ISSN (Print) 0473-453X
Discussion Paper No. 1160
ISSN (Online) 2435-0982

COMPARING BEHAVIOR BETWEEN A LARGE SAMPLE OF SMART STUDENTS AND A REPRESENTATIVE SAMPLE OF JAPANESE ADULTS

Nobuyuki Hanaki
Keigo Inukai
Takehito Masuda
Yuta Shimodaira

February 2022

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

Comparing behavior between a large sample of smart students and a representative sample of Japanese adults[*]

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

February 7, 2022

Abstract

We address a concern about the external validity, in particular, the representativeness of the sampled population, of an experiment conducted with university students. We do so by conducting large-scale (partly) incentivized online surveys of students at a Japanese university and of a sample of Japanese adults to measure individual characteristics such as cognitive ability, mentalizing skills, preferences for risk and distribution, and personality traits. While significant differences between these two samples are observed in many of these characteristics, the correlational structures among these characteristics are very similar in the two samples.

Keywords: external validity, individual characteristics

JEL codes: C91, D91

[*] 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 the 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

1 Introduction

An increasing number of studies document the relationship between measured cognitive ability and observed behavior in experiments. For example, people with higher cognitive ability tend to take more risks and are also more patient (Dohmen et al., 2010; Benjamin et al., 2013). Those with higher cognitive ability are not only more strategically sophisticated, in that their choices are closer to the prediction of the Nash equilibrium in p-beauty contest games (Burnham et al., 2009; Gill and Prowse, 2016; Carpenter et al., 2013), but are also more sensitive to the uncertainty regarding the behavior of others in a simple dominance-solvable coordination game (Hanaki et al., 2016) as well as in an asset market game (Akiyama et al., 2017). Furthermore, in interactive situations, the composition of participants in terms of their cognitive ability within a group or a pair matters for the outcomes. In repeated prisoner's dilemma games, Proto et al. (2019) show that a pair of higher cognitive ability participants can achieve a higher rate of cooperation than a pair of lower cognitive ability participants. In an experimental asset market à la Smith et al. (1988), Bosch-Rosa et al. (2018) report that while markets consisting of high cognitive ability participants result in little mispricing, those consisting of low cognitive ability participants result in large price bubbles.

These reported systematic relationships between measured cognitive ability and observed behavior suggest, not only individually but also collectively, that measuring and controlling the potential effect of such characteristics are of great importance when replicating published experimental results (for the importance of such a replication exercise see Camerer et al., 2016), or carrying out international comparative experimental studies such as those by Herrmann et al. (2008) and Gächter et al. (2010). These issues have motivated us (as discussed in Hanaki, 2020) to systematically measure cognitive ability, as well as other individual characteristics, of those who have registered in the participant database of the Institute of Social and Economic Research, Osaka University managed by ORSEE (Greiner, 2015).

At the same time, researchers employing controlled laboratory experiments are increasingly trying to address various concerns related to the "external validity" of the results obtained from laboratory experiments. There are two broad issues related to the notion of external validity. One relates to the "representativeness of the sampled population" and the other to the "representativeness of the environment" (List, 2007). The former relates to whether the results obtained from students are applicable to the general population. The latter relates to the effect on the results of the abstract nature of the task employed in the laboratory experiments, including neutral framing.

While researchers have used field experiments extensively to address both types of representativeness (see, for example, Harrison and List, 2004; List, 2007; Gangadharan et al., 2021), compared with field experiments, there are various advantages to running laboratory experiments using university student samples, including the cost, fewer logistical problems, and ease of replication.

The opportunity to run large-scale incentivized online surveys/experiments (Chapman et al., 2018) as well as nonincentivized surveys that are ex ante validated in the incentivized laboratory experiments (Falk et al., 2016, 2018) offers a new way to address the concerns related to the representativeness of the sampled population by comparing the data gathered from a sample of students and the representative population, as in Snowberg and Yariv (2021). Such comparisons not only enable researchers to compare their usual sample (students at their universities) with samples representing the broader population, but also to verify whether certain correlations between measured characteristics and behaviors are observed across sampled populations.

In this paper, we report the results of large-scale incentivized and non-incentivized online surveys (but with the latter ex ante validated with an incentivized experiment, Falk et al., 2016, 2018) that measure cognitive ability, mentalizing skills, preferences related to risk, time, and inequality, as well as personality traits (see Section 2 for details). The surveys comprised a large sample of students at Osaka University (involving 988 participants completing at least one set of tasks, and 526 completing all sets of tasks) and a sample of the Japanese adult population ($20 \leq \text{age} < 70$, involving 1855 participants completing at least one set of tasks, and 1023 completing all sets of tasks) registered in the panel of an online survey company.[1]

While significant differences between the two samples in many of these characteristics are observed, the correlational structures among these characteristics are very similar between the two samples. These results are similar to the finding of Snowberg and Yariv (2021), which compares Caltech and US representative samples. However, this does not mean that the results of the experiments conducted at Osaka University can be generalized to the Japanese population. Nevertheless, as far as the correlation between in-

[1] The age–sex composition of this sample matches that of the Japanese population.

dividual characteristics is concerned, the representativeness of the sampled population does not seem to be a major concern.

The remainder of the paper is organized as follows. Section 2 presents the individual characteristics we have measured. Section 3 discusses the implementation of our large-scale online surveys. The results are presented in Section 4, followed by concluding remarks in Section 5.

2 Measures

Our online surveys consist of 12 tasks. Through these tasks, we aim to measure (1) cognitive ability, (2) mentalizing skills, (3) preferences for risk, loss, ambiguity, and time, (4) distributional preferences, and (5) personality traits, such as trust, the Big Five, Grit (Duckworth et al., 2007), and overconfidence. We supplement these measures with the Global Preference Survey (GPS, Falk et al., 2018). Below, we provide a detailed explanation of each task.

2.1 Cognitive Ability

2.1.1 Cognitive Reflection Test

The Cognitive Reflection Test (CRT, Frederick, 2005) is a simple task to measure cognitive ability. The questions that comprise the CRT are questions that can be incorrectly answered if responded to 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.

Because the three questions proposed by Frederick (2005) have become well known, we replaced them with those used by Finucane and Gullion (2010), and also added three of the questions proposed by Toplak et al. (2014). We used the number of correct answers in these six questions as our CRT score (`V_Crt6`).

2.1.2 Ability to Do Backward Induction

We used a task called the *Race Game* (Gneezy et al., 2010) or the *Game of 21* (Dufwenberg et al., 2010). This game is a two-player, extensive-form perfect-information game. In our task, each participant plays against the computer. Players take turns, and when it is his/her turn, each player can choose an integer, 1, 2, or 3, to add to the sum of the previous integers chosen by both players. The player who reaches 21 is the winner. In our task, the participant moves first, followed by the computer. The computer chooses an integer randomly each time, and participants are informed about this behavior of the computer.

There is a winning strategy for participants, who are the first mover, in this game. It is to choose an integer such that after his/her choice, the sum of the previously chosen integers will be either 1, 5, 9, 13, 17, or 21. To see this, one needs to do backward induction (BI). Whatever the integer the computer chooses, the participant can win if he/she reaches 17. To do so for sure, one should reach 15, etc. We call these six numbers winning numbers.

We define a measure of ability to do BI, `V_BI_Gneezy`, as the number of times a participant successfully reached a winning number (the first time when he/she had a chance to do so) divided by the number of chances they faced.[2] `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 measures cognitive ability as proposed by Condon and Revelle (2014), and is maintained by the ICAR team (see their website, `https://icar-project.com`, for more information). The ICAR test was developed as a public domain tool, and researchers can choose which tasks to include as their experimental tasks.

We used the three-dimensional (3D) rotation measure (four questions) and the matrix reasoning measure (four questions) among those included in ICAR-16 (Condon and Revelle, 2014, Table 4). The 3D 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 asked the participants to answer each of the 3D rotation quizzes in 40 seconds, and the matrix reasoning quiz in 30 seconds. We then define the ICAR measure `V_ICARscore` as the score of the total of the eight items of the two measures.

[2] Ex ante, participants have six chances to reach the winning numbers. However, ex post, it can be less than six. For example, if a participant chose 3 on his/her first turn (and thus, missed the first chance), which was followed by the computer choosing 3 as well, then, the participant had no chance of reaching 5.

2.2 Mentalizing Skills (Theory of Mind)

Mentalizing skills (or theory of mind) refer to the ability to understand and infer the mental states, beliefs, intentions, desires, and perspectives of others. We consider two tasks to measure this skill.

2.2.1 False Belief Test

The task consists of stories describing false beliefs. A true/false question that follows the stories refers either to reality or to the false representation. To answer the belief task correctly, it is necessary to realize that the other person in the description has a different belief to oneself.

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 as follows. First, we conducted a preliminary experiment, using the Japanese translated version by Ogawa et al. (2017), to obtain the score distribution of the original 40-question version. We then selected a total of 10 questions (five questions each for belief and control tasks) so that the distribution of scores approximated the distribution of the 40-question version with respect to Kolmogorov–Smirnov statistics. Finally, we implemented only the five questions for the belief task.

We presented each statement for 14 seconds, followed by the question, and asked the participants to answer within 10 seconds. We defined the false belief test measure `V_ToMLscoreBelief` as the number of correct answers to these five questions.

2.2.2 Reading the Mind in the Eyes Test

The Reading the Mind in the Eyes Test (RMET), developed by Baron-Cohen et al. (1997) and Baron-Cohen et al. (2001), 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 showing only the eyes and asked them to choose a word from four options that best describes the person's emotions in the photograph. The RMET was originally developed to measure mentalizing skills in very high functioning (i.e., no cognitive impairment) adults with autism spectrum disorder.

The original problem set consisted of 36 questions. We chose 10 questions from the original 36 in the following way. First, we conducted a preliminary experiment, using the Japanese translated version by Yamada and Murai (2005), to obtain the score distribution of the 36-question version of the RMET. We then selected 10 questions so that the distribution of scores approximated the distribution of the 36-question version with respect to the Kolmogorov–Smirnov statistics. We defined the RMET measure `V_eyeTest` as the number of correct answers among the 10 questions. An online dictionary was available for participants to verify the meaning of the words from which they selected.[3]

[3] There was no online dictionary in our 2020 implementation (Hanaki et al., 2020). As a result, for some of the questions, the answers of the participants have diverged into two, the correct option and the other option, out of four available options.

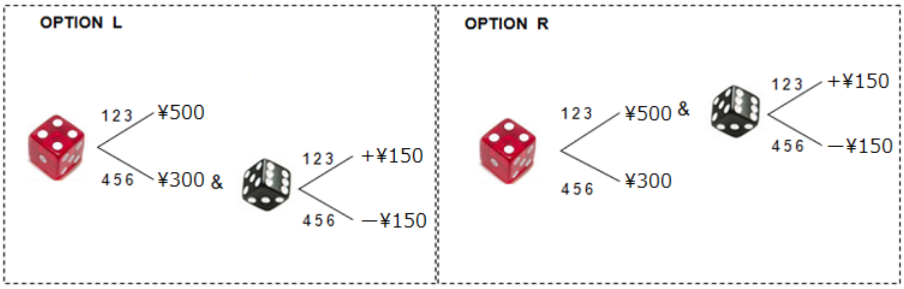


Figure 1: Graphical presentation of prudence tasks.

2.3 Risk, Ambiguity, and Loss Preferences

2.3.1 Risk Aversion and Attitude toward Higher Order Risk

We used the elicitation task originally proposed by Noussair et al. (2014) and also utilized by Masuda and Lee (2019) to measure risk aversion, prudence, and temperance. We asked five questions each for risk aversion, prudence, and temperance. In each task, participants were asked which of the two lotteries they would choose.

To measure risk aversion, we asked participants to choose between a risky lottery in which he/she gets 650 JPY with a 50% chance and 50 JPY with a 50% chance, or a sure payment of X JPY where X equals 200, 250, 300, 350, and 400. As we have only five values of X presented in increasing order, we approximate the certainty equivalent by the (middle of the) switching point from 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` as the number of safe options among the five questions.

To measure prudence, we present 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 disaggregates harms rather than selecting the low wealth state y. We define the prudence measure `V_PRUDscore` as the number of options R among the five questions.

To measure temperance, we present 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 disaggregating (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 temperate individual prefers lottery L to disaggregate the two risks. We define the temperance measure `V_TEMPscore` as the number of options L among the five questions.

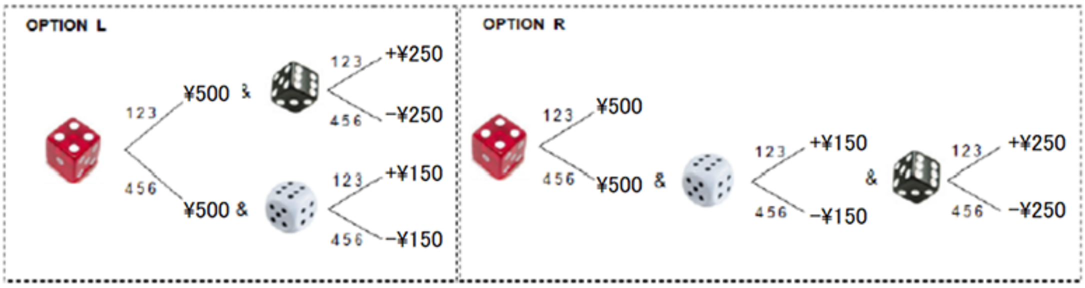


Figure 2: Graphical presentation of temperance tasks.

2.3.2 Index of Ambiguity Attitude

We used the elicitation task of Gneezy et al. (2015). First, participants answer the risk preference elicitation task (with known probability) using a multiple price list. Then, they answer a similar ambiguity aversion elicitation task (with unknown probability). The number of safe choices in the former task is a measure of risk aversion (`V_HLscore`) and the latter is a measure of ambiguity aversion (`V_AmbScore`).

2.3.3 Index of Loss Aversion

We used the experimental task proposed by Köbberling and Wakker (2005) to measure the degree of loss aversion. We asked participants to choose between a sure zero payment, and a lottery in which they would get 600 JPY with a 50% chance or lose X JPY with a 50% chance, where X equals 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 safe options among the six questions.

2.4 Distributional Preferences

2.4.1 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 themselves x and their opponent y. In the SVO framework,

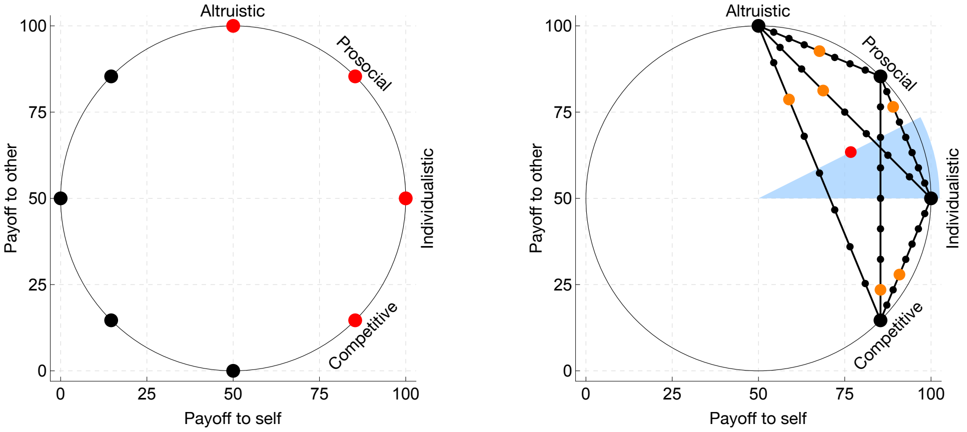


Figure 3: Graphical representation of the general SVO (left) and SVO slider measure task (right).

we assume that the decision-maker maximizes $ax+by$ where a is the weighting applied to the decision-maker and b is the weighting applied to his/her 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, or 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 the 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 dictator 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 endpoints, and thus participants are asked to make decisions in six dictator 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 oneself and the opponent, respectively. That is, we can obtain the measure as follows:

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

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°.

2.5 Personality Traits

2.5.1 General Trust Scale

The general trust scale measures participants' beliefs about the honesty and trustworthiness of others, in general. The six-item questionnaire, 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 the statement that "in general, most people are trustworthy." The second question asks "how much do you trust people you meet for the first time?" The latter question is based on the finding that being able to trust people one meets for the first time, who belong to an outgroup, means that one has a high tendency of general trust (Welzel, 2010). We required a response on 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` equals a real number between 1 and 4.

2.5.2 The Ten-Item Big Five Personality Inventory

The ten-item personality inventory (TIPI) measures personality using the Big Five. The Big Five model (also referred to as the five-factor model) is the most widely used personality trait model. The Big Five model 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 TIPI includes a total of 10 questions, two for each personality trait. Each question requires a response using 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).

2.5.3 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 Short Grit Scale (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 uses 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.

2.5.4 Overconfidence

Following Gillen et al. (2019), we implemented tasks to measure two dimensions of overconfidence, namely overestimation and overplacement, after the ICAR task.[4]

[4] In this paper, we do not measure another dimension of overconfidence, that is, overprecision.

Specifically, after each component of the ICAR task (i.e., 3D rotation task and matrix reasoning task), participants answered the following questions: "How many questions, out of four, do you think you have answered correctly?" and "Out of 100 randomly chosen other participants in this study, how many do you think have answered more questions, the same number of questions, and fewer questions correctly than yourself. Please enter numbers so that the total is 100. After entering your estimates of the numbers of participants who answered more questions and the same number of questions correctly, place click the "compute" button so that the number of those who answered fewer questions correctly will be computed automatically."

Overestimation measures (`V_RotOverEstimation` and `V_MatOverEstimation`) are computed simply as the difference between the estimated number of correct answers and the actual number of correct answers, and overplacement measures (`V_RotOverPlacement` and `V_MatOverPlacement`) are computed as the difference between the estimated percentile and the actual percentile.[5]

[5] In fact, we implemented a similar task after CRT. Specifically, after CRT, to estimate the degree of overestimation, we asked "how many questions, out of six, do you think you have answered correctly?" To estimate the degree of overplacement, we asked "Out of 100 randomly chosen other participants in this study, how many do you think answered more questions correctly than yourself?" However, because the answer to the latter question may depend on whether the participant has included those others who scored the same number of questions correctly as him/herself, we made it more explicit in the questions that followed the ICAR task, and decided not to use the responses to those questions after the CRT in our analyses.

2.6 Global Preference Survey (GPS Falk et al., 2018)

We implemented the preference survey of Falk et al. (2018).[6] Through non-incentivized self-assessment, but ex ante validated by an incentivized laboratory experiment (Falk et al., 2016), this survey measures the degree of patience, risk taking, positive and negative reciprocity, altruism, and trust. The survey also contains a question on tendency to procrastinate (WP13426) and self-reported math ability (WP13425). We use the same weighting method as Falk et al. (2018, Supplementary material p. 20) in constructing the measures. However, the z-score is derived within our sample (i.e., not in the same way as in Falk et al. (2018) because the information necessary to do so is unavailable). Note that because there are such options as "do not know" and "refuse to answer" (for patience, risk taking, reciprocity, and trust), as well as the possibility of participants proceeding to the next question without answering (questions related to altruism), there are missing observations in some of the variables.

3 Implementation

Online surveys were conducted between July and November 2021.[7] Because there are many tasks, we separated them into four blocks, as summarized in Table 1. Within each block, the order of tasks was randomized across par-

[6] We obtained the Japanese version of the survey from `https://www.briq-institute.org/global-preferences/home`. We modified the wording of a few questions (WP13422, WP13419, WP13420, and WP13458) as the original Japanese version sounded somewhat strange to us. Please contact us for further details.

[7] Surveys were implemented using the platform provided by Qualtrics `https://www.qualtrics.com`. Surveys for the Japanese adult sample were conducted by GMO Research `https://gmo-research.jp/`.

Table 1: Tasks in each block and number of participants

Block	Included tasks	N_{OU}	N_{nonOU}
1	RMET, TIPI, CRT, General trust scale, Ambiguity attitude	754	1023
2	False Belief Test, Higher Order Risk Preference, Grit, Game of 21	762	1855
3	SVO, ICAR, Loss aversion	745	1855
4	GPS	719	1023

N_{OU} and N_{nonOU} refer to the number of OU sample participants and the number of participants from the Japanese adult population (nonstudents) completing each block, respectively.

ticipants. Furthermore, in each block, to check if participants were carefully reading the material in the online experiment, we implemented instructional manipulation checks (IMC), which were proposed by Oppenheimer et al. (2009) as a tool to improve a dataset's reliability (see Appendix A for details).

At Osaka university (OU sample), these four blocks were implemented separately as four different online sessions. For our sample of the Japanese adult population (we refer to this sample as the "nonstudent sample" although there were a few students ($\approx 2\%$) in the sample), for logistical reasons, we further grouped two blocks into one session (Blocks 2&3 and Blocks 1&4), while randomizing the order of the two blocks across participants within a session.

For the OU sample, data were gathered over three separate waves. In the first wave, an invitation e-mail for each block (session) was sent two weeks apart in the months of July to August 2021. In the second wave, another invitation e-mail for each block (session) was sent one week apart in September 2021 to those who did not participate in their respective block in the first wave.[8] Finally, in the third wave, which was conducted in October 2021, yet another invitation e-mail for each block (session) was sent one week apart to those who did not participate in their respective block in the previous two waves.[9] A total of 988 participants completed at least one of the four blocks across the three waves (see Table 1 for the number of participants who completed each block); among them, 526 completed all four blocks.[10]

For the nonstudent sample, two sessions were conducted within the same week during November 2021. Following advice from the company that conducted the survey, we gathered responses from 1800 participants in the first sessions (reflecting the age–sex composition of the Japanese population).[11] In the end, we had 1855 complete responses because we allowed those participants who had already started answering the survey to complete it even after we reached the targeted number of complete responses. We reinvited these 1855 participants to the second session specifying that we would stop

[8] We shortened the interval between the two sessions because most participants responded to the invitation e-mail within three days.

[9] There were 2947 participants registered in our database (managed by ORSEE Greiner, 2015) as of December 2021. Among them, 2659 were registered prior to July 2021, and all were invited to participate in our first wave. The number of participants increased by 119 during the months of July and August, and by 69 in September 2021. These newly registered participants were invited to participate in our subsequent waves. Note that not all the registered participants took part in our laboratory experiments. In the 2020–2021 academic year (i.e., April 2020 to March 2021), 1457 students among the 2378 registered (about 63%) participated in laboratory experiments at least once.

[10] See Appendix F for the number of participants and descriptive statistics for each wave.

[11] We based our quota on the population estimates of October 2021, which are based on the census conducted in 2015. See `https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00200524&tstat=000000090001&cycle=1&year=20210&month=24101210&tclass1=000001011678`. This resulted in the following composition:

age group	20s	30s	40s	50s	60s	total
Male	154	164	212	201	176	907
Female	144	158	207	200	184	893

accepting responses when we reached a total of 1000 complete responses.[12]. Finally, we had 1023 complete responses, slightly exceeding our target, for the same reason as in the first session. See Appendix C for demographic information of the nonstudent sample.

In each block, in addition to a small participation fee,[13] each eligible participant had a 10% chance of being selected for a performance-based payment.[14] If a participant was selected for the performance-based payment in Blocks 1, 2, or 3, one of the tasks with a monetary incentive was chosen randomly and the selected participant was paid based on the amount they earned in the selected task,[15] and, in Block 4, a fixed reward of 1000 JPY was paid.

[12] The age–sex composition of the Japanese population was respected as follows:

age group	20s	30s	40s	50s	60s	total
Male	85	91	118	112	98	504
Female	80	88	115	111	102	496

[13] For OU sample participants, the participation fee was either a 10% chance to win 1000 JPY by completing all four blocks (July–August wave) or 100 JPY for completing each block (September and October waves). These differences in the way participation fees were paid, as well as the differences in the participated waves, did not result in statistically significant differences in the measured characteristics. See Appendix F. For the nonstudent sample participants, the participation fee was set to the standard amount determined by GMO Research, which depended on the number of questions included in a session.

[14] To be eligible for the performance-based payment, participants needed to complete all the questions in the block and pass the attention check question (IMC) presented in Appendix A. Note that for the nonstudent sample, there were two blocks within a session, and these blocks were considered independently for the bonus payment. Thus, although it did not occur, a participant could have been selected for the bonus payments for both blocks.

[15] OU sample participants were paid with Amazon gift cards (e-mail version) and the nonstudent sample participants were paid with equivalent amounts of GMO points

4 Results

We first compare the distributions of various measured characteristics between the OU and nonstudent samples using the entire dataset, i.e., using all the participants who completed the block that contains the respective task.

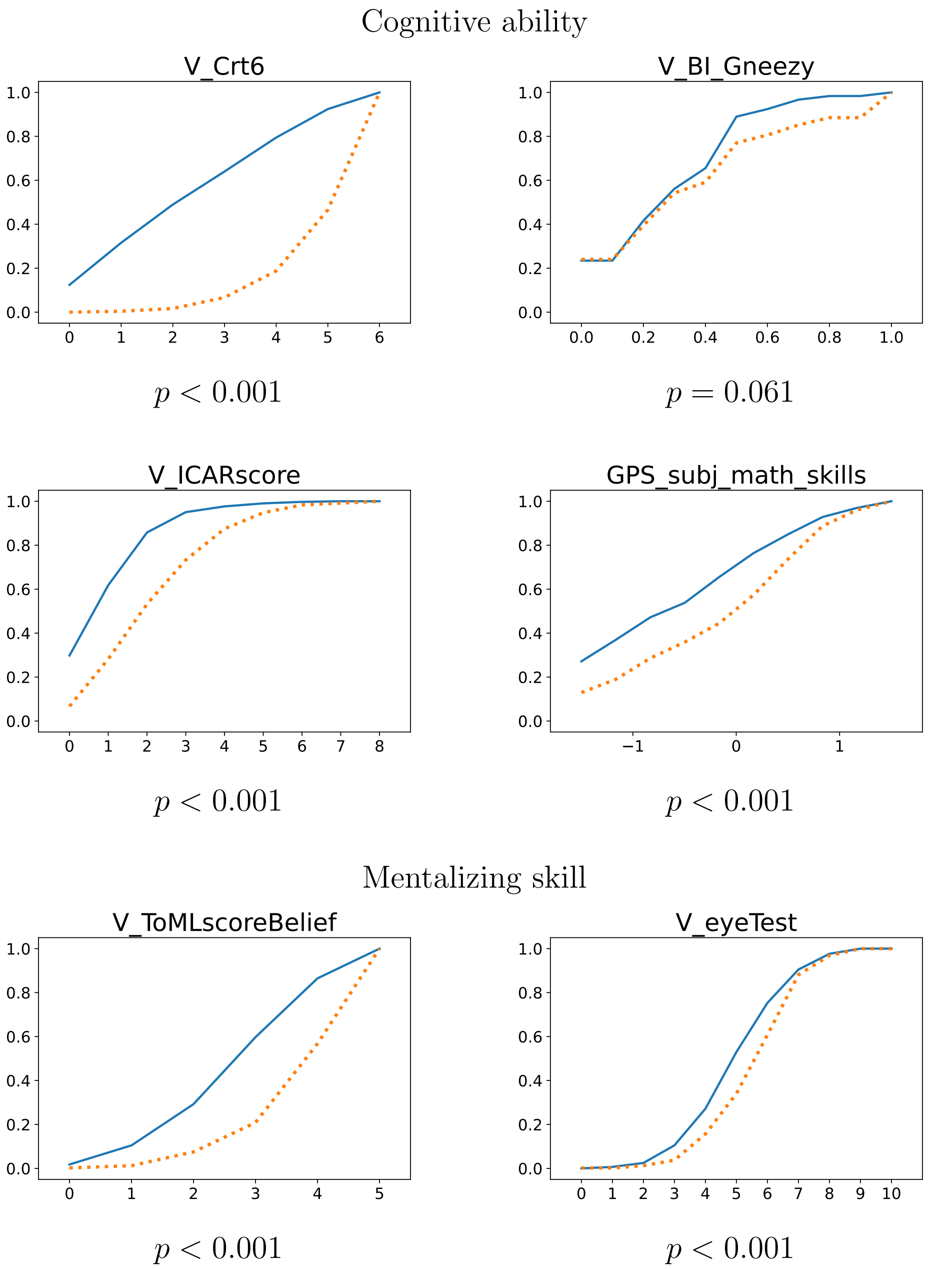
We group measures into those related to (i) skills (cognitive ability and mentalizing skills), (ii) preferences (risk, loss, ambiguity, time, and distributional), and (iii) personality traits. We use a 5% significance level in examining the difference between the two samples.

Because Osaka University is one of the most selective universities in Japan, we expect the measured cognitive ability of the OU sample to be significantly higher than that of the nonstudent sample. Based on the finding by Dohmen et al. (2010) regarding positive correlations between calculated risk taking as well as patience and cognitive ability, we also expect the OU sample to be significantly more risk taking and patient than the nonstudent sample.

4.1 Skills: Cognitive Ability and Mentalizing Skills

Figure 4 shows the distribution of the four cognitive ability measures for two samples. P-values from the two-sample Mann–Whitney tests are reported below each panel. We see that the measured cognitive abilities, except for the ability of BI, are significantly higher for the OU sample than for the nonstudent sample, which is as expected. For the two measures of mentalizing skills, the OU sample demonstrates significantly higher mentalizing skills than the nonstudent sample.

Figure 4: Distribution of measures of cognitive ability and mentalizing skills



P-values are based on the two-sample Mann–Whitney test.
The OU sample is indicated by the orange-dashed line and the nonstudent sample by the blue-solid line.

4.2 Preferences

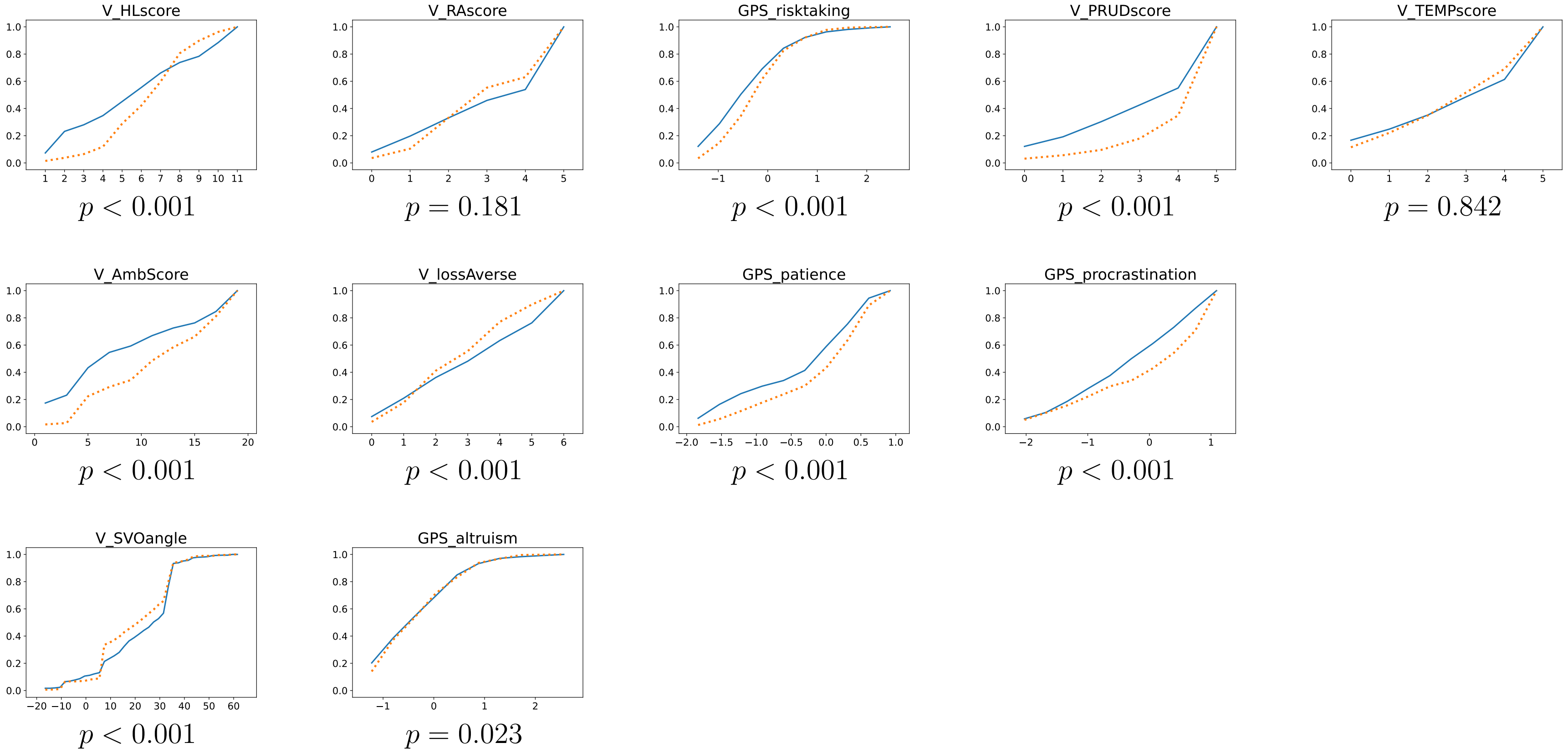
Figure 5 shows the distributions of the measures related to preferences for risk (and higher order risk), ambiguity, loss, and time. Among the three risk preference measures, the `V_RAscore` does not differ significantly between the OU sample and the nonstudent sample. It is possible that this measure, which is based on the number of safe choices in five binary choice tasks, may be too coarse. However, the remaining two measures show contradictory results. While `V_HLscore` shows that the OU sample is significantly more risk averse than the nonstudent sample, `GPS_risktaking`, which is the degree of *risk taking*, shows the opposite.

Figure 5 also shows that the OU sample is significantly more prudent (`V_RRUDscore`) and ambiguity averse (`V_AmbScore`), as well as significantly less loss averse (`V_lossAverse`) than the nonstudent sample. The degree of temperance (`V_TEMPscore`) is not significantly different between the two samples.

The distribution of the degree of patience as well as the tendency to procrastinate are both based on GPS. Interestingly, the OU sample is significantly more patient (`GPS_patience`) but has a greater tendency to procrastinate (`GPS_procrastination`) than the nonstudent sample.

The result of SVO angle (`V_SVOangle`) shows that the OU sample is significantly less prosocial than the nonstudent sample. At the same time, the OU sample demonstrates a slightly higher, but statistically significant, degree of altruism (mean $= 0.05$) than the nonstudent sample (mean $= -0.04$) based on `GPS_altruism`.

Figure 5: Distribution of the measures of preferences



$p < 0.001$ $p = 0.181$ $p < 0.001$ $p < 0.001$ $p = 0.842$

$p < 0.001$ $p < 0.001$ $p < 0.001$ $p < 0.001$

$p < 0.001$ $p = 0.023$

P-values are based on the two-sample Mann–Whitney test.
The OU sample is indicated by the orange-dashed line and the nonstudent sample by the blue-solid line.

4.3 Personality Traits

Figure 6 shows the distribution of trust, grit, positive and negative reciprocity, the Big Five personality traits, and overconfidence (overestimation and overplacement). The two measures of "general" trust provide opposing results. While according to `V_GTscale`, the OU sample is significantly more trusting than the nonstudent sample, the opposite is true based on `GPS_trust`. There are no statistically significant differences for the measure `V_grit`. In terms of reciprocity, the OU sample is significantly more reciprocal in the positive sense (`GPS_posrecip`) and significantly less reciprocal in the negative sense (`GPS_negrecip`) than the nonstudent sample.

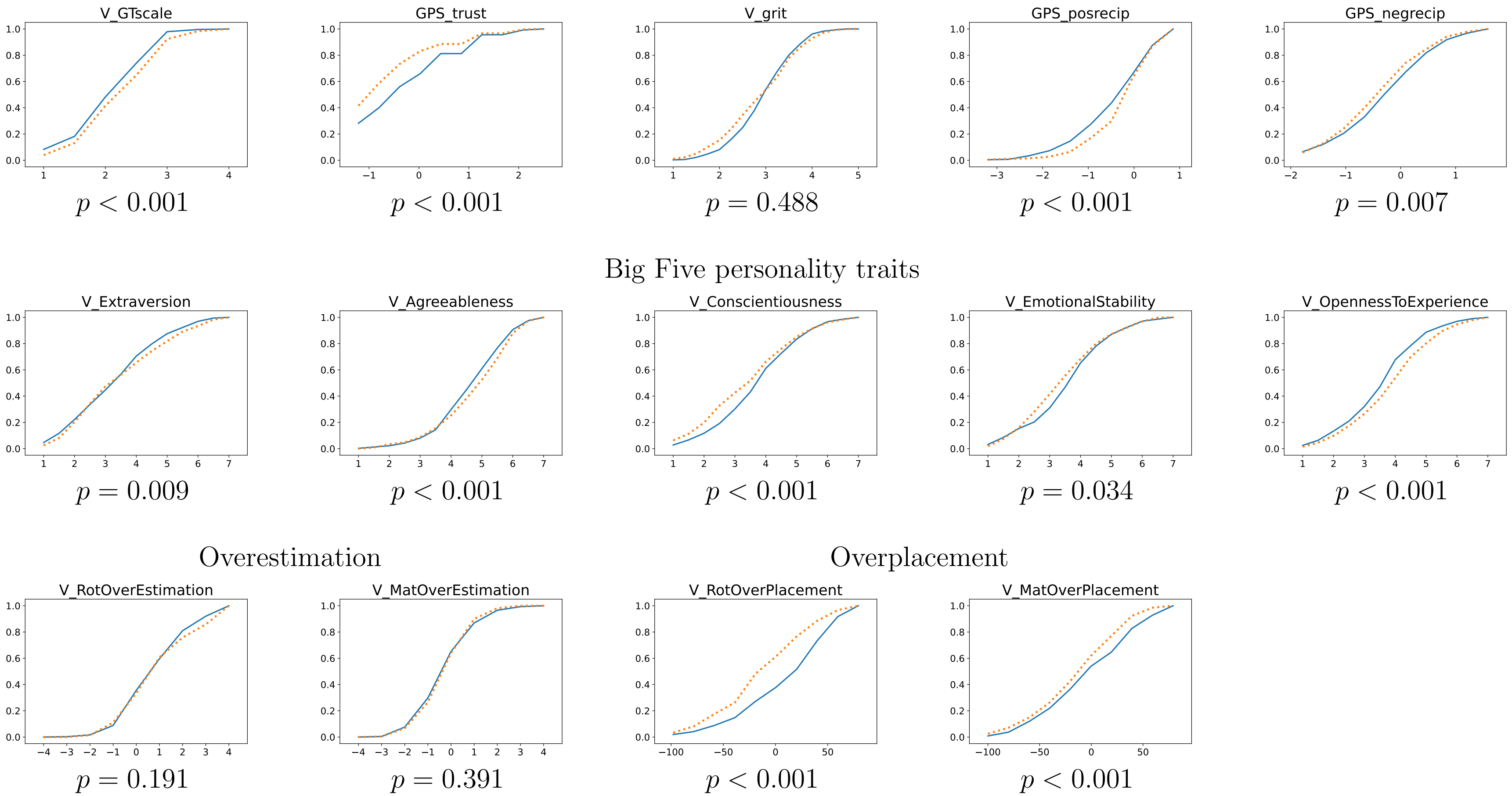
Among the Big Five personality traits, we observe that while the OU sample is significantly more extravert, agreeable, and open to experience than the nonstudent sample, they are also less conscientious and emotionally stable than the nonstudent sample.

It is interesting to note that there is no statistically significant difference between the two measures of overestimation between the two samples, and the nonstudent sample displays significantly stronger overplacement than the OU sample.

4.4 Correlation between Variables

In this subsection, we report the correlations between measured characteristics. We restrict our attention to those participants for whom we have data for all the measured characteristics (i.e., 526 from the OU sample and 1023 from the nonstudent sample). Because some of the measures in GPS are

Figure 6: Distribution of measures of personality traits



V_GTscale $p < 0.001$

GPS_trust $p < 0.001$

V_grit $p = 0.488$

GPS_posrecip $p < 0.001$

GPS_negrecip $p = 0.007$

Big Five personality traits

V_Extraversion $p = 0.009$

V_Agreeableness $p < 0.001$

V_Conscientiousness $p < 0.001$

V_EmotionalStability $p = 0.034$

V_OpennessToExperience $p < 0.001$

Overestimation

V_RotOverEstimation $p = 0.191$

V_MatOverEstimation $p = 0.391$

Overplacement

V_RotOverPlacement $p < 0.001$

V_MatOverPlacement $p < 0.001$

P-values are based on the two-sample Mann–Whitney test.
OU sample is orange-dashed. The nonstudent sample is blue-solid.

not available for those participants who answered "don't know" or "refuse to answer," the number of observations is further reduced to 479 for the OU sample and 725 for the nonstudent sample.[16]

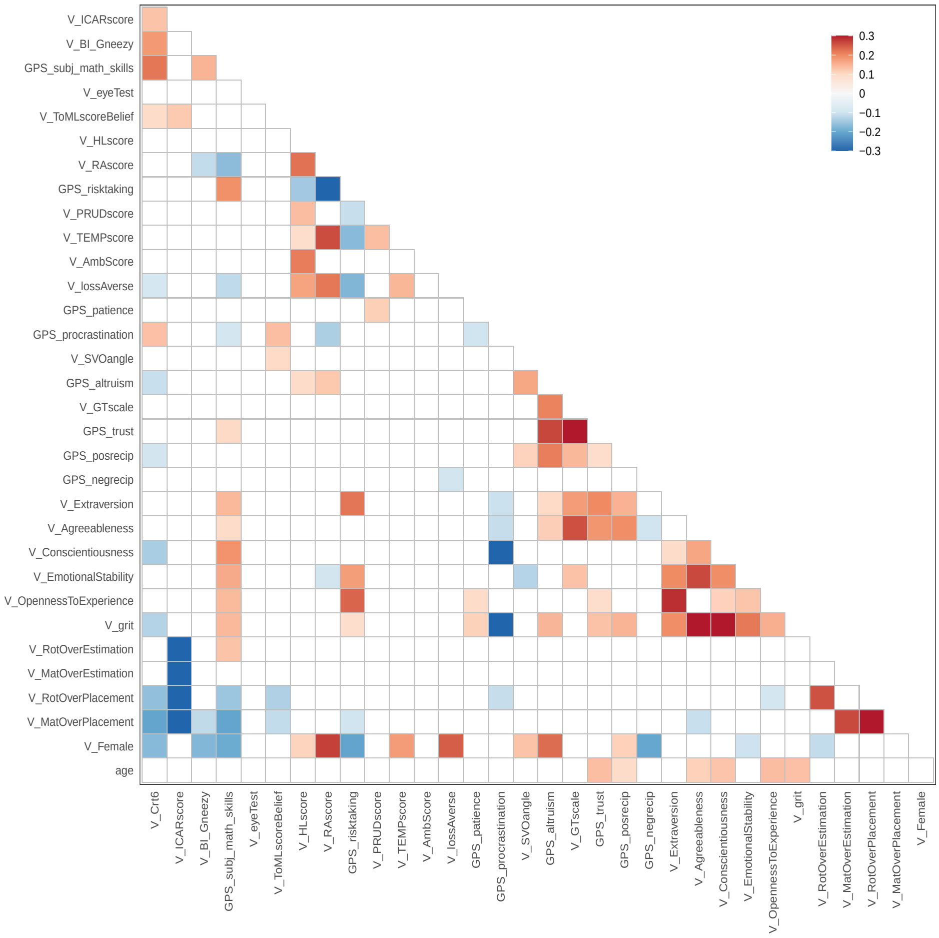
Figures 7 and 8 show the results of pairwise correlation analyses (Spearman correlation). Only those pairs with statistical significance at the 5% level are colored.[17] We observe reasonable correlation between measures within the same category (i.e., cognitive ability, risk-related preferences, time-related preferences, distributional preferences, personality, and overconfidence). Furthermore, we corroborate the existing literature regarding the correlation between sex and risk preferences, and the correlation between sex and prosociality, namely females are more risk averse (Booth and Nolen, 2012; Charness and Gneezy, 2012; Booth et al., 2014; Filippin, 2016) and more prosocial (Andreoni and Vesterlund, 2001; Kamas and Preston, 2015; Balafoutas et al., 2012).

We now discuss the correlations between measured cognitive ability (`V_Crt6` and `V_ICARscore`) and various individual characteristics. First, we have seen that while there is no clear relationship in terms of the degree of risk taking between the two samples, the OU sample is significantly more patient. Similarly, at the individual level, the correlations between the measures of cognitive ability and degree of risk taking (`V_HLscore`, `V_RAscore`, and `GPS_risktaking`) are mostly insignificant. The correlation between cognitive

[16] We did not implement multiple elicitations of the same task to address the measurement error problems by using the obviously related instrumental variables (ORIV) approach proposed by Gillen et al. (2019) and employed by Chapman et al. (2018) and Snowberg and Yariv (2021). Thus, our estimated correlations are biased downward.

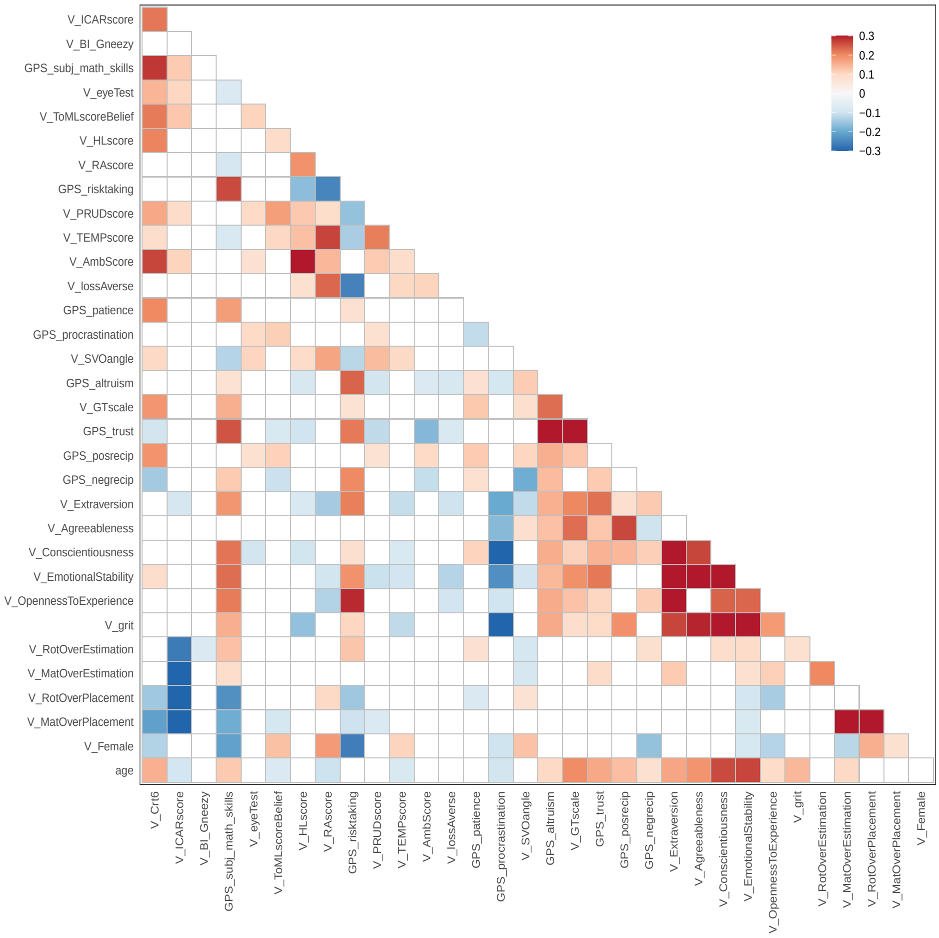
[17] Tables E.1 and E.2 in Appendix E show the corresponding values of (statistically significant) correlation coefficients.

Figure 7: Pairwise correlation between measures for the OU sample



Only those values that are significant at the 5% level are colored.

Figure 8: Pairwise correlation between measures for the nonstudent sample



Only those values that are significant at the 5% level are colored.

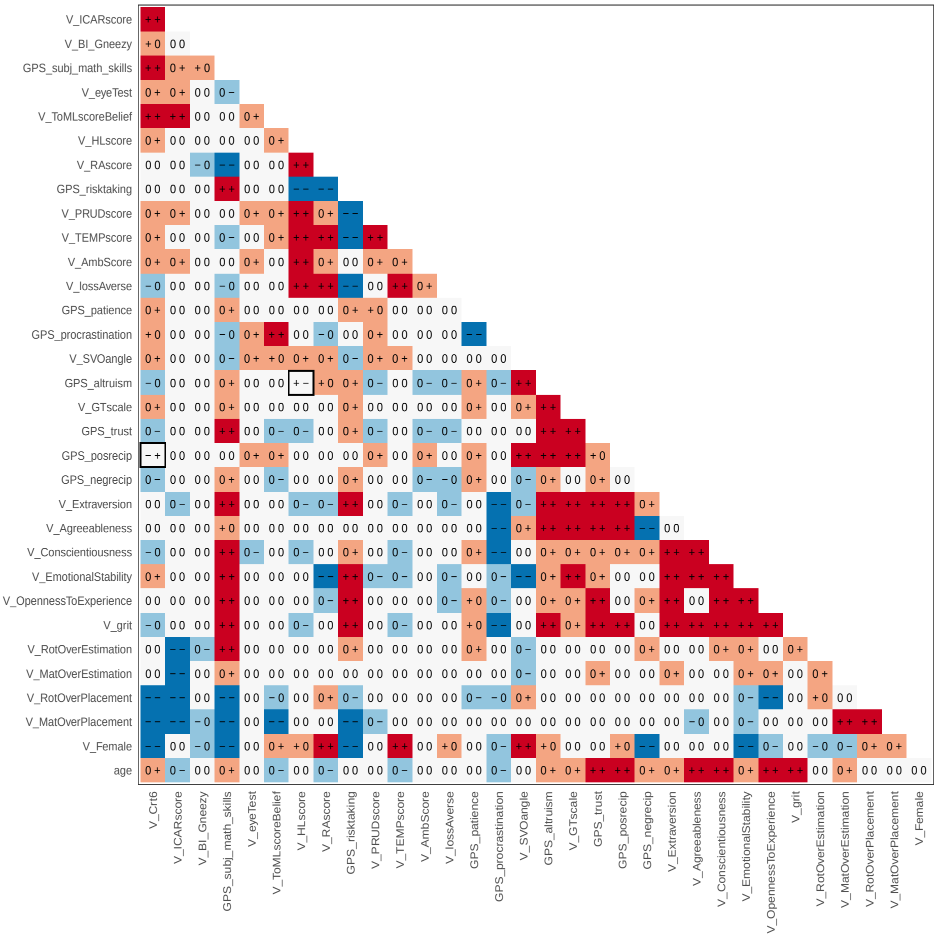
ability and `GPS_patience` is also weak, contrary to the sample comparison, within each sample; the only significant correlation is between `V_Crt6` and `GPS_patience` for the nonstudent sample. Interestingly, while the cognitive ability measures are not correlated with most of the personality traits, they are negatively correlated with our measures of overconfidence, and especially with overplacement.

Let us compare the correlation structure between the two samples. Figure 9 shows the results of comparing pairwise correlations between various measures in the two samples. We have category pairwise correlations depending on whether the results of the two samples have the same sign. As shown, many of the pairwise correlations are insignificant in both samples (shown as 00). There are also many pairwise correlations that are significant in both samples (shown as either ++ or $--$). There are only two pairwise correlations in which the two samples disagree (shown as either $+-$ or $-+$, and highlighted with bold squares). Thus, we conclude that, in terms of the correlational structure of the measures considered, the two samples generally show similar results.

5 Summary and Conclusion

In this paper, we report the results of large-scale online surveys (both incentivized and nonincentivized) that measure cognitive ability, mentalizing skills, preferences related to risk, time, and inequality, as well as personality traits of a large sample of students at Osaka University and a sample of the Japanese adult population ($20 \leq \text{age} < 70$) registered in the panel of an

Figure 9: Comparison of pairwise correlation between measures for the OU sample and the nonstudent sample



	V_Crt6	V_ICARscore	V_BI_Gneezy	GPS_subj_math_skills	V_eyeTest	V_ToMLscoreBelief	V_HLscore	V_RAscore	GPS_risktaking	V_PRUDscore	V_TEMPscore	V_AmbScore	V_lossAverse	GPS_patience	GPS_procrastination	V_SVOangle	GPS_altruism	V_GTscale	GPS_trust	GPS_posrecip	GPS_negrecip	V_Extraversion	V_Agreeableness	V_Conscientiousness	V_EmotionalStability	V_OpennessToExperience	V_grit	V_RotOverEstimation	V_MatOverEstimation	V_RotOverPlacement	V_MatOverPlacement	V_Female
V_ICARscore	++																															
V_BI_Gneezy	+0	00																														
GPS_subj_math_skills	++	0+	+0																													
V_eyeTest	0+	0+	00	0−																												
V_ToMLscoreBelief	++	++	00	00	0+																											
V_HLscore	0+	00	00	00	00	0+																										
V_RAscore	00	00	−0	−−	00	00	++																									
GPS_risktaking	00	00	00	++	00	00	−−	−−																								
V_PRUDscore	0+	0+	00	00	0+	0+	++	0+	−−																							
V_TEMPscore	0+	00	00	0−	00	0+	++	++	−−	++																						
V_AmbScore	0+	0+	00	00	0+	00	++	0+	00	0+	0+																					
V_lossAverse	−0	00	00	−0	00	00	++	++	−−	00	++	0+																				
GPS_patience	0+	00	00	0+	00	00	00	00	0+	+0	00	00	00																			
GPS_procrastination	+0	00	00	−0	0+	++	00	−0	00	0+	00	00	00	−−																		
V_SVOangle	0+	00	00	0−	0+	+0	0+	0+	0−	0+	0+	00	00	00	00																	
GPS_altruism	−0	00	00	0+	00	00	+−	+0	0+	0−	00	0−	0−	0+	0−	++																
V_GTscale	0+	00	00	0+	00	00	00	00	0+	00	00	00	00	0+	00	0+	++															
GPS_trust	0−	00	00	++	00	0−	0−	00	0+	0−	00	0−	0−	00	00	00	++	++														
GPS_posrecip	−+	00	00	00	0+	0+	00	00	00	0+	00	0+	00	0+	00	++	++	++	+0													
GPS_negrecip	0−	00	00	0+	00	0−	00	00	0+	00	00	0−	−0	0+	00	0−	0+	00	0+	00												
V_Extraversion	00	0−	00	++	00	00	0−	0−	++	00	0−	00	0−	00	−−	0−	++	++	++	++	0+											
V_Agreeableness	00	00	00	+0	00	00	00	00	00	00	00	00	00	00	−−	0+	++	++	++	++	−−	00										
V_Conscientiousness	−0	00	00	++	0−	00	0−	00	0+	00	0−	00	00	0+	−−	00	0+	0+	0+	0+	0+	++	++									
V_EmotionalStability	0+	00	00	++	00	00	00	−−	++	0−	0−	00	0−	00	0−	−−	0+	++	0+	00	00	++	++	++								
V_OpennessToExperience	00	00	00	++	00	00	00	0−	++	00	00	00	0−	+0	0−	00	0+	0+	++	00	0+	++	00	++	++							
V_grit	−0	00	00	++	00	00	0−	00	++	00	0−	00	00	+0	−−	00	++	0+	++	++	00	++	++	++	++	++						
V_RotOverEstimation	00	−−	0−	++	00	00	00	00	0+	00	00	00	00	0+	00	0−	00	00	00	00	0+	00	00	0+	0+	00	0+					
V_MatOverEstimation	00	−−	00	0+	00	00	00	00	00	00	00	00	00	00	00	0−	00	00	0+	00	00	0+	00	00	0+	0+	00	0+				
V_RotOverPlacement	−−	−−	00	−−	00	−0	00	0+	0−	00	00	00	00	0−	−0	0+	00	00	00	00	00	00	00	00	0−	−−	00	+0	00			
V_MatOverPlacement	−−	−−	−0	−−	00	−−	00	00	−−	0−	00	00	00	00	00	00	00	00	00	00	00	00	−0	00	0−	00	00	00	++	++		
V_Female	−−	00	−0	−−	00	0+	+0	++	−−	00	++	00	+0	00	0−	++	+0	00	00	+0	−−	00	00	00	−−	0−	00	−0	0−	0+	0+	
age	0+	0−	00	0+	00	0−	00	0−	00	00	0−	00	00	00	0−	00	0+	0+	++	++	0+	0+	++	++	0+	++	++	00	0+	00	00	00

++ :	both the OU and nonstudent samples show significant positive correlation
+0 (0+):	while the OU sample shows significant positive (no significant) correlation, the nonstudent sample shows no significant (significant positive) correlation
+− (−+):	while the OU sample shows significant positive (negative) correlation, the nonstudent sample shows significant negative (positive) correlation
00:	neither the OU sample nor the nonstudent sample shows significant correlation
−0 (0−):	while the OU sample shows significant negative (no significant) correlation, the nonstudent sample shows no significant (significant negative) correlation
−− :	both the OU and nonstudent samples show significant negative correlation

online survey company.

While significant differences between the two samples in many of these characteristics are observed, the correlational structures of these characteristics are very similar in the two samples. However, as noted, although these results are similar to previous findings Snowberg and Yariv (2021), caution is required in generalizing these results from an experiment at OU to the Japanese population in general. Nevertheless, as far as the correlations between individual characteristics are concerned, the external validity of the experimental results obtained at OU, in particular the representativeness of the sampled population, does not appear to be a major concern.

However, the tasks included in our online surveys are mostly individual and do not involve explicit interactions between participants. Future research should compare the results of interactive tasks between these subject pools.

References

Akiyama, E., N. Hanaki, and R. Ishikawa (2017): "It is not just confusion! Strategic uncertainty in an experimental asset market," *Economic Journal*, 127, F563–F580.

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

Balafoutas, L., R. Kerschbamer, and M. Sutter (2012): "Distributional preferences and competitive behavior," *Journal of Economic Behavior & Organization*, 83, 125–135.

Baron-Cohen, S., T. Jolliffe, C. Mortimore, and M. Robertson (1997): "Another Advanced Test of Theory of Mind: Evidence from Very High Functioning Adults with Autism or Asperger Syndrome," *Journal of Child Psychology and Psychiatry*, 38, 813–822.

Baron-Cohen, S., S. Wheelwright, J. Hill, Y. Raste, and I. Plumb (2001): "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*, 42, 241–251.

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

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

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

Bosch-Rosa, C., T. Meissner, and A. Bosch-Domènech (2018): "Cognitive Bubbles," *Experimental Economics*, 21, 132–153, doi:10.1007/s10683-017-9529-0.

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

Camerer, C. F., A. Dreber, T. 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 (2016): "Evaluating Replicability of Laboratory Experiments in Economics," *Sicence*, 351, 1433–1436.

Carpenter, J., M. Graham, and J. Wolf (2013): "Cognitive ability and strategic sophistication," *Games and Economic Behavior*, 80, 115–130.

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

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

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

Crosetto, P., O. Weisel, and F. Winter (2019): "A flexible z-Tree and oTree implementation of the Social Value Orientation Slider Measure," *Journal of Behavioral and Experimental Finance*, 23, 46–53.

Dodell-Feder, D., J. Koster-Hale, M. Bedny, and R. Saxe (2011): "fMRI item analysis in a theory of mind task," *NeuroImage*, 55, 705–712.

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

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

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

Duckworth, A. L. and P. D. Quinn (2009): "Development and Validation of the Short Grit Scale (Grit-S)," *Journal of Personality Assessment*, 91, 166–174, pMID: 19205937.

Dufwenberg, M., R. Sundaram, and D. J. Butler (2010): "Epiphany in the Game of 21," *Journal of Economic Behavior & Organization*, 75, 132–143.

Falk, A., A. Becker, T. Dohmen, B. Enke, D. Huffman, and U. Sunde (2016): "The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences," Discussion Paper 9674, IZA.

——— (2018): "Global evidence on economic preferences," *Quarterly Journal of Economics*, 133, 1645–1692.

Filippin, A. (2016): "Gender differences in risk attitudes," *IZA World of Labor*, ., .

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

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

Gächter, S., B. Herrmann, and C. Thöni (2010): "Culture and Cooperation," *Philosophical Transactions of The Royal Society B*, 365, 2651–2661.

Gangadharan, L., T. Jain, P. Maitra, and J. Vecci (2021): "Lab-in-the-field experiments: perspectives from research on gender," *Japanese Economic Review*, forthcoming.

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.

Gillen, B., E. Snowberg, and L. Yariv (2019): "Experimenting with measurement error: Techniques with applications to the Caltech Cohort Study," *Journal of Political Economy*, 127, 1826–1863.

Gneezy, U., A. Imas, and J. List (2015): "Estimating individual ambiguity aversion: A simple approach," Working paper 20982, National Bureau of Economic Research.

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

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

Hanaki, N. (2020): "Cognitive ability and observed behavior in laboratory experiments: implications for macroeconomic theory," *Japanese Economic Review*, 71, 355–378.

Hanaki, N., K. Inukai, T. Masuda, and Y. Shimodaira (2020): "Participants' characteristics at ISER-Lab in 2020," Discussion Paper 1141, Institute of Social and Economic Research, Osaka University.

Hanaki, N., N. Jacquemet, S. Luchini, and A. Zylbersztejn (2016): "Cognitive ability and the effect of strategic uncertainty," *Theory and Decision*, 81, 101–121.

Harrison, G. W. and J. A. List (2004): "Field experiments," *Journal of Economic Literature*, 42, 1009–1055.

Herrmann, B., C. Thöni, and S. Gächter (2008): "Antisocial Punishment Across Societies," *Science*, 319, 1362–1367.

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

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

List, J. A. (2007): "Field experiments: A bridge between lab and naturally occuring data," *The B.E. Journal of Economic Analysis & Policy*, 6, 0000102202153806371747.

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

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

Murphy, R. O., K. A. Ackermann, and M. J. J. Handgraaf (2011): "Measuring Social Value Orientation," *Judgment and Decision Making*, 6, 771–781.

Noussair, C. N., S. T. Trautmann, and G. van de Kuilen (2014): "Higher Order Risk Attitudes, Demographics, and Financial Decisions," *The Review of Economic Studies*, 81, 325–355.

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

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

Oshio, A., S. Abe, and P. Cutrone (2012): "Development, Reliability, and Validity of the Japanese Version of Ten Item Personality Inventory (TIPI-J)," *The Japanese Journal of Personality*, 21, 40–52.

Proto, E., A. Rustichini, and A. Sofianos (2019): "Intelligence, Personality and Gains from Cooperation in Repeated Interactions," *Journal of Political Economy*, 127, 1351–1390.

RAVEN, J. (2000): “The Raven’s Progressive Matrices: Change and Stability over Culture and Time,” *Cognitive Psychology*, 41, 1–48.

SMITH, V. L., G. L. SUCHANEK, AND A. W. WILLIAMS (1988): “Bubbles, Crashes, and Endogenous Expectations in Experimental Spot Asset Markets,” *Econometrica*, 56, 1119–1151.

SNOWBERG, E. AND L. YARIV (2021): “Testing the Waters: Behavior across participant pools,” *American Economic Review*, 111, 687–719.

TOPLAK, M. E., R. F. WEST, AND K. E. STANOVICH (2014): “Assessing miserly information processing: An expansion of the Cognitive Reflection Test,” *Thinking & Reasoning*, 20, 147–168.

WELZEL, C. (2010): “How Selfish Are Self-Expression Values? A Civicness Test,” *Journal of Cross-Cultural Psychology*, 41, 152–174.

YAMADA, M. AND T. MURAI (2005): “Seijinyou ‘Me kara Kokoro o Yomu Test’ Kaiteiban (Nihongoban) (The Revised Version of the Adult ‘Reading the Mind in the Eyes’ Test (Japanese version)),” `https://www.autismresearchcentre.com/tests/eyes-test-adult/`.

YAMAGISHI, T. AND M. YAMAGISHI (1994): “Trust and commitment in the United States and Japan,” *Motivation and Emotion*, 18, 129–166.

A Instructional Manipulation Checks

The IMCs are lengthy and thus participants may be somewhat hesitant to read them all. At the end of the statement, there are some questions. We implemented several IMCs because our online experiments were divided into several blocks.

In Block 1, the questions that follow the long description use a Likert scale. The statement asks the participant to ignore these questions, namely, not to answer the question. We used the Japanese translated version by Miura and Kobayashi (2015). We named the indicator of IMC success `V_good_W1`, and `V_good_W1` $= 1$ means that the check was successfully completed.

In Block 2, a question similar to the temperance measurement task was presented. After making their choices, participants were asked to unselect their choices. We named the indicator of IMC success `V_good_W2`, and `V_good_W2` $= 1$ means that the check was successfully completed.

In Block 3, a question similar to the SVO question was presented. Participants were asked to select all the options. We named the indicator of IMC success `V_good_W3`, and `V_good_W3` $= 1$ means that the check was successfully completed.

In Block 4, after questions regarding positive and negative reciprocity, subjective math skill, and tendency to procrastinate (WP13422–14326 of Falk et al., 2018), a task with an instruction similar to that used by Miura and Kobayashi (2015) with a choice set similar to those in WP13422–14326 was presented in the context of risk-taking. After making their choices, participants needed to unselect their choices by clicking the instruction. We named the indicator of IMC success `V_good_W4`, and `V_good_W4` = 1 means that the check was successfully completed.

B Questions for CRT

Questions 1 to 3 are from Finucane and Gullion (2010), and questions 4 to 6 are from Toplak et al. (2014).

1. If it takes 2 nurses 2 minutes to measure the blood pressure of 2 patients, how long would it take 200 nurses to measure the blood pressure of 200 patients? ___ minutes. [Correct answer: two minutes; intuitive answer: 200 minutes]

2. Soup and salad cost 5.50 euros in total. The soup costs 5 euros more than the salad. How much does the salad cost? ___ (in euros). [Correct answer: 0.25 euros; intuitive answer: 0.5 euros]

3. Sally is making sun tea. Every hour, the concentration of the tea doubles. If it takes 6 hours for the tea to be ready, how long would it take for the tea to reach half of the final concentration? ___ hours. [Correct answer: 5 hours; intuitive answer: 3 hours]

4. If John can drink one barrel of water in 6 days, and Mary can drink one barrel of water in 12 days, how long would it take them to drink one barrel of water together? ___ days. [correct answer: 4 days; intuitive answer: 9]

5. A man buys a pig for 60 euros, sells it for 70 euros, buys it back for 80 euros, and sells it finally for 90 euros. How much has he made? ___ euros. [correct answer: 20 euros; intuitive answer: 10 euros]

6. Simon decided to invest 8,000 euros in the stock market one day early in 2008. Six months after he invested, on July 17, the stocks he had purchased were down 50%. Fortunately for Simon, from July 17 to October 17, the stocks he had purchased went up 75%. At this point, Simon has:

   a. broken even in the stock market

   b. is ahead of where he began

   c. has lost money

   [correct answer: c, because the value at this point is 7,000 euros; intuitive response b].

Table C.1: Age-Sex composition

OU sample (All waves)

	10s	20s	30s	40s	50s	60s	70s	
Male	90	458	13	1	1	0	0	563
Female	53	325	9	2	2	0	0	391

Nonstudent sample

First session: targeted number of response 1800

	10s	20s	30s	40s	50s	60s	70s	total
Male	0	154	167	217	208	181	4	932
Female	1	144	156	210	208	187	1	906

Second session: targeted number of response 1000

	10s	20s	30s	40s	50s	60s	70s	
Male	0	84	95	123	115	98	3	518
Female	0	79	85	117	114	102	0	497

C Summary of Demographics

Table C.1 shows the age-sex composition of OU-sample and our nonstudent sample. Although we observe a few participants of the non-student sample have answered their ages that do not match the way company has recruited the participants, over all the age-sex composition of the participants are very similar to our target.

Tables C.2 to C.5 show the distribution of educational background, occupation, household annual income, and the residence (prefecture) of our non-student sample. Because we have conducted our online survey over two sessions, we report the distributions for each wave. Please recall that the participants to the second session is a subset of the participants to the first session.

Table C.2: Education: nonstudent sample

	First session		Second session	
Category	Male	Female	Male	Female
Primary or Junior High	13	5	5	2
High school dropout	22	14	9	4
High school completed	248	274	143	146
2-year college dropout	1	10	0	4
2-year college completed	66	216	41	122
4-year college dropout	44	14	27	9
4-year college completed	446	332	240	187
Master program dropout	1	1	9	1
Master program completed	56	16	34	10
Ph.D. program dropout	4	0	3	0
Ph.D. program completed	12	2	6	0
decline to answer	19	22	10	12

Table C.3: Occupation: nonstudent sample

	First session		Second session	
Category	Male	Female	Male	Female
Full time employee	459	202	248	115
Part time employee 1 (Paato)	28	148	13	81
Part time employee 2 (Arubaito)	37	39	18	27
Temporary employee (Haken)	14	17	9	6
Contract employee	53	25	22	13
Self employed	94	42	56	26
Executive	21	4	15	1
No job (inc. retired, housewife/husband)	161	373	99	204
Unemployed	20	4	12	2
Students	27	26	15	9
Others	11	11	7	3
decline to answer	7	15	4	10

Table C.4: Household annual income (Y): nonstudent sample

	First session		Second session	
Category*	Male	Female	Male	Female
$Y < 1$	58	40	24	22
$1 \leq Y < 2$	44	63	25	29
$2 \leq Y < 3$	104	90	62	45
$3 \leq Y < 4$	106	114	59	55
$4 \leq Y < 5$	91	93	45	56
$5 \leq Y < 6$	93	87	46	57
$6 \leq Y < 7$	66	55	38	34
$7 \leq Y < 8$	72	55	46	28
$8 \leq Y < 9$	48	45	26	19
$9 \leq Y < 10$	44	38	27	24
$10 \leq Y < 11$	27	28	14	16
$11 \leq Y < 12$	17	13	9	7
$12 \leq Y < 13$	17	5	13	3
$13 \leq Y < 14$	8	6	0	
$14 \leq Y < 15$	13	9	7	5
$15 \leq Y$	26	17	16	9
decline to answer	98	153	55	88

*: Y in million Yen.

Table C.5: Prefecture of residence: nonstudent sample

	First session		Second session			First session		Second session	
Prefecture	Male	Female	Male	Female	Prefecture	Male	Female	Male	Female
1. Hokkaido	45	53	23	23	25. Shiga	6	8	3	4
2. Aomori	5	3	0	0	26. Kyoto	16	27	9	20
3. Iwate	8	4	4	2	27. Osaka	78	78	42	49
4. Miyagi	20	19	11	10	28. Hyogo	37	38	23	20
5. Akita	5	3	4	1	29. Nara	3	6	3	5
6. Yamagata	5	7	2	5	30. Wakayama	8	4	5	1
7. Fukushima	9	14	7	7	31. Tottori	0	4	0	3
8. Ibaraki	16	18	7	11	32. Shimane	5	0	3	0
9. Tochigi	4	12	3	8	33. Okayama	14	12	9	5
10. Gunma	13	7	7	3	34. Hiroshima	18	23	9	11
11. Saitama	72	59	45	34	35. Yamaguchi	8	10	3	5
12. Chiba	46	48	31	25	36. Tokushima	4	2	2	0
13. Tokyo	164	141	90	75	37. Kagawa	6	3	2	3
14. Kanagawa	106	80	62	50	38. Ehime	8	10	5	6
15. Niigata	12	16	5	8	39. Kochi	3	5	2	3
16. Toyama	8	5	6	2	40. Fukuoka	31	39	16	21
17. Ishikawa	5	3	4	1	41. Saga	1	1	0	1
18. Fukui	6	2	3	1	42. Nagasaki	4	11	1	7
19. Yamanashi	6	1	3	1	43. Kumamoto	6	4	2	3
20. Nagano	11	9	5	6	44. Oita	6	1	2	1
21. Gifu	9	12	4	6	45. Miyazaki	3	8	2	2
22. Shizuoka	25	18	13	7	46. Kagoshima	7	5	6	1
23. Aichi	40	52	20	24	47. Okinawa	8	5	2	4
24. Mie	12	16	8	12	48. decline to answer	0	0	0	0

D Summary Statistics

Tables D.1 to D.4 report the mean and the standard deviations of characteristics measured in Blocks 1 to 4 for the OU and nonstudent samples. As noted above, the number of observations differs across the four blocks for the OU sample, and they differ between Blocks 1&4 and 2&3 for the nonstudent sample.

Table D.1: Mean (standard deviation). Measured in Block 1

	Full sample			Only V_good_1W= 1		
	OU	nonstudent	P-values	OU	nonstudent	P-values
V_GTscale	2.33	2.27	<0.001	2.46	2.33	0.012
	(0.66)	(0.62)		(0.65)	(0.62)	
V_eyeTest	5.94	5.35	<0.001	5.99	5.65	<0.001
	(1.46)	(1.62)		(1.44)	(1.43)	
V_Crt6	5.18	2.50	<0.001	5.28	3.30	<0.001
	(1.07)	(1.83)		(0.96)	(1.77)	
V_Extravesion	3.72	3.49	0.009	3.69	3.24	<0.001
	(1.53)	(1.37)		(1.54)	(1.39)	
V_Agreeableness	4.99	4.79	<0.001	4.98	4.86	0.084
	(1.19)	(1.11)		(1.19)	(1.17)	
V_Conscientiousness	3.53	3.89	<0.001	3.47	3.78	<0.001
	(1.46)	(1.29)		(1.45)	(1.34)	
V_EmotaionalStability	3.68	3.77	0.034	3.65	3.64	0.593
	(1.32)	(1.25)		(1.36)	(1.31)	
V_OpennessToExperience	4.14	3.70	<0.001	4.10	3.66	<0.001
	(1.31)	(1.19)		(1.32)	(1.26)	
V_HLscore	6.65	5.72	<0.001	6.88	6.57	0.241
	(2.20)	(3.31)		(2.00)	(2.95)	
V_AmbScore	12.78	9.40	<0.001	13.08	11.21	<0.001
	(5.78)	(6.59)		(5.68)	(6.43)	
V_good_1W	0.79	0.38	<0.001			
	(0.41)	(0.49)				
N	754	1023		592	387	

P-values are based on Mann-Whitney test, two-tailed.

Table D.2: Mean (standard deviation). Measured in Block 2

	Full sample			Only V_good_2W= 1		
	OU	nonstudent	P-values	OU	nonstudent	P-values
V_BI_Gneezy	0.39	0.34	0.061	0.39	0.36	0.217
	(0.31)	(0.24)		(0.32)	(0.27)	
V_RAscore	3.36	3.36	0.181	3.41	3.49	0.108
	(1.50)	(1.83)		(1.47)	(1.78)	
V_PRUDscore	4.27	3.30	<0.001	4.32	3.91	<0.001
	(1.26)	(1.84)		(1.23)	(1.63)	
V_TEMPscore	3.12	3.03	0.842	3.21	3.29	0.062
	(1.72)	(1.92)		(1.68)	(1.91)	
V_ToMLscoreBelief	4.02	3.06	<0.001	3.21	3.40	<0.001
	(1.01)	(1.22)		(1.68)	(1.16)	
V_grit	3.04	3.08	0.488	4.16	3.08	0.290
	(0.77)	(0.64)		(0.91)	(0.64)	
V_good_2W	0.77	0.22	<0.001			
	(0.42)	(0.41)				
N	762	1855		584	405	

P-values are based on Mann-Whitney test, two-tailed.

Table D.3: Mean (standard deviation). Measured in Block 3

	Full sample			Only V_good_3W= 1		
	OU	nonstudent	P-values	OU	nonstudent	P-values
V_rotScoreTot	0.86	0.29	<0.001	0.92	0.32	<0.001
	(1.09)	(0.64)		(1.13)	(0.71)	
V_matScoreTot	1.76	0.94	<0.001	1.81	1.10	<0.001
	(1.09)	(0.92)		(1.08)	(0.95)	
V_ICARscore	2.62	1.22	<0.001	2.73	1.42	<0.001
	(1.64)	(1.18)		(1.66)	(1.26)	
V_SVOangle	22.24	24.56	<0.001	22.42	26.40	<0.001
	(15.28)	(15.96)		(15.19)	(15.80)	
V_lossAverse	3.10	3.51	<0.001	3.10	3.66	<0.001
	(1.66)	(2.05)		(1.64)	(1.98)	
V_RotOverEstimation	1.24	1.13	0.191	1.17	1.17	0.830
	(1.59)	(1.45)		(1.58)	(1.50)	
V_RotOverPlacement	6.58	31.10	<0.001	4.97	31.54	<0.001
	(44.58)	(45.64)		(45.40)	(43.41)	
V_MatOverEstimation	0.18	0.16	0.391	0.15	0.03	0.056
	(1.12)	(1.25)		(1.12)	(1.16)	
V_MatOverPlacement	4.04	18.01	<0.001	2.45	14.12	<0.001
	(42.00)	(43.83)		(42.06)	(42.97)	
V_good_3W	0.72	0.36	<0.001			
	(0.45)	(0.48)				
N	745	1855		538	671	

P-values are based on Mann-Whitney test, two-tailed.

Table D.4: Mean (standard deviation). Measured in Block 4

	Full sample			Only V_good_4W= 1		
	OU	nonstudent	P-values	OU	nonstudent	P-values
GPS_patience	0.18	-0.13	<0.001	0.19	-0.10	<0.001
	(0.72)	(0.85)		(0.71)	(0.84)	
n	696	879		669	629	
GPS_risktaking	0.15	-0.11	<0.001	0.14	-0.20	<0.001
	(0.65)	(0.86)		(0.66)	(0.79)	
n	697	939		667	674	
GPS_posrecip	0.15	-0.97	<0.001	0.16	0.06	0.014
	(0.69)	(0.84)		(0.68)	(0.75)	
n	715	967		685	695	
GPS_negrecip	-0.05	0.05	0.007	-0.07	-0.02	0.212
	(0.82)	(0.86)		(0.82)	(0.88)	
n	695	873		665	616	
GPS_altruism	0.05	-0.04	0.023	0.03	-0.09	0.003
	(0.80)	(0.86)		(0.79)	(0.84)	
n	705	904		676	649	
GPS_trust	-0.20	0.15	<0.001	-0.22	-0.06	<0.001
	(0.95)	(1.01)		(0.94)	(0.94)	
n	713	978		682	697	
GPS_subj_math_skills	0.29	-0.21	<0.001	0.30	-0.27	<0.001
	(0.94)	(0.99)		(0.94)	(1.00)	
n	716	996		684	705	
GPS_procrastination	0.21	-0.15	<0.001	0.23	-0.15	<0.001
	(1.02)	(0.95)		(1.02)	(0.99)	
n	716	996		686	706	
V_good_4W	0.95	0.70	<0.001			
	(0.21)	(0.46)				
n	719	1023				

P-values are based on Mann-Whitney test, two-tailed.

E Correlation Coefficients for OU and Non-student Samples

Table E.1: Spearman correlation coefficients among measures for OU sample

		1	2	3	4	5	6	7	8	9	10	11
1	V_Crt6											
2	V_ICARscore	0.13										
3	V_BI_Gneezy	0.18	0									
4	GPS_subj_math_skills	0.22	0	0.15								
5	V_eyeTest	0	0	0	0							
6	V_ToMLscoreBelief	0.09	0.12	0	0	0						
7	V_HLscore	0	0	0	0	0	0					
8	V_RAscore	0	0	-0.11	-0.17	0	0	0.22				
9	GPS_risktaking	0	0	0	0.19	0	0	-0.15	-0.3			
10	V_PRUDscore	0	0	0	0	0	0	0.14	0	-0.11		
11	V_TEMPscore	0	0	0	0	0	0	0.09	0.26	-0.17	0.14	
12	V_AmbScore	0	0	0	0	0	0	0.21	0	0	0	0
13	V_lossAverse	-0.09	0	0	-0.12	0	0	0.17	0.22	-0.18	0	0.15
14	GPS_patience	0	0	0	0	0	0	0	0	0	0.12	0
15	GPS_procrastination	0.13	0	0	-0.09	0	0.14	0	-0.14	0	0	0
16	V_SVOangle	0	0	0	0	0	0.1	0	0	0	0	0
17	GPS_altruism	-0.11	0	0	0	0	0	0.1	0.12	0	0	0
18	V_GTscale	0	0	0	0	0	0	0	0	0	0	0
19	GPS_trust	0	0	0	0.1	0	0	0	0	0	0	0
20	GPS_posrecip	-0.09	0	0	0	0	0	0	0	0	0	0
21	GPS_negrecip	0	0	0	0	0	0	0	0	0	0	0
22	V_Extraversion	0	0	0	0.14	0	0	0	0	0.22	0	0
23	V_Agreeableness	0	0	0	0.09	0	0	0	0	0	0	0
24	V_Conscientiousness	-0.14	0	0	0.19	0	0	0	0	0	0	0
25	V_EmotionalStability	0	0	0	0.16	0	0	0	-0.1	0.18	0	0
26	V_OpennessToExperience	0	0	0	0.14	0	0	0	0	0.24	0	0
27	V_grit	-0.13	0	0	0.14	0	0	0	0	0.09	0	0
28	V_RotOverEstimation	0	-0.45	0	0.13	0	0	0	0	0	0	0
29	V_MatOverEstimation	0	-0.37	0	0	0	0	0	0	0	0	0
30	V_RotOverPlacement	-0.16	-0.62	0	-0.15	0	-0.13	0	0	0	0	0
31	V_MatOverPlacement	-0.21	-0.67	-0.12	-0.21	0	-0.12	0	0	-0.1	0	0
32	V_Female	-0.17	0	-0.18	-0.2	0	0	0.11	0.27	-0.21	0	0.18
33	age	0	0	0	0	0	0	0	0	0	0	0

		12	13	14	15	16	17	18	19	20	21	22
12	V_AmbScore											
13	V_lossAverse	0										
14	GPS_patience	0	0									
15	GPS_procrastination	0	0	-0.1								
16	V_SVOangle	0	0	0	0							
17	GPS_altruism	0	0	0	0	0.16						
18	V_GTscale	0	0	0	0	0	0.21					
19	GPS_trust	0	0	0	0	0	0.27	0.4				
20	GPS_posrecip	0	0	0	0	0.11	0.21	0.14	0.09			
21	GPS_negrecip	0	-0.1	0	0	0	0	0	0	0		
22	V_Extraversion	0	0	0	-0.11	0	0.1	0.18	0.2	0.15	0	
23	V_Agreeableness	0	0	0	-0.11	0	0.12	0.26	0.19	0.19	-0.1	0
24	V_Conscientiousness	0	0	0	-0.5	0	0	0	0	0	0	0.1
25	V_EmotionalStability	0	0	0	0	-0.13	0	0.13	0	0	0	0.2
26	V_OpennessToExperience	0	0	0.1	0	0	0	0	0.09	0	0	0.29
27	V_grit	0	0	0.11	-0.45	0	0.15	0	0.13	0.15	0	0.2
28	V_RotOverEstimation	0	0	0	0	0	0	0	0	0	0	0
29	V_MatOverEstimation	0	0	0	0	0	0	0	0	0	0	0
30	V_RotOverPlacement	0	0	0	-0.11	0	0	0	0	0	0	0
31	V_MatOverPlacement	0	0	0	0	0	0	0	0	0	0	0
32	V_Female	0	0.24	0	0	0.13	0.23	0	0	0.11	-0.2	0
33	age	0	0	0	0	0	0	0	0.14	0.09	0	0

		23	24	25	26	27	28	29	30	31	32
23	V_Agreeableness										
24	V_Conscientiousness	0.17									
25	V_EmotionalStability	0.26	0.19								
26	V_OpennessToExperience	0	0.11	0.13							
27	V_grit	0.31	0.58	0.22	0.15						
28	V_RotOverEstimation	0	0	0	0	0					
29	V_MatOverEstimation	0	0	0	0	0	0				
30	V_RotOverPlacement	0	0	0	-0.09	0	0.26	0			
31	V_MatOverPlacement	-0.11	0	0	0	0	0	0.26	0.33		
32	V_Female	0	0	-0.1	0	0	-0.11	0	0	0	
33	age	0.11	0.13	0	0.14	0.13	0	0	0	0	0

Only those that are statistically significant at 5% are shown.

Table E.2: Spearman correlation coefficients among measures for nonstudent sample

		1	2	3	4	5	6	7	8	9	10	11
1	V_Crt6											
2	V_ICARscore	0.22										
3	V_BI_Gneezy	0	0									
4	GPS_subj_math_skills	0.28	0.12	0								
5	V_eyeTest	0.15	0.11	0	-0.07							
6	V_ToMLscoreBelief	0.22	0.12	0	0	0.11						
7	V_HLscore	0.21	0	0	0	0	0.09					
8	V_RAscore	0	0	0	-0.09	0	0	0.19				
9	GPS_risktaking	0	0	0	0.26	0	0	-0.17	-0.25			
10	V_PRUDscore	0.16	0.1	0	0	0.1	0.18	0.12	0.09	-0.16		
11	V_TEMPscore	0.09	0	0	-0.08	0	0.1	0.13	0.27	-0.14	0.21	
12	V_AmbScore	0.27	0.11	0	0	0.08	0	0.41	0.14	0	0.12	0.09
13	V_lossAverse	0	0	0	0	0	0	0.08	0.23	-0.26	0	0.1
14	GPS_patience	0.2	0	0	0.18	0	0	0	0	0.08	0	0
15	GPS_procrastination	0	0	0	0	0.1	0.12	0	0	0	0.08	0
16	V_SVOangle	0.1	0	0	-0.13	0.11	0	0.1	0.17	-0.12	0.14	0.1
17	GPS_altruism	0	0	0	0.08	0	0	-0.08	0	0.24	-0.09	0
18	V_GTscale	0.19	0	0	0.15	0	0	0	0	0.08	0	0
19	GPS_trust	-0.1	0	0	0.25	0	-0.08	-0.1	0	0.22	-0.12	0
20	GPS_posrecip	0.19	0	0	0	0.08	0.11	0	0	0	0.08	0
21	GPS_negrecip	-0.15	0	0	0.12	0	-0.11	0	0	0.2	0	0
22	V_Extraversion	0	-0.09	0	0.18	0	0	-0.08	-0.14	0.21	0	-0.11
23	V_Agreeableness	0	0	0	0	0	0	0	0	0	0	0
24	V_Conscientiousness	0	0	0	0.22	-0.09	0	-0.1	0	0.08	0	-0.08
25	V_EmotionalStability	0.09	0	0	0.23	0	0	0	-0.1	0.19	-0.11	-0.1
26	V_OpennessToExperience	0	0	0	0.21	0	0	0	-0.13	0.29	0	0
27	V_grit	0	0	0	0.15	0	0	-0.16	0	0.11	0	-0.12
28	V_RotOverEstimation	0	-0.27	-0.08	0.13	0	0	0	0	0.13	0	0
29	V_MatOverEstimation	0	-0.5	0	0.09	0	0	0	0	0	0	0
30	V_RotOverPlacement	-0.15	-0.36	0	-0.24	0	0	0	0.1	-0.15	0	0
31	V_MatOverPlacement	-0.21	-0.69	0	-0.19	0	-0.09	0	0	-0.1	-0.08	0
32	V_Female	-0.13	0	0	-0.21	0	0.13	0	0.18	-0.26	0	0.11
33	age	0.15	-0.09	0	0.12	0	-0.07	0	-0.1	0	0	-0.08

		12	13	14	15	16	17	18	19	20	21	22
12	V_AmbScore											
13	V_lossAverse	0.11										
14	GPS_patience	0	0									
15	GPS_procrastination	0	0	-0.12								
16	V_SVOangle	0	0	0	0							
17	GPS_altruism	-0.07	-0.08	0.08	-0.09	0.12						
18	V_GTscale	0	0	0.12	0	0.09	0.23					
19	GPS_trust	-0.17	-0.08	0	0	0	0.32	0.36				
20	GPS_posrecip	0.1	0	0.12	0	0.1	0.15	0.13	0			
21	GPS_negrecip	-0.11	0	0.08	0	-0.19	0.14	0	0.12	0		
22	V_Extraversion	0	-0.1	0	-0.2	-0.12	0.15	0.2	0.23	0.09	0.12	
23	V_Agreeableness	0	0	0	-0.17	0.08	0.13	0.23	0.12	0.27	-0.1	0
24	V_Conscientiousness	0	0	0.11	-0.47	0	0.16	0.11	0.15	0.14	0.11	0.32
25	V_EmotionalStability	0	-0.13	0	-0.24	-0.1	0.14	0.19	0.22	0	0	0.31
26	V_OpennessToExperience	0	-0.1	0	-0.1	0	0.16	0.13	0.1	0	0.12	0.36
27	V_grit	0	0	0	-0.35	0	0.16	0.09	0.1	0.19	0	0.27
28	V_RotOverEstimation	0	0	0.08	0	-0.09	0	0	0	0	0.08	0
29	V_MatOverEstimation	0	0	0	0	-0.09	0	0	0.09	0	0	0.12
30	V_RotOverPlacement	0	0	-0.07	0	0.07	0	0	0	0	0	0
31	V_MatOverPlacement	0	0	0	0	0	0	0	0	0	0	0
32	V_Female	0	0	0	-0.1	0.13	0	0	0	0	-0.16	0
33	age	0	0	0	-0.1	0	0.1	0.2	0.16	0.14	0.08	0.17

		23	24	25	26	27	28	29	30	31	32
23	V_Agreeableness										
24	V_Conscientiousness	0.27									
25	V_EmotionalStability	0.31	0.45								
26	V_OpennessToExperience	0	0.24	0.24							
27	V_grit	0.29	0.51	0.32	0.18						
28	V_RotOverEstimation	0	0.09	0.09	0	0.08					
29	V_MatOverEstimation	0	0	0.08	0.11	0	0.2				
30	V_RotOverPlacement	0	0	-0.09	-0.14	0	0	0			
31	V_MatOverPlacement	0	0	-0.08	0	0	0	0.31	0.37		
32	V_Female	0	0	-0.09	-0.13	0	0	-0.12	0.15	0.08	
33	age	0.19	0.26	0.27	0.09	0.14	0	0.1	0	0	0

Only those that are statistically significant at 5% are shown.

F Comparison Across Three Sessions at Osaka University

As noted in the main text, the data at Osaka University were gathered during three sessions conducted over four months. Tables F.1 to F.4 report the means and standard deviations of the characteristics measured in Blocks 1 to 4, respectively, across three waves. For all the measures except for `V_good_3W` in Block 3, and `GPS_patience` and `V_good_4W` in Block 4, there are no statistically significant differences across the three waves at the 5% significance level.

Table F.1: Mean (standard deviation). Measured in Block 1

	Wave 1	Wave 2	Wave 3	P-values
V_GTscale	2.45	2.50	2.35	0.09
	(0.66)	(0.68)	(0.62)	
V_eyeTest	5.96	5,88	6.01	0.83
	(1.45)	(1.48)	(1.44)	
V_Crt6	5.25	5.21	4.98	0.09
	(1.03)	(0.97)	(1.27)	
V_Extravesion	3.73	3.74	3.68	0.95
	(1.54)	(1.52)	(1.51)	
V_Agreeableness	5.00	4.99	4.94	0.77
	(1.21)	(1.71)	(1.15)	
V_Conscientiousness	3.44	3.69	3.52	0.10
	(1.49)	(1.46)	(1.38)	
V_EmotaionalStability	3.65	3.65	3.78	0.59
	(1.35)	(1.28)	(1.29)	
V_OpennessToExperience	4.18	4.18	4.00	0.24
	(1.33)	(1.31)	(1.23)	
V_HLscore	6.65	6.86	6.36	0.09
	(2.12)	(2.32)	(2.20)	
V_AmbScore	12.68	12.34	13.64	0.08
	(5.73)	(6.09)	(5.40)	
V_good_1W	0.79	0.77	0.80	0.81
	(0.41)	(0.42)	(0.40)	
N	365	227	162	

P-values are based on Kruskal-Wallis test.

Table F.2: Mean (standard deviation). Measured in Block 2

	Wave 1	Wave 2	Wave 3	P-values
V_BI_Gneezy	0.41	0.36	0.35	0.07
	(0.32)	(0.32)	(0.28)	
V_RAscore	3.33	3.42	3.34	0.74
	(1.53)	(1.46)	(1.52)	
V_PRUDscore	4,28	4.22	4.30	0.69
	(1.17)	(1.4)	(1.24)	
V_TEMPscore	2.99	3.27	3.20	0.14
	(1.76)	(1.67)	(1.68)	
V_ToMLscoreBelief	4.08	3.97	3.97	0.34
	(0.99)	(1.03)	(1.04)	
V_grit	3.06	3.01	3.04	0.78
	(0.79)	(0.76)	(0.73)	
V_good_2W	0.77	0.73	0.80	0.20
	(0.42)	(0.45)	(0.40)	
N	366	217	179	

P-values are based on Kruskal-Wallis test.

Table F.3: Mean (standard deviation). Measured in Block 3

	Wave 1	Wave 2	Wave 3	P-values
V_rotScoreTot	0.90	0.88	0.73	0.18
	(1.11)	(1.10)	(1.03)	
V_matScoreTot	1.76	1.75	1.76	0.96
	(1.07)	(1.20)	(1.00)	
V_ICARscore	2.67	2.63	2.49	0.58
	(1.65)	(1.73)	(1.50)	
V_SVOangle	22.14	20.61	24.38	0.08
	(15.10)	(15.77)	(15.06)	
V_lossAverse	3.04	3.23	3.13	0.48
	(1.56)	(1.75)	(1.75)	
V_RotOverEstimation	1.19	1.25	1.35	0.45
	(1.60)	(1.61)	(1.52)	
V_RotOverPlacement	5.99	4.35	10.73	0.32
	(46.05)	(41.13)	(44.25)	
V_MatOverEstimation	0.19	0.18	0.12	0.75
	1.14	(1.15)	(1.05)	
V_MatOverPlacement	4.02	4.65	3.39	0.89
	40.59	(47.59)	(39.27)	
V_good_3W	0.75	0.64	0.74	0.03
	0.43	(0.48)	(0.44)	
N	419	173	153	

P-values are based on Kruskal-Wallis test.

Table F.4: Mean (standard deviation). Measured in Block 4

	Wave 1	Wave 2	Wave 3	P-values
GPS_patience	0.21	0.22	0.04	0.03
	(0.72)	(0.71)	(0.74)	
n	344	207	145	
GPS_risktaking	0.19	0.09	0.14	0.18
	(0.65)	(0.70)	(0.61)	
n	346	205	146	
GPS_posrecip	0.17	0.14	0.10	0.75
	(0.66)	(0.73)	(0.72)	
n	354	210	151	
GPS_negrecip	-0.08	-0.01	-0.04	0.84
	(0.84)	(0.82)	(0.78)	
n	346	205	144	
GPS_altruism	0.02	0.09	0.04	0.42
	(0.81)	(0.77)	(0.80)	
n	348	209	148	
GPS_trust	-0.22	-0.13	-0.26	0.34
	(0.96)	(0.95)	(0.93)	
n	351	211	151	
GPS_subj_math_skills	0.31	0.28	0.26	0.76
	(0.96)	(0.94)	(0.88)	
n	355	211	150	
GPS_procrastination	0.23	0.19	0.20	0.70
	(1.05)	(1.01)	(0.99)	
n	354	211	151	
V_good_4W	0.97	0.93	0.93	0.03
	(0.16)	(0.25)	(0.25)	
n	356	212	151	

P-values are based on Kruskal-Wallis test.