ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Hanaki, Nobuyuki; Inukai, Keigo; Masuda, Takehito; Shimodaira, Yuta

Working Paper

Participants' characteristics at ISER-lab in 2020

ISER Discussion Paper, No. 1141

Provided in Cooperation with:
The Institute of Social and Economic Research (ISER), Osaka University

Suggested Citation: Hanaki, Nobuyuki; Inukai, Keigo; Masuda, Takehito; Shimodaira, Yuta (2021) : Participants' characteristics at ISER-lab in 2020, ISER Discussion Paper, No. 1141, Osaka University, Institute of Social and Economic Research (ISER), Osaka

This Version is available at:
https://hdl.handle.net/10419/248601

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Discussion Paper No. 1141

ISSN (Print) 0473-453X
ISSN (Online) 2435-0982

PARTICIPANTS' CHARACTERISTICS AT ISER-LAB IN 2020

Nobuyuki Hanaki
Keigo Inukai
Takehito Masuda
Yuta Shimodaira

September 2021

The Institute of Social and Economic Research
Osaka University
6-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan

Participants' Characteristics at ISER-Lab in 2020*

Nobuyuki Hanaki†, Keigo Inukai‡, Takehito Masuda§, Yuta Shimodaira¶

September 8, 2021

Abstract

We summarize the experimentally measured characteristics of the registered participants of the experiments conducted at the Institute of Social and Economic Research, Osaka University. Measured characteristics include fluid intelligence, risk preference (risk aversion, prudence, and temperance), social value orientation, theory of mind, personality (Big Five and Grit), ability to backward induct, as well as their general trust. We discuss reliability of these measures and correlation among them.

Keywords: Cognitive Ability, Personality, Theory of Mind, Higher Order Risk Preferences

JEL codes: C91, D91

1 Introduction

Increasingly more research documents the relationships among such measured individual characteristics as sex, cognitive ability, risk and time preference, and behavior in various strategic situations.

The literature on the sex difference is large. For example, it has been demonstrated that females tend to be (a) more risk averse (Booth and Nolen, 2012; Charness and Gneezy, 2012; Booth et al., 2014; Filippin, 2016), (b) less competitive

*This work was supported by Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research, Nos. 18K19954, and 20H05631, and the Joint Usage/Research Center, Institute of Social and Economic Research, Osaka University. The experiments reported in this paper have been approved by the Institutional Review Board of Institute of Social and Economic Research, Osaka University.

†Institute of Social and Economic Research, Osaka University. E-mail: nobuyuki.hanaki@iser.osaka-u.ac.jp

‡Faculty of Economics, Meiji Gakuin University. E-mail: inukai@eco.meijigakuin.ac.jp

§Faculty of Economics and Law, Shinshu University. E-mail: tmasuda@shinshu-u.ac.jp

¶Graduate School of Economics, Osaka University. E-mail: u064674i@ecs.osaka-u.ac.jp

(Niederle and Vesterlund, 2007; Andersen et al., 2013; List and Gneezy, 2014; Kuhn and Villeval, 2015), (c) more pro-social (Andreoni and Vesterlund, 2001; Kamas and Preston, 2015; Balafoutas et al., 2012) and to believe more in reciprocity (Chaudhuri and Gangadharan, 2003; Dittrich, 2015), and (d) considered more trustworthy (Alesina and La Ferrara, 2002; Buchan et al., 2008) than males.

In terms of cognitive ability, Dohmen et al. (2010) and Benjamin et al. (2013) report those with higher cognitive ability tends to take more risk and also patient. Burnham et al. (2009), Gill and Prowse (2016) and Carpenter et al. (2013) study the relationship between cognitive ability and observed behavior in p-beauty contest games and report those with higher cognitive ability submit numbers closer to the Nash equilibrium. Basteck and Mantovani (2018) report those with lower cognitive ability receive lower payoff in school choice experiments. Hanaki et al. (2016) and Akiyama et al. (2017) report that cognitive ability is positively related to the how much participants consider the behavioral uncertainty of other players playing a simply coordination game and an asset market game, respectively.

Many of them rely on ex-post correlational analyses between the individual characteristics and observed behavior (for example, Corgnet et al., 2018). However, increasingly more experimental research involve selectively recruiting participants with certain characteristics, such as cognitive ability class, to better identify the relationships between the observed behavior and the characteristics in question. Some of these latter studies measures the targeted characteristics at the beginning of the experiment, and then separate participants into groups based on the outcome (Gill and Prowse, 2016; Proto et al., 2019), others measures various characteristics in a separate experiment, and then recruit participants back based on the measured characteristics (see, among others, Bosch-Rosa et al., 2018)

If systematic relationships among the individual characteristics and observed behavior in experiments exist as these studies suggest, it is of great importance to measure and understand these characteristics of participants when, for example,

trying to replicate the experiment conducted elsewhere (see, Camerer et al., 2016, for an importance of such replication exercise), or comparing and interpreting the results of the same experiment conducted in various countries, such as those done by Herrmann et al. (2008) and Gächter et al. (2010).

Acknowledging the need, researchers have started to measure and to compare individual characteristics in the various pools of participants. For example, Snowberg and Yariv (2021) compare measured characteristics of participants from California Institute of Technology (Caltech), US-Based Amazon Mechanical Turk (MTurk) users, and the sample from the representative US population. The results show that while the average scores are different across these subjects pools, the correlations among measures are similar suggesting the robustness of the relationships between individual characteristics and observed behavior. Indeed, there are several attempt to re-examine the previously reported correlations among individual characteristics using a large scale survey in a national representative sample (Chapman et al., 2018; Dean and Ortoleva, 2019).

We have started to systematically measure various individual characteristics of those who have registered to the participants database of the Institute of Social and Economic Research, Osaka University managed by ORSEE (Greiner, 2015). This paper summarizes the measured individual characteristics as well as correlation among them during the 2020-2021 academic year.

Rest of the paper is organized as follows. Section 2 provides the description of measures used. Section 3 describes the experimental procedure, followed by the results in Section 4.

2 Measure

2.1 Cognitive Ability

2.1.1 Cognitive Reflection Test

The Cognitive Reflection Test (CRT) is a simple task to measure cognitive ability. The questions that comprise the CRT are questions that can be incorrectly answered if responded intuitively. Therefore, the cognitive ability measured by the CRT is the ability to control one's intuitive response and derive the correct answer by deliberation.

The three questions of CRT proposed by Frederick (2005) have already become well known. Therefore, we used three of the questions from Finucane and Gullion (2010) instead of original questions. Each of these three questions has the same logical structure. The number of correct answers to these three questions can be used as the score for the conventional CRT.

We also selected and added three of the questions proposed by Toplak et al. (2014). We used the number of correct answers in the six questions, together with the three mentioned above, as the CRT score (`V_crt6`).

2.1.2 Ability of Backward Induction

We used a task called *the Race Game* or *the Game of 21*, which was used by Gneezy et al. (2010) and Dufwenberg et al. (2010) in their research on learning in experimental games, to measure the participants' ability to perform backward induction.

The game we used is an extensive form perfect information game. The number of players in this game is two, and the participants play against the computer. There is a state variable, a number between 0 and 21, and the initial state is 0. We require the players to add a number, either 1, 2, or 3, to the state variable. The participant is a first mover, and the computer is a follower. They take turn to choose a number. The winner is who gets 21 first. The computer chose an action randomly, and we

explain that to the participants. In this game, a state variable is a winning number for the player if and only if it is 1, 5, 9, 13, 17, and 21. The winning strategy is to move to the one among those six numbers when possible. Note that, since the computer's behavior is random, there can be cases where the participant cannot choose a winning number.

We observed whether the participants have correctly chosen the winning number at the first chance to select each number. We define a measure of backward induction ability `V_BI_Gneezy` as the number of times a participant successfully chooses a winning number divided by the number of chances they face. Thus `V_BI_Gneezy` is a real number between 0 and 1.

2.1.3 International Cognitive Ability Resource Test

The International Cognitive Ability Resource (ICAR) test is a measurement of cognitive ability proposed by Condon and Revelle (2014), and maintained by the ICAR Team (see, their website, `https://icar-project.com`, for more information). The ICAR test has been developed as a public domain tool, and researchers can choose which tasks to include as their experimental tasks.

We used the Three-dimensional Rotation measure and the Matrix Reasoning measure among those provided by the ICAR. The Three-dimensional Rotation items present participants with cube renderings and ask participants to identify which of the response choices is a possible rotation of the target stimuli. The Matrix Reasoning items contain stimuli that are similar to those used in Raven's Progressive Matrices (Raven, 2000). The stimuli are 3×3 arrays of geometric shapes with one of the nine shapes missing. Participants are instructed to identify which of the six geometric shapes presented as response choices will best complete the stimuli.

We selected five items for the Three-dimensional Rotation measure and five items for the Matrix Reasoning measure in the following way for simplified implementation. First, we conducted a preliminary experiment at both Osaka University and Meiji

Gakuin University. For our preliminary experiment, we selected 11 of the 24 Three-dimensional Rotation items in advance, balancing the difficulty level based on the percentage of correct answers reported by Condon and Revelle (2014). For the Matrix Reasoning measure, we used all 11 items. Then, we obtained the distribution of scores in a set of 22 tasks from the preliminary experiments. Finally, we selected questions so that the distribution of scores is approximated to the distribution of the 22-question version with respect to Kolmogorov-Smirnov statistics.

We asked the participants to answer five questions in three minutes for each of the two measures: the Three-dimensional Rotation and the Matrix Reasoning. We then define the ICAR measure `V_ICARscore` as the score of the total ten items of the two measures.

2.2 Risk Attitude

2.2.1 Risk Aversion and Attitude Over Higher Order Risk

We used the same method as Masuda and Lee (2019) to measure risk aversion, prudence, and temperance. We ask five questions for risk aversion, ten questions for prudence, and five questions for temperance, asking which of the two lotteries they would choose. The elicitation task was originally proposed by Noussair et al. (2014). Prudence and temperance are known as higher-order risk attitudes (Eeckhoudt and Schlesinger, 2006).

To measure risk aversion, we asked participants to choose between a risky lottery in which she gets 650 JPY with 50% chance and 50 JPY with 50% chance, and, on the other hand, a sure payment of X JPY where X takes 200, 250, 300, 350, and 400. Since we have only five values of X presented in an increasing order, we approximate the certain equivalent by the (middle of) switching point to the risky lottery to the sure payment. Assuming that individuals consistently choose the risky option only when X is less than their certainty equivalent, the fewer times they choose the risky option, the more risk averse they are. We define the risk aversion measure `V_RAscore`

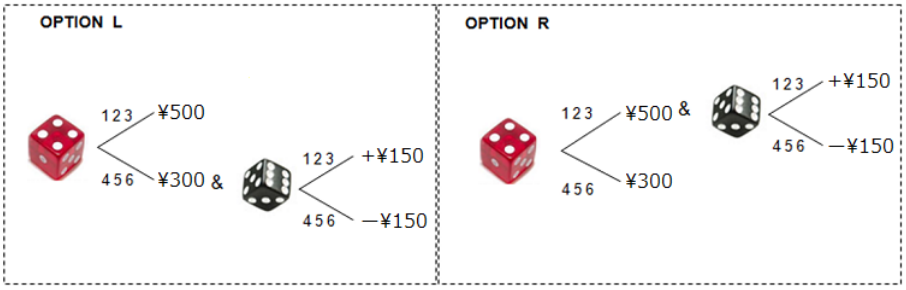


Figure 1: Graphical presentation of prudence tasks.

as the number of safe option among the five questions.

To measure prudence, we present the options L and R illustrated in Figure 1. Assume that realizations x and y with $x > y > 0$, as well as $+z$ and $-z$, are equally likely, and that the chance outcomes are all independent within, and between, lotteries L and R. In the example shown in Figure 1, $x = 500$, $y = 300$, and $z = 150$. In lottery R, a zero-mean risk occurs in the high wealth state x, while in lottery L, it occurs in the low wealth state y. A prudent individual prefers lottery R over lottery L because accepting the risk in the high wealth state x dis-aggregates harms than taking in the low wealth state y. We define the prudence measure `V_PRUDscore` as the number of option R among the ten questions.

To measure temperance, we present the options L and R illustrated in Figure 2. As in the case of prudence, the decision maker has the choice between aggregating (lottery R) or dis-aggregating (lottery L) two harms. The harms are two zero mean lotteries of sizes z_1 and z_2, both of which have equally likely positive and negative realizations. In the example shown in Figure 2, $z_1 = 250$ and $z_2 = 150$. A

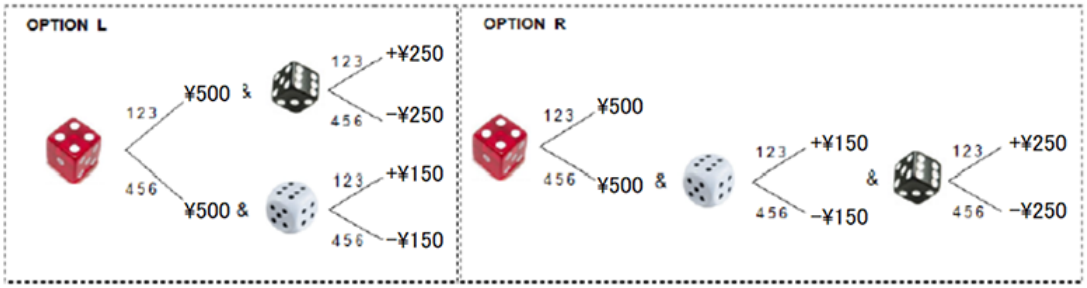


Figure 2: Graphical presentation of temperance tasks.

temperate individual prefers lottery L to dis-aggregate of the two risks. We define the temperance measure `V_TEMPscore` as the number of option L among the five questions.

2.2.2 Index of Loss Aversion

We used the experimental task proposed by Köbberling and Wakker (2005) to measure a degree of loss aversion. We asked participants to choose between a sure zero payment, and, on the other hand, a lottery in which they would get 600 JPY with 50% chance and lose X JPY with 50% chance where X takes 120, 240, 360, 480, 600, or 720.

We assume that loss averse individuals tend to choose the sure zero payment option. Then, we define the measure of loss aversion `V_lossAverse` as the number of choices of safe option among the six questions.

2.3 Personality Trait

2.3.1 Ten Item Big Five Personality Inventory

The Ten Item Personality Inventory (TIPM) is a task to measure personality called the Big Five. The Big Five (also referred to as the five-factor) is the most widely used personality trait model. The Big Five consists of the following five traits: openness to experience (`V_OpennessToExperience`), conscientiousness (`V_Conscientiousness`), extraversion (`V_Extraversion`), agreeableness (`V_Agreeableness`), and emotional stability (`V_EmotionalStability`).

The TIPM includes a total of ten questions, two for each personality trait. Each question has the form of a seven-point Likert scale. We use the average of the scores of the two questions for each personality trait as the measure. Thus, the measures `V_OpennessToExperience`, `V_Conscientiousness`, `V_Extraversion`, `V_Agreeableness`, and `V_EmotionalStability` are each defined by a real number between 1 and 7. We used the Japanese translated version by Oshio et al. (2012).

2.3.2 General Trust Scale

The General Trust scale measures participants' beliefs about the honesty and trustworthiness of others, in general. The six-item questionnaire, comes from Yamagishi and Yamagishi (1994), is often used as a General Trust scale.

To measure the General Trust simply, instead, we used the following two questions. The first question asks whether one agrees or disagrees with that "in general, most people are trustworthy." The second question asks, "how much does one trust people one meets for the first time?" The latter question based on the following finding: being able to trust people you meet for the first time, who belong to an out-group, means that you have a high tendency of General Trust (Welzel, 2010). We required a response with a four-point Likert scale for each question and defined the General Trust scale `V_GTscale` as the mean of two scores. Thus, the measure `V_GTscale` takes a real number between 1 and 4.

2.3.3 Grit Scale

The Grit scale measures an individual's Grit, i.e., perseverance and passion for long-term goals (Duckworth et al., 2007).

We used the Grit-S developed by Duckworth and Quinn (2009), which consists of four items that measure perseverance and four items that measure passion. We used the Japanese translated version by Kanzaki (Duckworth, 2016). Each question has the form of a five-point Likert scale. We define the Grit measure `V_grit` as the average of the scores, and thus `V_grit` takes a real number between 1 and 5.

2.3.4 Slider Measure of Social Value Orientation

The Social Value Orientation (SVO) slider measure, proposed by Murphy et al. (2011), is a measure of social preference defined on a one-dimensional continuum. SVO characterizes the decision-maker's weighting of the allocation of payoffs between themself x and their opponent y. In the SVO framework, we assume that the

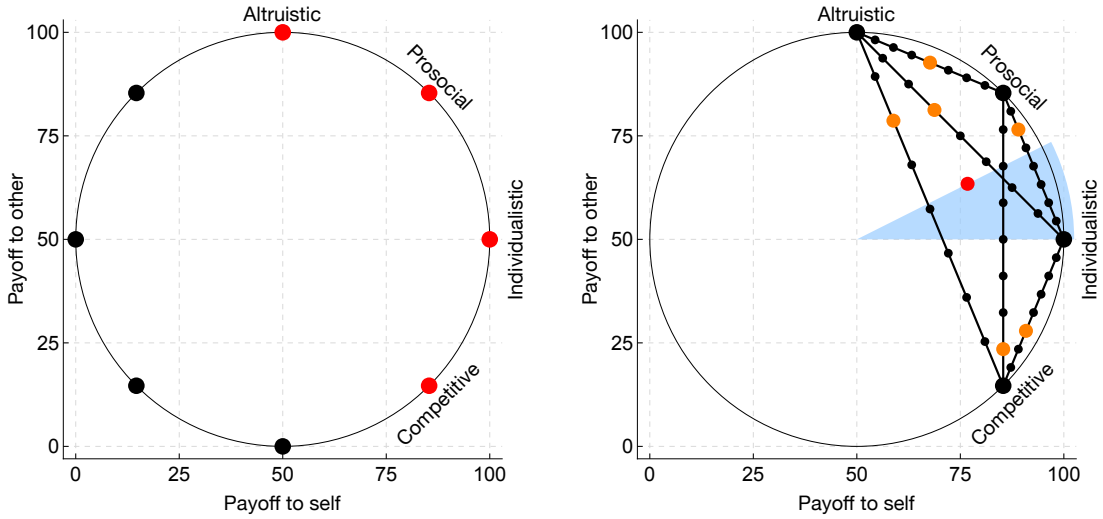


Figure 3: A graphical representation of the general SVO (the left) and SVO slider measure task (the right).

decision-maker maximizes $ax + by$ where a is the weighting to the self and b is to the opponent. For example, the individualistic SVO corresponds to $(a, b) = (1, 0)$, the prosocial SVO to $(a, b) = (1, 1)$, the altruistic SVO to $(a, b) = (0, 1)$, and the competitive SVO to $(a, b) = (1, -1)$. These four SVOs are often considered typical. Figure 3 shows a two-dimensional graph where the horizontal axis measures self payoff x and the vertical axis measures other's payoff y. On the circle drawn in the left panel of Figure 3, each decision-maker with individualistic, prosocial, altruistic, and competitive SVO chooses each red point (or payoff bundle). Arraying the four typical SVOs on a one-dimensional scale, the SVO slider measure maps the decision-maker on this scale.

The task of SVO slider measure consists of six generalized dictator games, which vary in the conversion rates between tokens allocated to the participant and their opponent. The allocation bundle that participants can choose is on the line segment bounded by two endpoints. The endpoints in each game are any two of the four typical bundles on the circle in Figure 3. There are six ways to choose two of the four end points, and thus participants are asked to make decisions in six games. We used

the discrete-choice implementation based on Crosetto et al. (2019), in which participants choose between nine evenly aligned points on each line. In the right panel of Figure 3, the black and orange points on each line are the options participants can choose. On the decision screen, the nine options are aligned horizontally.

The SVO slider measure `V_SVOangle` is defined as the central angle on the circle between the individualistic point $(x, y) = (100, 50)$ and the geometric center of the bundles chosen by the participants $(x, y) = (\overline{x}, \overline{y})$ where $\overline{x}$ and $\overline{y}$ are the averages of the allocations to the oneself and opponent, respectively. That is, we can obtain the measure as following:

$$\texttt{V_SVOangle} = \arctan\left(\frac{\overline{y} - 50}{\overline{x} - 50}\right)\frac{180}{\pi}\,[\text{deg}].$$

See the right panel of Figure 3. The orange points show examples of decisions made in the six dictator game tasks, and the red point is the geometric center of these six points. In this example, `V_SVOangle` is the central angle marked in blue. The minimum value of `V_SVOangle` is $-16.26°$ and the maximum is $61.39°$. Murphy et al. (2011) also proposed the following classification: altruists would have an angle greater than $57.15°$; prosocials would have angles between $22.45°$ and $57.15°$; individualists would have angles between $-12.04°$ and $22.45°$; and competitive types would have an angle less than $-12.04°$.

2.4 Theory of Mind

2.4.1 False Belief Test

The false belief test is a task that measures whether an individual has a theory of mind. The theory of mind refers to the ability to understand and infer the mental states, beliefs, intentions, desires, and perspectives of others.

The task consists of stories describing false beliefs. A true/false question that follows the stories refers either to reality or false representation. To correctly answer

the belief task, it is necessary to realize that the other person in the description has a different belief than the self.

The original problem set of Dodell-Feder et al. (2011) consists of 20 belief tasks, and 20 control tasks that do not require a theory of mind to be answered correctly. We chose five questions for the belief task and five questions for the control task in the following way. First, we conducted a preliminary experiment to obtain the score distribution of the original 40-question version. We selected questions so that the distribution of scores is approximated to the distribution of the 40-question version with respect to Kolmogorov-Smirnov statistics.

We presented each statement for 14 seconds, followed by the question, and asked the participants to answer within 10 seconds. We used the Japanese translated version by Ogawa et al. (2017). We defined the false belief test measure `V_ToMLScoreTot` as the score of only the belief tasks, and thus `V_ToMLScoreTot` takes between 0 and 5.

2.4.2 Reading the Mind in the Eyes Test

The Reading the Mind in the Eyes Test (RMET) is a task that measures an individual's ability to understand the words that describe their mental state and map them to facial expressions. In each question, experimenters presented participants with a photograph of a person's face cropped only for the eyes and asked them to choose a word of the four options that was likely to describe the person's emotion in the photograph. The RMET was developed by Baron-Cohen et al. (1997) and Baron-Cohen et al. (2001) to measure theory of mind abilities in very high functioning (i.e., no cognitive impairment) adults with autism spectrum disorder.

The original problem set consisted of 36 questions, but conducting all questions online would be demanding on the participants. Thus, we chose ten questions from the original 36 in the following way. First, we conducted a preliminary experiment to obtain the score distribution of the 36-question version of the RMET. We selected ten questions so that the distribution of scores is approximated to the distribution of

the 36-question version with respect to Kolmogorov-Smirnov statistics. We defined the RMET measure `V_eyeTest` as the number of correct answers among the ten questions. We used the Japanese translated version by Yamada and Murai (2005).

2.4.3 Heider and Simmel Movie Test

We used the Heider and Simmel movie test, as in the Bruguier et al. (2010), to measure participants' theory of mind abilities. The movie presented in this task, which Heider and Simmel (1944) developed, displays geometric shapes whose movements imitated social interaction. Heider and Simmel (1944) found that when people watch videos that appear mere figures rather than persons, they can think that the geometric shapes have intentions or emotions. According to Baron-Cohen (1995), this ability is related to the intentionality detector, which is one of the modules that form the theory of mind.[1]

The details of the task are described below. In the movie, two triangles, one circle, and a partially opened rectangular frame appear. The video's length is about a minute and a half, and we paused the movie every five seconds. We asked the participants whether the distance between the two triangles would get closer, farther, or not change after each pause of the movie. For each question, respondents were given five seconds to respond, and if they did not respond within that time, they were forced to move on to the next question. There are 16 opportunities for participants to make predictions, and thus we define the Heider and Simmel movie test measure `V_HeiderScoreTot` as the number of correct answers among the 16 questions.

[1]In previous studies by psychologists using the Heider and Simmel movie, participants were asked to describe the video's content and scored by analyzing their narratives (e.g., Klin, 2000). Unlike the methods that are usually applied in psychology, Bruguier et al. (2010) asked participants to predict the motion of objects in the video and choose an answer from options; that is easier for researchers to score. Bruguier et al. (2010) found that investors who scored higher on the Heider and Simmel movie test were better able to forecast markets where insiders exist.

2.5 Instructional Manipulation Checks

The Instructional Manipulation Checks (IMC) detect whether the participants carefully read the material and answer the questions appropriately. The IMC was proposed by Oppenheimer et al. (2009) as a tool to improve a dataset's reliability.

The IMC has such a lengthy statement that participants may be somewhat hesitant to read them all. At the end of the statement, there are questions. The statement asks the participant to ignore these questions. When the participant nevertheless responds, we judged that the participant lacks attention[2]. If the participant did not understand the statement's requirements, they should be tempted to answer.

The questions that follows the long description are similar to the TIPM with a Likert scale. We used the Japanese translated version by Miura and Kobayashi (2015). We named the indicator of IMC success `V_good`, and `V_good` $= 1$ means that the check was successfully completed.

3 Design

As explained in the previous section, we have 12 experimental tasks in total. We divided the experimental tasks into three waves. The first wave experiment included five tasks: CRT, Big Five questionnaire, General Trust questionnaire, RMET, and IMC. The second wave included four tasks: the game of backward induction, the false belief test, the risk attitude elicitation, and Grit questionnaire. The third wave included four tasks: ICAR test, the loss aversion elicitation, SVO slider measure elicitation, and Heider–Simmel movie test. The tasks composing each wave are summarized in Table 1.

We paid respondents at a 10% chance. Immediately after the participants com-

[2]To be precise, we detected when the participant answers all three questions. Since the answer interface was in the form of radio buttons, it could not be undone once the participant had clicked on it. If the participants could notice the true meaning of the question by the first two of the three items, we did not detect it as a satisficing behavior.

Table 1: Summary of measures.

	scale	incentive	category
1st wave			
V_crt6	$\{0, 1, \ldots, 6\}$		Cognitive Ability
V_Extraversion	$[1, 7]$		Personality Trait
V_Agreeableness	$[1, 7]$		Personality Trait
V_Conscientiousness	$[1, 7]$		Personality Trait
V_EmotionalStability	$[1, 7]$		Personality Trait
V_OpennessToExperience	$[1, 7]$		Personality Trait
V_GTscale	$[1, 4]$		Personality Trait
V_eyeTest	$\{0, 1, \ldots, 10\}$		Theory of Mind
V_eyeTest_extd	$\{0, 1, \ldots, 5\}$		Theory of Mind
2nd wave			
V_BI_Gneezy	$[0, 1]$	In case of winning	Cognitive Ability
V_RAscore	$\{0, 1, \ldots, 5\}$	One answer randomely chosen	Risk Attitude
V_PRUDscore	$\{0, 1, \ldots, 10\}$	One answer randomely chosen	Risk Attitude
V_TEMPscore	$\{0, 1, \ldots, 5\}$	One answer randomely chosen	Risk Attitude
V_grit	$[1, 5]$		Personality Trait
V_ToMLscoreBelief	$\{0, 1, \ldots, 5\}$		Theory of Mind
3rd wave			
V_ICARscore	$\{0, 1, \ldots, 10\}$	The number of correct answers	Cognitive Ability
V_lossAverse	$\{0, 1, \ldots, 6\}$	One answer randomely chosen	Risk Attitude
V_SVOangle	$[-16.26, 61.39]$	One answer randomely chosen	Personality Trait
V_HeiderScoreTot	$\{0, 1, \ldots, 16\}$	The number of correct answers	Theory of Mind

pleted all the tasks for each wave, the computer selected the winners by generating random numbers. The results of the lottery were fed back to the participants on the final screen. In the first wave experiment, we paid a fixed reward of 1,000 JPY to the winner. In the second and third wave experiments, the amount we paid to the winners differed for each respondent. Respondents who participated in all three waves were rewarded a 1,000 JPY bonus with a 10% chance. We implemented payments by sending an Amazon gift card (e-mail version).

In the second wave experiment, the reward depended on the respondent's decision in the Risk Attitude elicitation task and the respondent's performance in the Game of Backward Induction. For Risk Attitude elicitation task, one of the 20 questions is randomly selected, and its payoff is realized. The minimum possible payoff is 50 JPY, and the maximum is 1,700 JPY. For the Game of Backward Induction, we added 500 JPY to the payoff for winning the game against the computer, i.e.,

successfully choosing 21.

In the third wave experiment, the reward depended on the score in the ICAR test and the Heider–Simmel movie test, and on the decisions in the Loss Aversion elicitation and the SVO slider elicitation. For the ICAR test and the Heider–Simmel movie test, we added 30 JPY to the reward for each correct answer. For the SVO slider elicitation, one answer was chosen at random. Tokens which the respondent held to themselves, as a dictator, were added to their reward. At this time, a recipient is randomly chosen from among all other respondents. If the respondent decided to pass some amount of tokens on their recipient as a dictator, this amount was the recipient's payoff. In other words, the winner respondent received not only the amount they held as a dictator but also the amount passed by others as a recipient. The minimum amount a respondent can hold is 150 JPY, the maximum is 300 JPY, and the minimum amount a respondent may be passed is 45 JPY, the maximum is 300 JPY. For the Loss Aversion elicitation, one of the six answers was randomly selected and realized. If a participant loses the lottery in the Loss Aversion task, the realized amount is actually deducted from the rewards in the other tasks.

We conducted the online experiment, using Qualtrics, an online survey software, from August to October 2020. The subject pool at the Institute of Social and Economic Research, Osaka University, managed on ORSEE (Greiner, 2015), has 2378 registered participants[3], and we recruited all of them.

A total of 927 (female: 359, male: 561) people participated in the first wave experiment, 864 (female: 328, male: 539) in the second wave, and 810 (female: 298, male: 501) in the third wave[4]. Of the three experimental waves, 667 (female: 243, male: 418) people participated in all waves, and 1093 (female: 422, male: 659) participated in at least one wave.

Among participants who registered at the pool until October 2020, 1457 peo-

[3]At the date we closed the survey, October 23, 2020.

[4]The sum of the number of male and female participants is not equal to the total number of participants due to the respondents who did not answer their gender.

ple participated in experiments at least once between April 2020 and March 2021, excluding our online survey. Of those, 929 participants responded to our survey.

We asked respondents about their gender, academic field, and hometown (their hometown prefecture for Japanese residents) at the end of the first wave experiment. These data are missing for the respondents who did not participate in the first wave experiment. The gender and academic field data are also recorded in the ORSEE database, and using this data, we complemented them as well as possible.

4 Results

In this section, we first present the results of analyses checking the reliability of the data we have gathered. We then present correlation among individual characteristics we have gathered.

4.1 Reliability of measures

In this subsection, we present the results of various tests we used in order to check the reliability of measures included in the current experiment.

4.1.1 Comparing with binomial distribution

In order to check the reliability of the measured outcomes in those tasks that involve multiple choices, we have compared the experimental outcome with the binomial distribution. When the distribution of the experimental outcome in a given task is similar to the binomial distribution, it is possible that participants have made random choices in the experiment. The results are summarized in Figure 4.

The histogram of the scores of the Heider–Simmel movie test and the plot of the binomial distribution has a similar shape. Although the score distribution significantly differs from the binomial distribution ($p < .001$) for the two-sided Kolmogorov-Smirnov test, the effect size $r = Z/\sqrt{N} = .22$ is not large. The one-

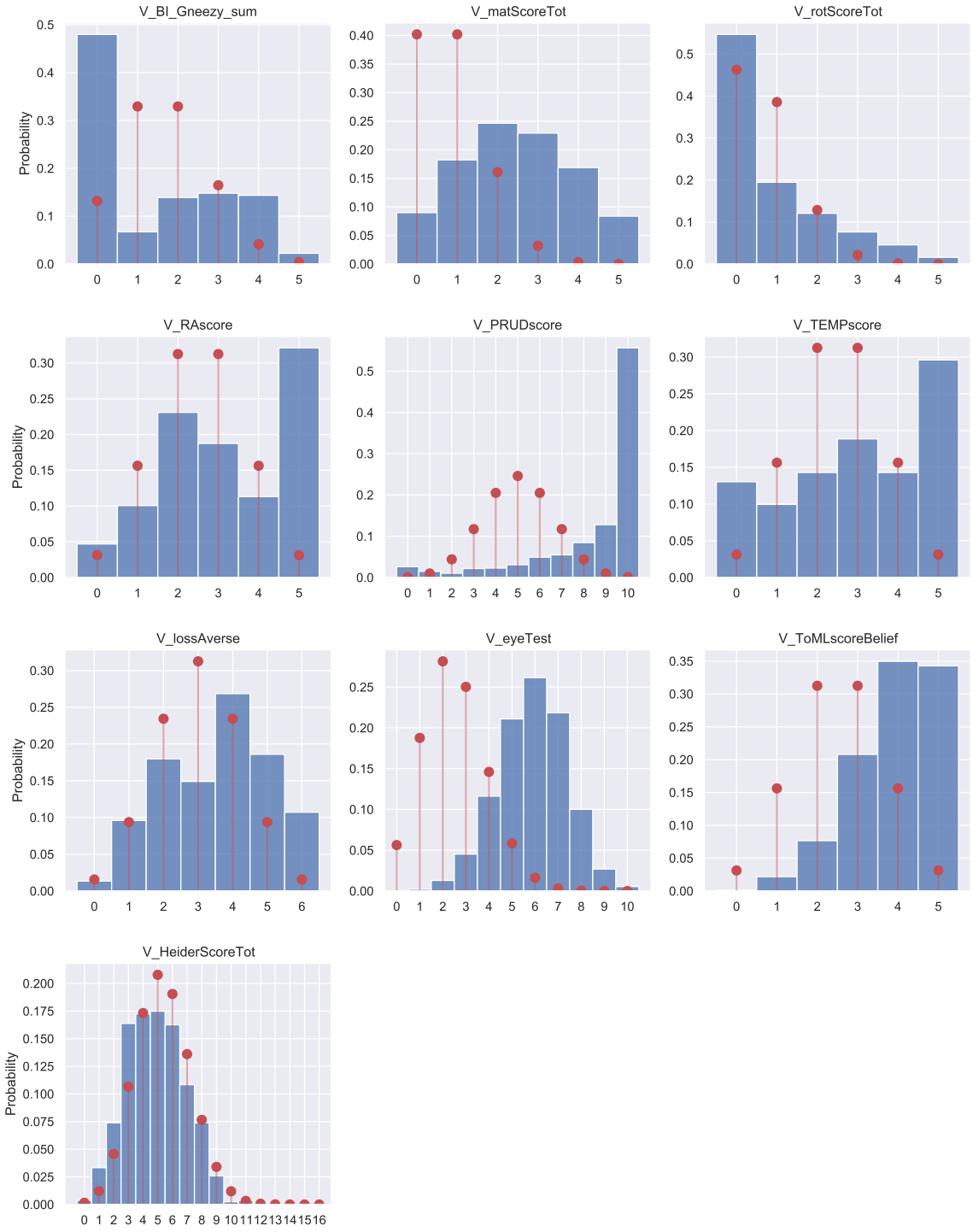


Figure 4: Comparison with binomial distribution.

sided binomial test for each problem showed that at the 5% significance level, only 7 out of 16 problems had a higher percentage of correct answers than the random answer by the three-sided die. Therefore, we conclude that the reliability of the score of the Heider–Simmel movie test is not high.

4.1.2 Uni-dimensionality of measures

To assume that a variable measures a single latent trait, the data composing the variable must have uni-dimensionality. In order to check the uni-dimensionality of the measure, we compute the number of factors by using the scree plot. The scree plots (Figure 5) show, in the decreasing order, the eigenvalues of the correlation matrix of the scores of each underlying question. The number of factor is equal to the number of eigenvalues that are greater than those values based on the correlation matrix among random scores (shown as red dotted line in Figure 5).

According to Figure 5, `V_ICARscore`, `V_lossAverse`, `V_grit`, `V_SVOangle`, `V_eyeTest`, and `V_HeiderScoreTot` have more than one factor. While this is natural for `V_ICARscore` (constructed based on two different measures) and `V_grit` (should have two factors), other measures require a discussion.

The number of factors in the SVO slider measure task is two. These factors can be divided into (a) those based on downward sloping segments, and (b) others (upward sloping and vertical line). Note that, in the former there is a strict trade-off between the payoffs of the other and the participant themself, while such a trade-off is absent in the latter.

The score distribution of the Heider–Simmel movie test was, as we have noted above, close to randomly generated distribution. Thus, this measure, combined with the fact that this measure is not uni-dimensional, seems to be not a reliable measure.

Finally, for the RMET, we have already seen that the score distribution is significantly different from the binomial distribution based on four choices. However, if looking at the frequency distribution of choices made in each question (Table 2), we

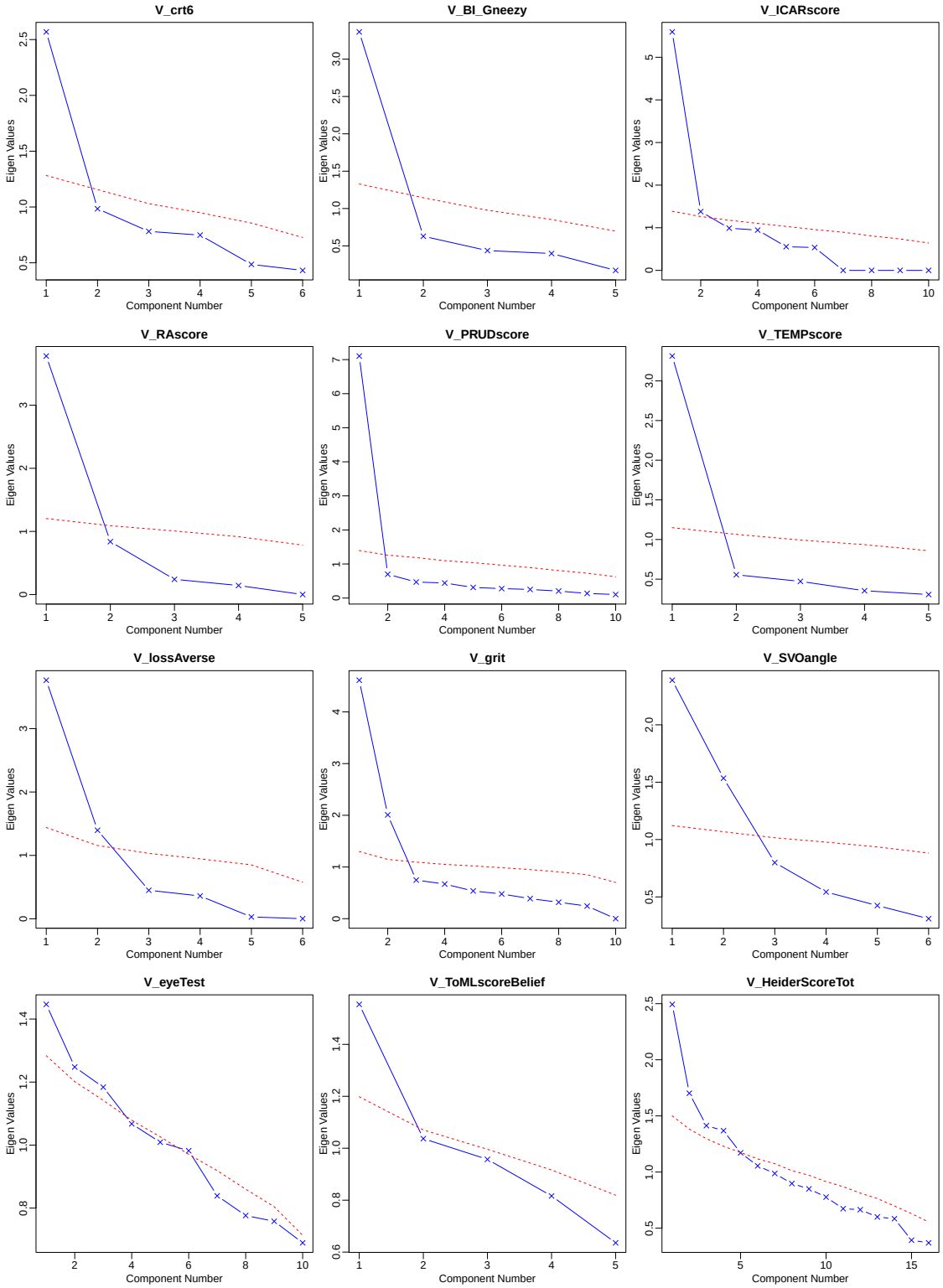


Figure 5: Scree plots.

Table 2: The number of answers for each option in each item of the RMET.

	opt. 1	opt. 2	opt. 3	opt. 4	correct opt.	rate of correct answer	mode in incorrect opts.	rate of incorrect mode
1	708	1	49	169	1	0.76	4	0.18
2	127	72	361	367	3	0.39	4	0.40
3	702	70	44	111	1	0.76	4	0.12
4	31	205	67	624	4	0.67	2	0.22
5	450	3	449	25	3	0.48	1	0.49
6	10	722	139	56	2	0.78	3	0.15
7	14	736	35	142	2	0.79	4	0.15
8	485	6	26	410	1	0.52	4	0.44
9	165	280	70	412	4	0.44	2	0.30
10	58	253	531	85	2	0.27	3	0.57

notice that there are several questions in which responses are divided between two of the four choices. This suggests that, for these questions, participants may have chosen randomly between these two options. We should, therefore, exclude these items because they add noise to the estimation of the traits.[5]

For those measures that are based on the average of two questions (Big Five and General Trust) using the Likert scale, it is more straightforward to verify the measure's consistency, to check the correlation between these two scores than to conduct a factor analysis. Figure 6 shows the polychoric correlation heat map. As we can observe, relatively speaking, the correlation coefficient for Agreeableness and Emotional Stability are small.

4.1.3 Analysis using item response theory

The probability with which an individual with potential trait θ ($\sim N(0,1)$) correctly answers the item j of the task set is called the item characteristic function.[6] Here,

[5]One possible reason for this result is the absence of the dictionary. In the original test, participants could refer to a dictionary that contains the definition of each word used in the test. For this reason, We have decided to introduce a similar dictionary (on line) in the future.

[6]Here "correctly" is defined broadly. For those tasks that have a correct answer, the definition is straight forward. For those tasks without such an answer, such as personal characteristics or risk preference measures, it is defined as an answer that is more likely to be chosen by those individual with potential trait θ.

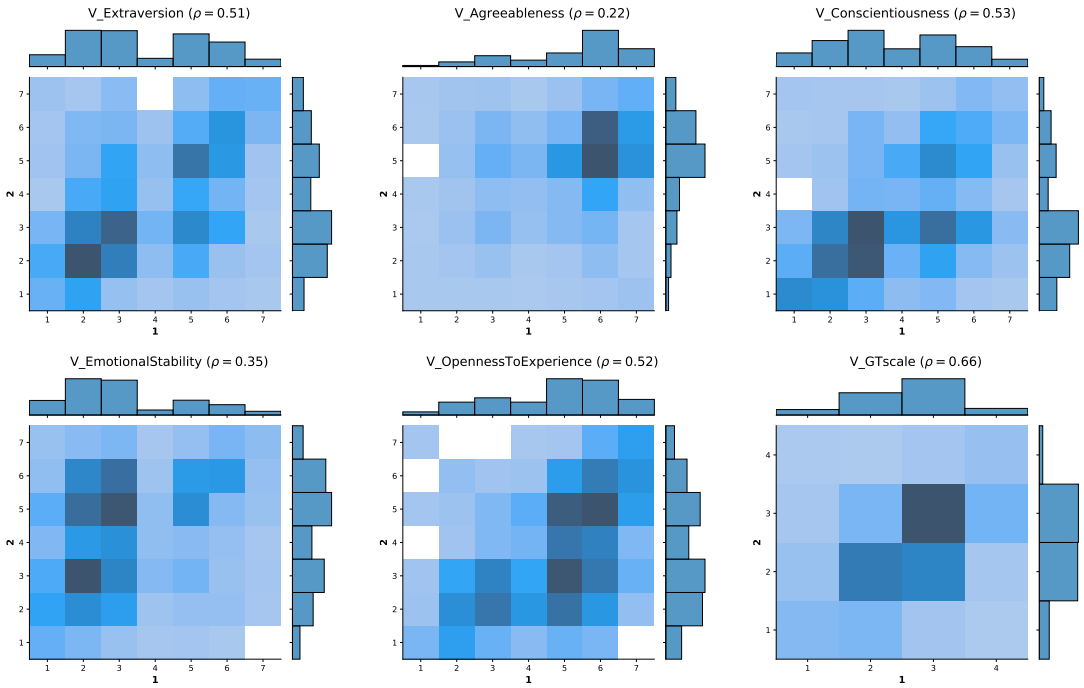


Figure 6: Correlation matrix heat map between two Likert scales that constitute a single measure.

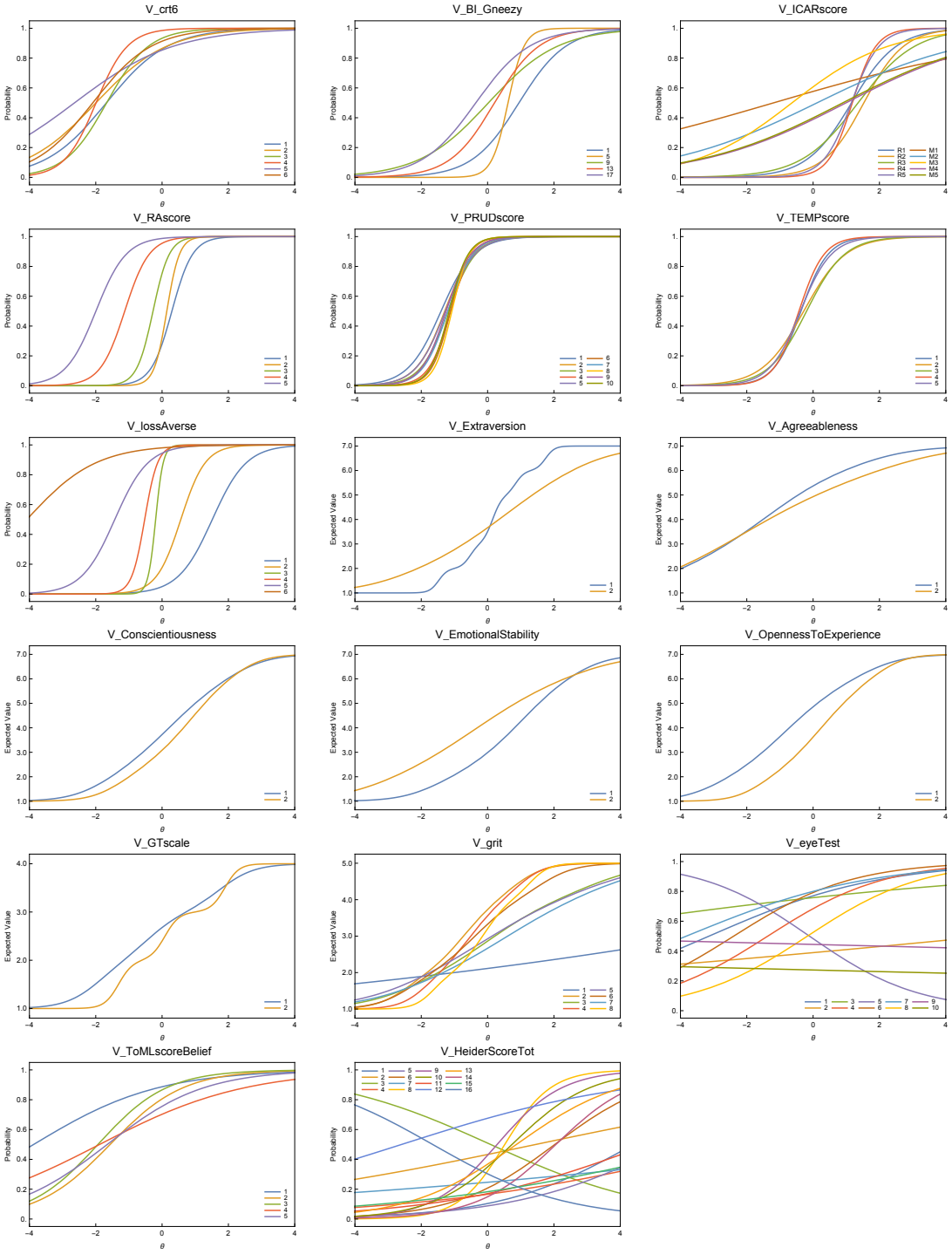
by fitting the item characteristic function,

$$P_j(\theta) = \frac{1}{1 + \exp\big(-a_j(\theta - b_j)\big)},$$

using the logistic model which proposed by Birnbaum (1968), we aim to estimate the discrimination parameter a_j and the difficulty parameter b_j of item j. When a Likert scale is used, such as in Big Five measure, the graded response model proposed by Samejima (1968) is used to estimate the parameter. The Figure 7 shows the item characteristic functions for each task. For tasks using the Likert scale, the expected score is plotted for each item.

The difficulty parameter b is equal to the value of θ for which the probability is 0.5. The more right the curve is located, the more difficult the item is. Furthermore, an individual who is an average trait of the population will correctly answer the item with a probability corresponding to $\theta = 0$.

The slope of the curve corresponds to the discrimination parameter a. An item



Notes. For tasks using the Likert scale, the expected value of the score for each item was plotted.

Figure 7: Item characteristic function.

with a mild slope (i.e., a is small) means no difference in the probability of correct answers resulting from differences in latent traits. A small discrimination power means that the item does not contribute to the discrimination between high and low traits.

An item with a downward curve means that the probability of a correct answer decreases as the characteristic increases. Such items should be excluded from the analysis. We found items with negative discrimination power in the RMET and the Heider–Simmel movie test.

4.1.4 Excluding Data

In the previous subsection, we found that the reliability of the Heider–Simmel movie test measure was not high. Therefore, we decided to exclude this measure from the analysis in the following sections.

We also found that some items in the RMET should be excluded from the analysis. We excluded the five items for which the estimated discrimination parameter of the item characteristic function was negative or close to zero. Thus, in the following analysis, we defined refined measure of RMET `V_eyeTest_extd` as the sum of the scores among the remaining five items.

4.2 Correlations among measured characteristics

The figure 8 shows the heat map of the computed Spearman's correlation matrix after excluding items that are considered unreliable.[7] In addition to the measures described in the previous sections, we include dummy variables indicating gender (`OR_V_Female`), economics students (`OR_econstudent`), medical and nursing students (`OR_medicalstudent`), and graduate students (`OR_gradstudent`), respectively.

In our survey, the sample size of about a thousand is large, and the p-value of the correlation coefficient tends to be small. Therefore, we use the correlation coefficient

[7] See Appendix A for the correlation between these excluded measures and those that have been kept.

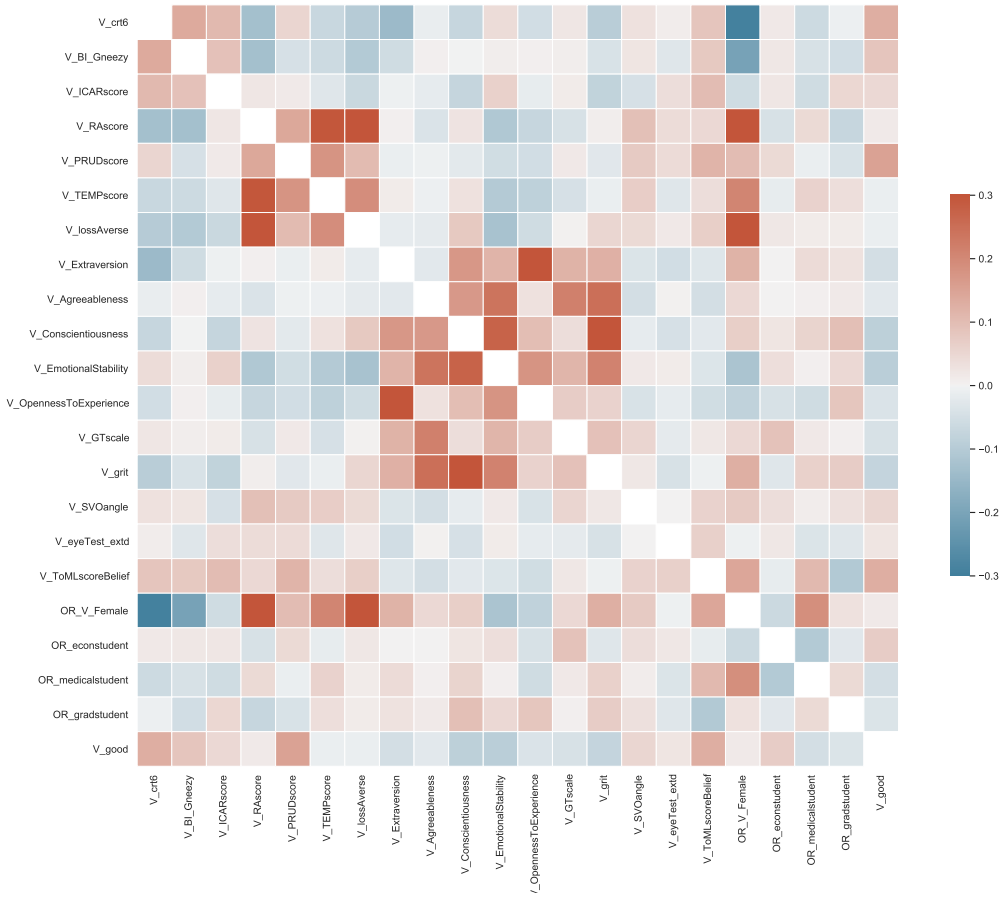


Figure 8: Correlation matrix heat map between measures.

itself as an effect size to determine whether or not a correlation practically exists. Cohen (1988) has defined as conventional criteria that coefficients of .1, .3, and .5 are "small", "medium", and "large", respectively. For the "medium" effect size, Cohen has explained that "this degree of relationship would be perceptible to the naked eye of a reasonably sensitive observer (Cohen, 1988, p. 80)." This Cohen's criterion is widely accepted in studies with behavioral experiments (Field and Hole, 2003). We judged a pair of variables to be effectively correlated if they had an effect size greater than Cohen's "medium" effect size, or if the absolute value of the correlation coefficient was greater than .3. The colors in the heat map are saturated at where the absolute value of the correlation coefficient is .3.

We make following observations.

Observation 1. The pairwise correlation among the measures for which the effect size r is greater than or equal to .3 are as follows: `V_OpennessToExperience`–`V_Extraversion` ($r = .32$); `V_grit`–`V_Conscientiousness` ($r = .54$); `V_TEMPscore`–`V_RAscore` ($r = .30$); `V_lossAverse`–`V_RAscore` ($r = .33$).

Observation 2. The three measures in the cognitive ability category are positively correlated with each other at the 5% significance level, although the effect sizes are small.

Observation 3. The four measures in the risk attitude category are positively correlated with each other at the 5% significance level.

Observation 4. In the theory of mind category, the null hypothesis that `V_eyeTest_extd` and `V_ToMLscoreBelief` are not correlated is not rejected at the 5% significance level ($p = .07$).

Observation 5. The gender indicator is significantly correlated with several scales across the categories. Actually, for 13 out of the 17 measures, except for `V_GTscale`, `V_eyeTest_extd`, `V_Agreeableness`, and `V_ICARscore`, we rejected uncorrelation at the 5% significance level.

Because of Observation 5, we provide more detailed comparisons of the differences between male and female.

Figure 9 plots the empirical cumulative distribution of scores by gender. Tables 3 summarizes the results of the Wilcoxon rank-sum test. The first to third columns show the mean (the standard deviation in parentheses) of each measure for all, male, and female sample, respectively. The last column shows the effect size $r = Z/\sqrt{N}$, and p-value in the parentheses. The effect size corresponds to the correlation coefficient between each variable and a dummy variable representing gender (Rosenthal, 1991). Following Cohen's criterion described above, we conclude that

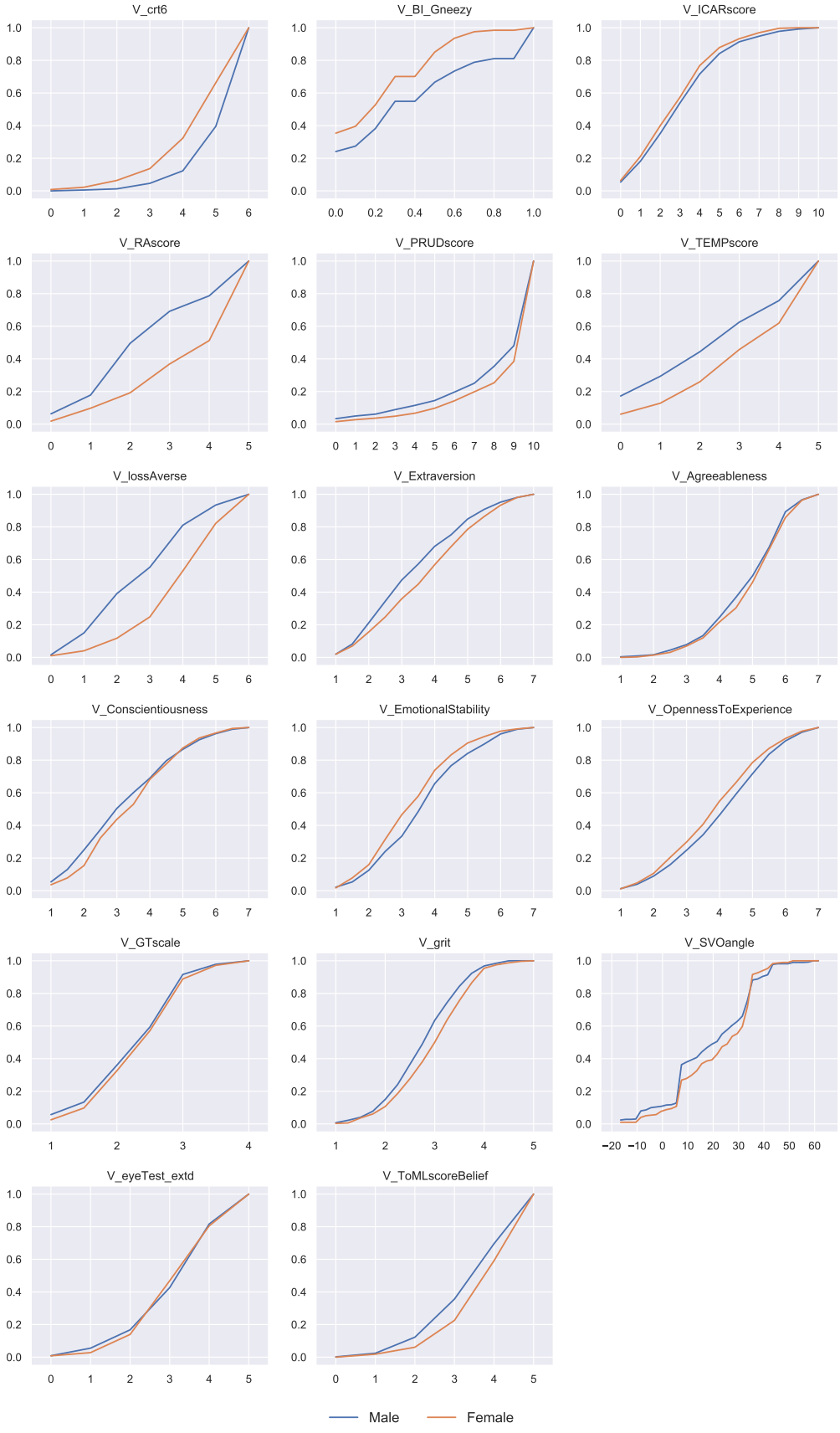


Figure 9: Empirical cumulative distribution of measures by gender.

Table 3: Summary statistics and results of the Wilcoxon's rank-sum test.

	Wave	All	Male	Female	Effect size
Cognitive Ability					
V_crt6	1	5.16	5.42	4.78	0.29
		(1.10)	(0.89)	(1.26)	(8.71×10^{-19})
V_BI_Gneezy	2	0.38	0.44	0.28	0.21
		(0.33)	(0.35)	(0.26)	(1.07×10^{-9})
V_ICARscore	3	3.39	3.48	3.20	0.06
		(2.06)	(2.11)	(1.96)	(1.04×10^{-1})
Risk Attitude					
V_RAscore	2	3.18	2.78	3.81	−0.33
		(1.56)	(1.50)	(1.42)	(1.47×10^{-13})
V_PRUDscore	2	8.42	8.22	8.73	−0.10
		(2.53)	(2.68)	(2.26)	(2.48×10^{-3})
V_TEMPscore	2	3.00	2.71	3.48	−0.21
		(1.75)	(1.79)	(1.55)	(6.32×10^{-10})
V_lossAverse	3	3.55	3.15	4.23	−0.35
		(1.54)	(1.50)	(1.35)	(8.48×10^{-14})
Personality Trait					
V_Extraversion	1	3.73	3.59	3.94	−0.12
		(1.48)	(1.47)	(1.50)	(2.47×10^{-4})
V_Agreeableness	1	5.08	5.03	5.15	−0.05
		(1.14)	(1.15)	(1.11)	(1.51×10^{-1})
V_Conscientiousness	1	3.50	3.43	3.61	−0.07
		(1.44)	(1.48)	(1.36)	(3.98×10^{-2})
V_EmotionalStability	1	3.69	3.81	3.50	0.12
		(1.32)	(1.34)	(1.27)	(3.72×10^{-4})
V_OpennessToExperience	1	4.22	4.31	4.07	0.08
		(1.41)	(1.42)	(1.40)	(1.10×10^{-2})
V_GTscale	1	2.51	2.48	2.56	−0.05
		(0.66)	(0.67)	(0.64)	(1.49×10^{-1})
V_grit	2	3.01	2.94	3.13	−0.13
		(0.72)	(0.70)	(0.73)	(1.81×10^{-4})
V_SVOangle	3	22.18	21.16	23.78	−0.08
		(16.20)	(16.91)	(14.99)	(3.31×10^{-2})
Theory of Mind					
V_eyeTest_extd	1	3.53	3.53	3.55	0.01
		(1.08)	(1.11)	(1.04)	(8.77×10^{-1})
V_ToMLscoreBelief	2	3.91	3.80	4.10	−0.14
		(1.03)	(1.06)	(0.95)	(2.13×10^{-5})

gender differences effectively exist for variables whose effect sizes are greater than .3.

We can make following observations:

Observation 6. For all four measures of the risk attitude category, the tendency was stronger for women. In particular, the effect size is larger than .3 on risk aversion and loss aversion.

Observation 7. Males scored higher on CRT and BI included in the cognitive ability category. Especially for CRT, the effect size is close to .3.

Observation 8. There is no difference between male and female for the other measures, or if there is a difference, the effect size is limited.

5 Summary

This paper summarize the set of individual characteristics that are measured for about one third of 2378 people registered in the participants database of the Institute of Social and Economic Research at Osaka University.

While we have found that two measures of theory of mind we have employed, namely, the measures of RMET and the Heider–Simmel movie test, are not reliable, other measures are.

Consistent with the literature, we find males are less risk and loss averse than females and score better in CRT and backward induction task.

References

Akiyama, Eizo, Nobuyuki Hanaki, and Ryuichiro Ishikawa, "It is not just confusion! Strategic uncertainty in an experimental asset market," *Economic Journal*, 2017, *127*, F563–F580.

Alesina, Alberto and Eliana La Ferrara, "Who trusts others?," *Journal of Public Economics*, 2002, *85* (2), 207–234.

Andersen, Steffen, Seda Ertac, Uri Gneezy, John A List, and Sandra Maximiano, "Gender, competitiveness, and socialization at a young age: Evidence from a matrilineal and a patriarchal society," *Review of Economics and Statistics*, 2013, *95* (4), 1438–1443.

Andreoni, James and Lise Vesterlund, "Which is the fair sex? Gender differences in altruism," *Quarterly Journal of Economics*, 2001, *116* (1), 293–312.

Balafoutas, Loukas, Rudolf Kerschbamer, and Matthias Sutter, "Distributional preferences and competitive behavior," *Journal of Economic Behavior & Organization*, 2012, *83* (1), 125–135.

Baron-Cohen, Simon, *Mindblindness: An essay on autism and theory of mind.* Learning, development, and conceptual change., Cambridge, MA, US: The MIT Press, 1995.

_ , Sally Wheelwright, Jacqueline Hill, Yogini Raste, and Ian Plumb, "The 'Reading the Mind in the Eyes' Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism," *Journal of Child Psychology and Psychiatry*, 2001, *42* (2), 241–251.

_ , Therese Jolliffe, Catherine Mortimore, and Mary Robertson, "Another Advanced Test of Theory of Mind: Evidence from Very High Functioning Adults with Autism or Asperger Syndrome," *Journal of Child Psychology and Psychiatry*, 1997, *38* (7), 813–822.

Basteck, Christian and Marco Mantovani, "Cognitive ability and games of school choice," *Games and Economic Behavior*, 2018, *109*, 156–183.

Benjamin, Daniel J., Sebastian A. Brown, and Jesse M. Shapiro, "Who is 'behavioral'? Cognitive ability and anomalous preferences," *Journal of European Economic Association*, 2013, *11* (6), 1231–1255.

Birnbaum, Allan, "Some latent trait models and their use in inferring an examinee's ability," in F. M. Lord and M. R. Novick, eds., *Statistical Theories of Mental Test Scores*, Addison-Wesley Pub. Co, 1968, pp. 395–479.

Booth, Alison L and Patrick Nolen, "Gender differences in risk behaviour: Does nurture matter?," *The Economic Journal*, 2012, *122* (558), F56–F78.

Booth, Alison, Lina Cardona-Sosa, and Patrick Nolen, "Gender differences in risk aversion: Do single-sex environments affect their development?," *Journal of Economic Behavior & Organization*, 2014, *99*, 126–154.

Bosch-Rosa, Ciril, Thomas Meissner, and Antoni Bosch-Domènech, "Cognitive Bubbles," *Experimental Economics*, 2018, *21*, 132–153. doi:10.1007/s10683-017-9529-0.

Bruguier, Antoine J., Steven R. Quartz, and Peter Bossaerts, "Exploring the Nature of "Trader Intuition"," *The Journal of Finance*, 2010, *65* (5), 1703–1723.

Buchan, Nancy R, Rachel TA Croson, and Sara Solnick, "Trust and gender: An examination of behavior and beliefs in the Investment Game," *Journal of Economic Behavior & Organization*, 2008, *68* (3), 466–476.

Burnham, Terence C., David Cesarini, Magnus Johannesson, Paul Lichtenstein, and Björn Wallace, "Higher cognitive ability is associated with lower entries in a p-beauty contest," *Journal of Economic Behavior and Organization*, 2009, *72*, 171–175.

Camerer, C. F., A. Dreber, T.H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Forsell, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeisser, M. Razen, and H. Wu, "Evaluating Replicability of Laboratory Experiments in Economics," *Sicence*, 2016, *351* (6280), 1433–1436.

Carpenter, Jeffrey, Michael Graham, and Jesse Wolf, "Cognitive ability and strategic sophistication," *Games and Economic Behavior*, 2013, *80*, 115–130.

Chapman, J., M. Dean, P. Ortoleva, E. Snowberg, and C. Camerer, "Econographics," Working paper w24931, National Bureau of Economic Research 2018.

Charness, Gary and Uri Gneezy, "Strong evidence for gender differences in risk taking," *Journal of Economic Behavior & Organization*, 2012, *83* (1), 50–58.

Chaudhuri, Ananish and Lata Gangadharan, "Gender differences in trust and reciprocity," *Department of Economics - Working Papers Series 875, The University of Melbourne*, 2003.

Cohen, J., *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, 1988.

Condon, David M. and William Revelle, "The international cognitive ability resource: Development and initial validation of a public-domain measure," *Intelligence*, 2014, *43*, 52–64.

Corgnet, Brice, Mark DeSantis, and David Porter, "What makes a good trader? On the role of intuition and reflection on trader performance," *Journal of Finance*, 2018, *73* (3), 1113–1137. Economic Science Institute, Chapman University.

Crosetto, Paolo, Ori Weisel, and Fabian Winter, "A flexible z-Tree and oTree implementation of the Social Value Orientation Slider Measure," *Journal of Behavioral and Experimental Finance*, 2019, *23*, 46–53.

Dean, M. and P. Ortoleva, "The empirical relationship between nonstandard econmic behavior," *Proceedings of National Academy of Science*, 2019, *116* (33), 16262–16267.

Dittrich, Marcus, "Gender differences in trust and reciprocity: evidence from a large-scale experiment with heterogeneous subjects," *Applied Economics*, 2015, *47* (36), 3825–3838.

Dodell-Feder, David, Jorie Koster-Hale, Marina Bedny, and Rebecca Saxe, "fMRI item analysis in a theory of mind task," *NeuroImage*, 2011, *55* (2), 705–712.

Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde, "Are risk aversion and impatience related to cognitive ability?," *American Economic Review*, 2010, *100*, 1238–1260.

Duckworth, Angela, *Yarinuku Chikara: Jinsei no Arayuru Seikou o Kimeru 'Kyukyoku no Nouryoku' o Minitsukeru (Grit: the power of passion and perseverance)*, Tokyo, Japan: Diamond, Inc., 2016. Translated by Akiko Kanzaki.

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

Duckworth, Angela Lee and Patrick D. Quinn, "Development and Validation of the Short Grit Scale (Grit-S)," *Journal of Personality Assessment*, 2009, *91* (2), 166–174. PMID: 19205937.

Dufwenberg, Martin, Ramya Sundaram, and David J. Butler, "Epiphany in the Game of 21," *Journal of Economic Behavior & Organization*, 2010, *75* (2), 132–143.

Eeckhoudt, Louis and Harris Schlesinger, "Putting Risk in Its Proper Place," *American Economic Review*, March 2006, *96* (1), 280–289.

Field, Andy and Graham Hole, *How to design and report experiments*, London, UK: Sage, 2003.

Filippin, Antonio, "Gender differences in risk attitudes," *IZA World of Labor*, 2016, ., .

Finucane, Melissa L and Christina M Gullion, "Developing a tool for measuring the decision-making competence of older adults.," *Psychology and Aging*, 2010, *25* (2), 271–288.

Frederick, Shane, "Cognitive Reflection and Decision Making," *Journal of Economic Perspectives*, December 2005, *19* (4), 25–42.

Gächter, Simon, Benedikt Herrmann, and Christian Thöni, "Culture and Cooperation," *Philosophical Transactions of The Royal Society B*, 2010, *365*, 2651–2661.

Gill, David and Victoria Prowse, "Cognitive Ability, Character Skills, and Learning to Play Equilibrium: A Level-k Analysis," *Journal of Political Economy*, 2016, *124* (6), 1619–1676.

Gneezy, Uri, Aldo Rustichini, and Alexander Vostroknutov, "Experience and insight in the Race game," *Journal of Economic Behavior & Organization*, 2010, *75* (2), 144–155.

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

Hanaki, Nobuyuki, Nicolas Jacquemet, Stéphane Luchini, and Adam Zylbersztejn, "Cognitive ability and the effect of strategic uncertainty," *Theory and Decision*, 2016, *81* (1), 101–121.

Heider, F and M Simmel, "An experimental study of apparent behavior.," *The American Journal of Psychology*, 1944, *57*, 243–259.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter, "Antisocial Punishment Across Societies," *Science*, 2008, *319*, 1362–1367.

Kamas, Linda and Anne Preston, "Can social preferences explain gender differences in economic behavior?," *Journal of Economic Behavior & Organization*, 2015, *116*, 525–539.

Klin, Ami, "Attributing social meaning to ambiguous visual stimuli in higher-functioning autism and Asperger syndrome: The Social Attribution Task.," *Journal of Child Psychology and Psychiatry*, 2000, *41* (7), 831–846.

Köbberling, Veronika and Peter P. Wakker, "An index of loss aversion," *Journal of Economic Theory*, 2005, *122* (1), 119–131.

Kuhn, Peter and Marie Claire Villeval, "Are Women More Attracted to Cooperation Than Men?," *Economic Journal*, 2015, *125* (582), 115–140.

List, John and Uri Gneezy, *The why axis: hidden motives and the undiscovered economics of everyday life*, Random House, 2014.

Masuda, Takehito and Eungik Lee, "Higher order risk attitudes and prevention under different timings of loss," *Experimental Economics*, 2019, *22* (1), 197–215.

Miura, Asako and Tetsuro Kobayashi, "Monitors are not monitored: How satisficing among online survey monitors can distort empirical findings," *Japanese journal of social psychology*, 2015, *31* (2), 120–127.

Murphy, Ryan O., Kurt A. Ackermann, and Michel J. J. Handgraaf, "Measuring Social Value Orientation," *Judgment and Decision Making*, Dec 2011, *6* (8, SI), 771–781.

Niederle, Muriel and Lise Vesterlund, "Do women shy away from competition? Do men compete too much?," *Quaterly Journal of Economics*, 2007, *122* (3), 1067–1101.

Noussair, Charles N., Stefan T. Trautmann, and Gijs van de Kuilen, "Higher Order Risk Attitudes, Demographics, and Financial Decisions," *The Review of Economic Studies*, 2014, *81* (1), 325–355.

Ogawa, Akitoshi, Ryoichi Yokoyama, and Tatsuya Kameda, "Development of a Japanese version of a theory-of-mind functional localizer for functional magnetic resonance imaging," *The Japanese Journal of Psychology*, 2017, *88* (4), 366–375.

Oppenheimer, Daniel M., Tom Meyvis, and Nicolas Davidenko, "Instructional manipulation checks: Detecting satisficing to increase statistical power," *Journal of Experimental Social Psychology*, 2009, *45* (4), 867–872.

Oshio, Atsushi, Shingo Abe, and Pino Cutrone, "Development, Reliability, and Validity of the Japanese Version of Ten Item Personality Inventory (TIPI-J)," *The Japanese Journal of Personality*, 2012, *21* (1), 40–52.

Proto, Eugenio, Aldo Rustichini, and Andis Sofianos, "Intelligence, Personality and Gains from Cooperation in Repeated Interactions," *Journal of Political Economy*, 2019, *127* (3), 1351–1390.

Raven, John, "The Raven's Progressive Matrices: Change and Stability over Culture and Time," *Cognitive Psychology*, 2000, *41* (1), 1–48.

Rosenthal, Robert, *Meta-analytic procedures for social research, Rev. ed.* Applied social research methods series, Vol. 6., Thousand Oaks, CA, US: Sage Publications, Inc, 1991.

Samejima, Fumi, "ESTIMATION OF LATENT ABILITY USING A RESPONSE PATTERN OF GRADED SCORES1," *ETS Research Bulletin Series*, 1968, *1968* (1), i–169.

Snowberg, Erik and Leeat Yariv, "Testing the Waters: Behavior across participant pools," *American Economic Review*, 2021, *111* (2), 687–719.

Toplak, Maggie E., Richard F. West, and Keith E. Stanovich, "Assessing miserly information processing: An expansion of the Cognitive Reflection Test," *Thinking & Reasoning*, 2014, *20* (2), 147–168.

Welzel, Christian, "How Selfish Are Self-Expression Values? A Civicness Test," *Journal of Cross-Cultural Psychology*, 2010, *41* (2), 152–174.

Yamada, Makiko and Toshiya Murai, "Seijinyou 'Me kara Kokoro o Yomu Test' Kaiteiban (Nihongoban) (The Revised Version of the Adult 'Reading the Mind in the Eyes' Test (Japanese version))," 2005. `https://www.autismresearchcentre.com/tests/eyes-test-adult/`.

Yamagishi, Toshio and Midori Yamagishi, "Trust and commitment in the United States and Japan," *Motivation and Emotion*, jun 1994, *18* (2), 129–166.

A Correlation based on all the data

Figure 10 shows the Spearman's correlations between our all measures and unreliable measures. Table 4 summarizes the results of the Wilcoxon rank-sum test for unreliable measures. The qualitative results are the same as the one presented in the main text.

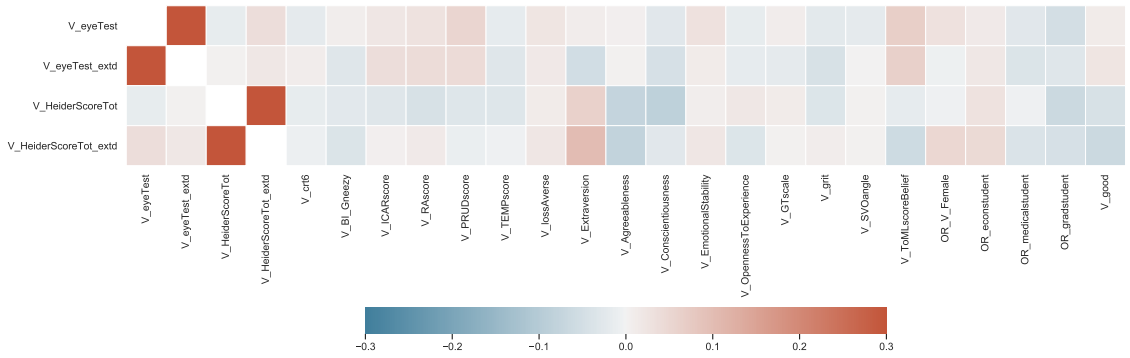


Figure 10: Correlation matrix heat map for unreliable measures.

Table 4: Summary statistics for unreliable measures.

	Wave	All	Male	Female	Effect size
`V_eyeTest`	1	5.88	5.84	5.96	−0.03
		(1.51)	(1.52)	(1.48)	(3.24×10^{-1})
`V_eyeTest_extd`	1	3.53	3.53	3.55	0.01
		(1.08)	(1.11)	(1.04)	(8.77×10^{-1})
`V_HeiderScoreTot`	3	4.87	4.87	4.84	0.00
		(2.01)	(2.03)	(1.97)	(9.25×10^{-1})
`V_HeiderScoreTot_extd`	3	1.96	1.91	2.03	−0.05
		(1.43)	(1.46)	(1.36)	(1.64×10^{-1})