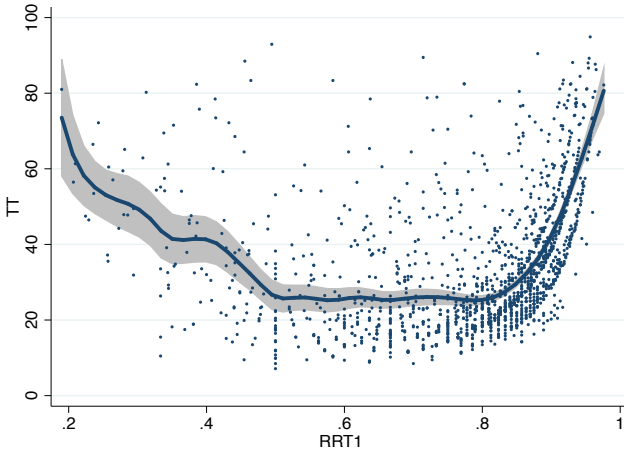


Figure 20: Estimation results for puzzle 2.2.2.4

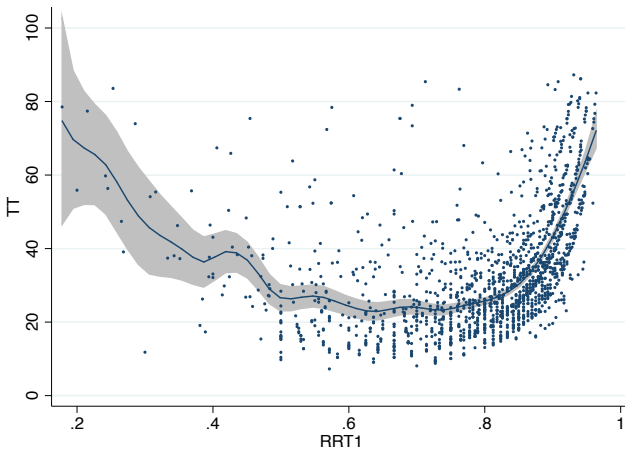


Figure 21: Estimation results for puzzle 2.2.2.2.2

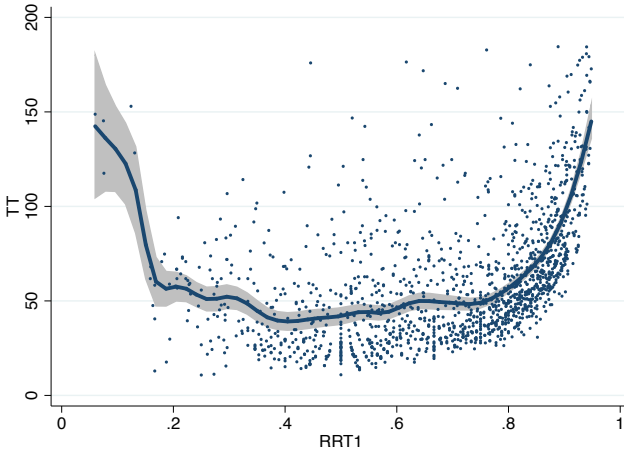


Figure 22: Estimation results for puzzle 3.2.2.2.2

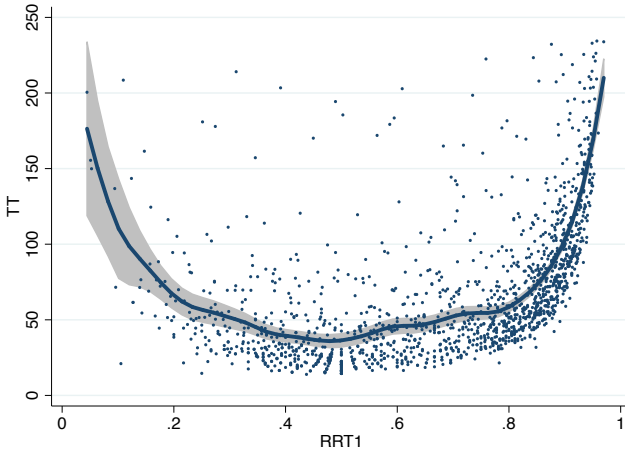


Figure 23: Estimation results for puzzle 4.2.2.2.2

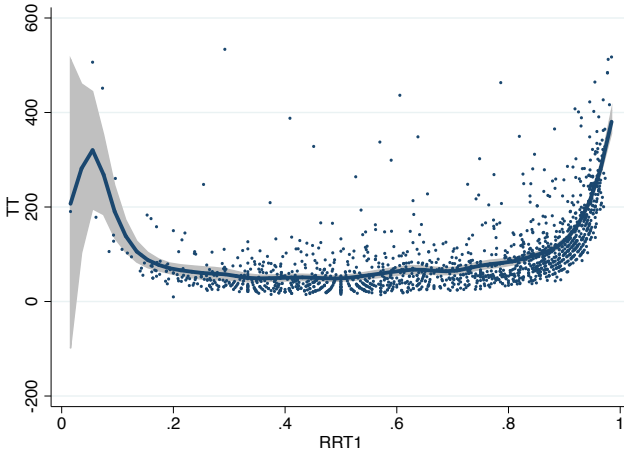


Figure 24: Estimation results for puzzle 2.2.2.2.2.2

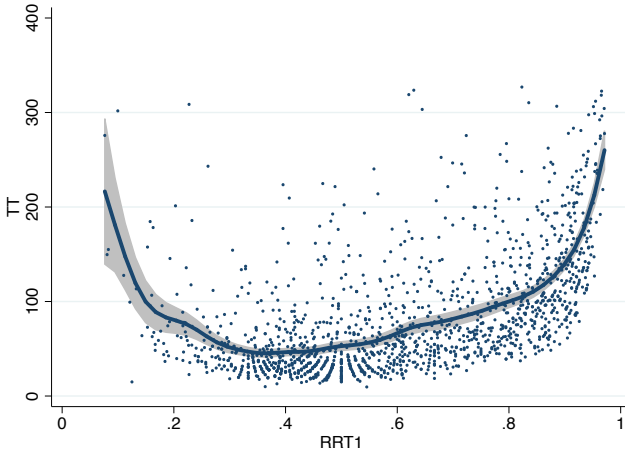


Table 1: List of puzzles.

3 rounds	4 rounds	5 rounds	6 rounds
2.2.2	2.2.2.2	2.2.2.2.2	2.2.2.2.2.2
2.2.3	3.2.2.2	3.2.2.2.2	
2.3.2	4.2.2.2	4.2.2.2.2	
2.3.3	2.3.2.2		
3.2.2	2.4.2.2		
3.2.3	2.2.3.2		
3.3.2	2.2.4.2		
3.3.3	2.2.2.3		
4.2.2	2.2.2.4		

Table 2: Top 20 most popular countries in the sample.

country	% sample
Mexico	13.84%
India	12.25%
Argentina	6.65%
Brazil	6.21%
Colombia	4.86%
USA	4.23%
Spain	4.04%
Chile	3.76%
Poland	3.41%
Thailand	2.77%
Peru	1.71%
Ukraine	1.43%
Iraq	1.27%
China	1.24%
Russia	1.22%
Ecuador	1.18%
Egypt	1.13%
Italy	1.08%
Turkey	1.08%
Iran	1.01%

Table 3: Effort and skills: summary statistics.

puzzle	N	effort (TT)				skills (RRT1)			
		Mean	Std. Dev.	Min	Max	Mean	Std. Dev.	Min	Max
2.2.2	1,638	19.05	7.17	5	50	0.71	0.11	0.13	0.93
2.2.3	1,729	21.07	7.22	6	49	0.74	0.10	0.06	0.93
2.3.2	1,630	21.04	7.86	8	50	0.71	0.11	0.10	0.92
2.3.3	1,637	22.98	9.82	8	62	0.78	0.11	0.13	0.96
3.2.2	1,666	21.03	7.76	7	49	0.75	0.11	0.15	0.94
3.2.3	1,628	21.94	8.44	8	56	0.77	0.11	0.10	0.96
3.3.2	1,647	23.18	9.56	8	61	0.77	0.12	0.12	0.96
3.3.3	1,638	25.64	9.63	5	58	0.80	0.12	0.18	0.96
4.2.2	1,717	22.68	9.79	6	62	0.77	0.12	0.11	0.95
2.2.2.2	1,660	30.10	14.60	6	85	0.72	0.19	0.12	0.96
3.2.2.2	1,575	38.44	20.65	5	118	0.79	0.16	0.14	0.98
4.2.2.2	1,614	38.63	21.69	8	121	0.79	0.16	0.14	0.98
2.3.2.2	1,606	35.98	18.44	6	101	0.73	0.20	0.12	0.98
2.4.2.2	1,641	35.84	17.94	9	96	0.78	0.14	0.18	0.96
2.2.3.2	1,674	34.55	16.91	9	90	0.77	0.16	0.12	0.98
2.2.4.2	1,673	40.44	20.88	7	112	0.77	0.17	0.07	0.97
2.2.2.3	1,610	33.24	17.01	6	93	0.78	0.15	0.19	0.98
2.2.2.4	1,602	32.57	15.96	7	87	0.79	0.13	0.18	0.96
2.2.2.2.2	1,545	59.43	32.15	10	184	0.69	0.19	0.06	0.95
3.2.2.2.2	1,550	66.26	42.55	12	235	0.68	0.22	0.04	0.97
4.2.2.2.2	1,566	95.08	78.10	10	534	0.67	0.23	0.02	0.98
2.2.2.2.2.2	1,580	81.55	60.14	11	328	0.61	0.21	0.08	0.97

Notes. N denotes the number of subjects who played a given puzzle. For effort (TT) and skills (RRT1), this table provides the mean, standard deviation, and minimal and maximal values.

Table 4: Percentage of winners in RRT1-terciles.

puzzle	L	M	H
2.2.2	84%	98%	99%
2.2.3	84%	97%	99%
2.3.2	82%	96%	99%
2.3.3	82%	99%	99%
3.2.2	75%	98%	98%
3.2.3	76%	98%	99%
3.3.2	74%	98%	98%
3.3.3	71%	98%	99%
4.2.2	72%	98%	98%
2.2.2.2	18%	85%	98%
3.2.2.2	23%	89%	97%
4.2.2.2	28%	86%	97%
2.3.2.2	6%	65%	98%
2.4.2.2	52%	92%	99%
2.2.3.2	39%	93%	99%
2.2.4.2	30%	89%	98%
2.2.2.3	44%	94%	98%
2.2.2.4	58%	93%	98%
2.2.2.2.2	40%	80%	97%
3.2.2.2.2	30%	76%	96%
4.2.2.2.2	13%	42%	89%
2.2.2.2.2.2	21%	37%	83%

Notes. In each puzzle, observations were divided into RRT1-terciles (L, M, and H). Next, for each RRT1-tercile, the percentage of winners was computed.

Table 5: Percentage of winners in RRT1-quintiles conditional on TT-quintiles.

		RRT1 conditional on TT							RRT1 conditional on TT				
	TT	1	2	3	4	5		TT	1	2	3	4	5
2.2.2	1	93%	96%	99%	100%	100%	2.3.2.2	1	10%	22%	71%	95%	98%
	2	93%	98%	100%	99%	100%		2	4%	9%	71%	98%	100%
	3	78%	98%	99%	99%	100%		3	5%	8%	74%	97%	100%
	4	76%	97%	97%	99%	100%		4	3%	4%	73%	98%	100%
	5	48%	75%	98%	100%	98%		5	2%	10%	56%	100%	100%
2.2.2.2	1	44%	79%	96%	97%	97%	3.2.2	1	70%	97%	98%	100%	100%
	2	15%	57%	89%	96%	98%		2	86%	99%	98%	100%	100%
	3	7%	8%	78%	99%	100%		3	55%	97%	97%	99%	98%
	4	3%	21%	89%	100%	99%		4	62%	92%	100%	100%	100%
	5	5%	16%	70%	98%	97%		5	38%	82%	94%	97%	98%
2.2.2.2.2	1	46%	65%	76%	85%	100%	2.2.2.3	1	35%	75%	93%	98%	100%
	2	24%	60%	94%	98%	100%		2	35%	79%	98%	100%	100%
	3	16%	40%	93%	98%	100%		3	30%	87%	97%	100%	100%
	4	16%	53%	97%	100%	100%		4	17%	78%	97%	100%	100%
	5	16%	48%	84%	98%	100%		5	17%	51%	95%	97%	97%
3.3.2	1	65%	93%	98%	100%	100%	2.2.3	1	80%	98%	100%	99%	100%
	2	75%	95%	100%	100%	100%		2	85%	100%	98%	100%	100%
	3	73%	98%	98%	100%	99%		3	81%	99%	99%	100%	100%
	4	60%	97%	98%	97%	97%		4	76%	94%	95%	100%	100%
	5	36%	70%	91%	97%	98%		5	51%	90%	100%	96%	99%
2.2.2.4	1	43%	71%	80%	94%	97%	2.2.3.2	1	29%	67%	89%	97%	100%
	2	52%	88%	91%	97%	100%		2	25%	74%	97%	100%	100%
	3	38%	92%	97%	96%	98%		3	22%	69%	97%	99%	100%
	4	39%	84%	97%	100%	98%		4	23%	85%	93%	100%	100%
	5	32%	92%	97%	97%	98%		5	9%	58%	97%	100%	98%
3.3.3	1	58%	98%	100%	100%	100%	2.2.2.2.2.2	1	28%	22%	24%	31%	48%
	2	68%	97%	98%	99%	100%		2	26%	19%	39%	45%	63%
	3	63%	100%	100%	100%	100%		3	16%	27%	32%	59%	79%
	4	52%	91%	99%	99%	100%		4	15%	25%	66%	89%	98%
	5	33%	85%	100%	99%	97%		5	14%	46%	73%	94%	97%
4.2.2.2	1	14%	41%	75%	89%	98%	2.4.2.2	1	48%	62%	86%	86%	91%
	2	6%	55%	93%	96%	100%		2	36%	77%	93%	99%	98%
	3	10%	58%	93%	94%	99%		3	37%	93%	95%	100%	97%
	4	17%	80%	98%	100%	98%		4	27%	93%	97%	100%	100%
	5	6%	43%	95%	95%	100%		5	22%	85%	100%	100%	99%

Continued on next page

Table 5 – continued from previous page

		RRT1 conditional on TT							RRT1 conditional on TT				
	TT	1	2	3	4	5		TT	1	2	3	4	5
2.3.3	1	92%	99%	100%	100%	100%	3.2.2.2.2	1	19%	49%	61%	73%	89%
	2	79%	100%	100%	100%	100%		2	10%	28%	75%	97%	97%
	3	84%	99%	100%	100%	98%		3	8%	38%	80%	100%	98%
	4	54%	100%	98%	100%	100%		4	15%	62%	93%	95%	100%
	5	40%	88%	100%	98%	100%		5	21%	76%	94%	98%	97%
3.2.3	1	83%	100%	99%	100%	100%	4.2.2	1	77%	100%	99%	100%	100%
	2	67%	100%	98%	96%	100%		2	75%	98%	100%	100%	98%
	3	77%	99%	97%	100%	100%		3	60%	94%	94%	99%	100%
	4	63%	95%	100%	100%	100%		4	53%	93%	98%	100%	100%
	5	46%	66%	97%	98%	100%		5	31%	70%	93%	99%	100%
3.2.2.2	1	8%	35%	69%	95%	97%	2.2.4.2	1	27%	69%	86%	99%	100%
	2	8%	46%	86%	100%	100%		2	23%	62%	96%	99%	100%
	3	12%	79%	97%	98%	99%		3	20%	68%	96%	100%	100%
	4	8%	67%	93%	97%	100%		4	12%	55%	97%	99%	100%
	5	6%	54%	94%	97%	98%		5	7%	22%	86%	97%	100%
2.3.2	1	85%	100%	98%	100%	100%	4.2.2.2.2	1	14%	20%	29%	47%	63%
	2	75%	100%	100%	100%	100%		2	3%	11%	21%	57%	87%
	3	83%	96%	100%	95%	100%		3	2%	5%	19%	84%	92%
	4	73%	89%	98%	100%	100%		4	0%	32%	81%	97%	98%
	5	37%	71%	97%	97%	100%		5	11%	57%	84%	90%	95%

Notes. In each puzzle, observations were divided into TT-quintiles (rows: from the first TT-quintile denoted as 1 to the fifth TT-quintile denoted as 5). Next, for each TT-quintile, observations were divided into RRT1-quintiles (columns: from the first RRT1-quintile denoted as 1 to the fifth quintile denoted as 5). Fina1y, for each conditional RRT1-quintile, the percentage of winners was computed.

Table 6: Logit estimations.

puzzle	N	RRT1	Pseudo R^2	puzzle	N	RRT1	Pseudo R^2
2.2.2	1,638	10.786***	0.266	3.2.2	1,666	13.103***	0.350
		(0.859)				(1.103)	
2.3.2	1,630	11.329***	0.283	2.2.3	1,729	11.926***	0.306
		(0.925)				(0.991)	
3.3.2	1,647	10.613***	0.278	2.3.3	1,637	12.919***	0.376
		(0.885)				(1.142)	
3.2.3	1,628	12.584***	0.343	3.3.3	1,638	13.287***	0.398
		(1.011)				(1.059)	
4.2.2	1,717	13.204***	0.401	2.2.2.2	1,660	14.261***	0.507
		(0.887)				(0.790)	
3.2.2.2	1,575	15.861***	0.489	2.3.2.2	1,606	17.002***	0.570
		(0.940)				(1.037)	
2.2.3.2	1,674	13.426***	0.441	2.2.2.3	1,610	12.090***	0.389
		(0.851)				(0.762)	
4.2.2.2	1,614	15.683***	0.446	2.4.2.2	1,641	11.757***	0.347
		(0.847)				(0.706)	
2.2.4.2	1,673	12.983***	0.442	2.2.2.4	1,602	11.325***	0.304
		(0.707)				(0.771)	
2.2.2.2.2	1,545	7.832***	0.283	3.2.2.2.2	1,550	7.982***	0.347
		(0.421)				(0.381)	
4.2.2.2.2	1,566	8.080***	0.346	2.2.2.2.2.2	1,580	5.980***	0.217
		(0.407)				(0.317)	

Notes. N denotes the number of observations. Robust standard errors are provided in parentheses. Stars denote the significance level (*** 1%, ** 5%, and * 10%).

Table 7: Results from regressions.

		Model RRT1		Model TT	
puzzle	N	*Seq*	Adj R^2	*Seq*	Adj R^2
2.2.2	1,535	-0.00096***	0.006	-0.09835***	0.011
		(0.00034)		(0.02345)	
2.2.3	1,618	-0.00118***	0.010	-0.12995***	0.018
		(0.00031)		(0.02350)	
2.3.2	1,500	0.00028	0.001	-0.1006***	0.011
		(0.00312)		(0.02514)	
2.3.3	1,527	0.00004	0.000	-0.07907**	0.002
		(0.00030)		(0.03418)	
3.2.2	1,513	0.00077**	0.004	-0.13683***	0.018
		(0.00031)		(0.02560)	
3.2.3	1,483	-0.00039	0.001	-0.06037**	0.003
		(0.00033)		(0.02910)	
3.3.2	1,480	-0.00015	0.000	-0.09406***	0.006
		(0.00034)		(0.03058)	
3.3.3	1,469	-0.00025	0.001	-0.17014***	0.017
		(0.00028)		(0.03376)	
4.2.2	1,536	0.00027	0.001	-0.10035***	0.007
		(0.00028)		(0.02960)	
2.2.2.2	1,106	-0.00013	0.000	-0.25926***	0.018
		(0.00046)		(0.05730)	
3.2.2.2	1,098	0.0004	0.001	-0.30507***	0.012
		(0.00031)		(0.07531)	
4.2.2.2	1,136	0.0012***	0.011	-0.11723	0.004
		(0.00033)		(0.08524)	
2.3.2.2	907	0.00046	0.002	-0.3539***	0.021
		(0.00038)		(0.75574)	
2.4.2.2	1,323	0.00129***	0.009	-0.0999	0.002
		(0.00038)		(0.06604)	
2.2.3.2	1,292	0.00069*	0.003	-0.25675***	0.013
		(0.00040)		(0.05980)	
2.2.4.2	1,217	0.00203***	0.023	-0.08999	0.001
		(0.00034)		(0.07523)	
2.2.2.3	1,273	0.00004	0.000	-0.34115***	0.023
		(0.00041)		(0.05788)	
2.2.2.4	1,329	0.00027	0.000	-0.3414***	0.026
		(0.00038)		(0.05071)	
2.2.2.2.2	1,118	0.003***	0.023	0.1629	0.002
		(0.00059)		(0.13040)	
3.2.2.2.2	1,039	0.00351***	0.029	0.49873***	0.008
		(0.00063)		(0.17000)	
4.2.2.2.2	751	0.00436***	0.049	1.203***	0.012
		(0.00066)		(0.37900)	
2.2.2.2.2.2	741	0.00605***	0.059	1.09558***	0.018
		(0.00086)		(0.29890)	

Notes. N denotes the number of observations. Robust standard errors are provided in parentheses. Stars denote the significance level (*** 1%, ** 5%, and * 10%).