Interactive experiments in Toloka

Chapkovski, Philipp

Higher School of Economics

3 February 2022

Interactive experiments in Toloka

This is a preliminary draft.
Please do not cite without the permission of the author..

Philipp Chapkovski*

February 3, 2022

Abstract

The popularity of online behavioral experiments grew steadily even before the COVID-19 pandemic. With the start of lockdowns, online studies were often the only available option for the behavioral economists, sociologists and political scientists. The usage of most well-known platforms such as mTurk was so intensive that it harmed the quality of data. But even before the pandemics-induced quality crisis, online studies were limited in scope, since real-time interactions between participants were hard to achieve due to the large proportion of drop-outs and issues with creating stable groups. Using the crowdsourcing platform Toloka, we successfully ran several multi-round interactive experiments. Toloka's large online audience, relatively low exposure of participants to sociological surveys and behavioral studies, and a convenient application programming interface makes it a perfect tool to run behavioral studies that require real-time interactions of participants.

Keywords: Crowdsourcing, experiments, MTurk, online research, survey research
JEL classification: C90; C92; C81; C88; B41

*Philipp Chapkovski (fchapkovskiy@hse.ru) is a Research Fellow at the International Laboratory for Experimental and Behavioural Economics at the Higher School of Economics in Moscow.

1 Introduction

Many experimental designs that assume interdependence of decisions and payoffs, still allow for asynchronous decision-making. In one-shot games, in order to avoid synchronicity, experimentalists widely use the strategy method, in which a responder makes conditional decisions for each possible information set [1]. This method has proven to be quite reliable compared to the direct interaction of multiple players [2]. And yet numerous behavioral studies require interactions among their participants in real-time. Lab experiments on auctions [3], wage negotiations [4], voting patterns [5], and voluntary contributions with peer punishment [6] require the simultaneous presence of participants in a relatively stable group, where they can receive real-time feedback about their partners' behavior across time.

Despite the recent booming popularity of crowdsourcing platforms for conducting behavioral studies, so far the obstacles to running experiments with real-time interactions were too high to make them feasible. In a review of one crowdsourcing platform widely used for behavioral studies, Prolific, the authors mentioned that "[s]imultaneous experiments are currently not a focus of Prolific (and neither of other crowdworking platforms, for that matter)" [7].

The main problem with conducting interactive experiments using existing crowdsourcing platforms is the high drop-out rate. Arechar et al. [8] described their experience with replicating the study of a public good game with and without peer punishment conditions [6] using Amazon Mechanical Turk crowdsourcing platform (further on: mTurk). Over the course of that study, about 18% of the participants left the experiment before completing it. Each drop-out in turn affected the composition of the entire group and thus the behavior of his/her partners: in the end just 53% of the groups finished with all four members.

The drop-outs problem appears first at a matching phase: before proceeding to the decision stage, people need to be matched with other participants. If this stage takes too long, participants may tire of waiting and leave. This is particularly relevant when participants join a study at irregular intervals, in such a way that early and late arriving participants are at a particular risk of waiting too long for a match. Furthermore, compared to offline populations, mTurkers self-reportedly participate more often in surveys while multitasking, and they often leave a study page to return to it later[9]. Arechar and colleagues convincingly demonstrated that it is possible to conduct experiments using crowdsourcing platforms, when the research design requires real-time interaction in a stable group. Other studies have been conducted since Arechar's, such as [9], but it is clear that handling the high drop-out rate for behavioral games that span several rounds can be a daunting, if not insurmountable issue.

With an onset of the COVID-19 pandemic, however, running interactive experiments online became essential, as most on-campus labs were closed. Even before the disruption caused by COVID-19, there was an unmet demand for running interactive experiments using online audiences. Online studies provide access to audiences beyond typical white, educated, industrialized, rich, and democrat-leaning population attainable in the university pool [10]. For instance, mTurk online population is closer in their behavior to the general US population than students from large US universities with particularly salient differences regarding behavioral measures of honesty and altruism [11].

While COVID-19 boosted interest and demand for interactive experiments, it simultaneously exacerbated the problem with existing crowdsourcing platforms because of a substantial increase of new studies on the limited pool of participants, which decreased the quality of the data. The data quality crisis in mTurk started spreading even before the pandemic,caused by the easy accessibility of VPS (virtual private servers) that let non-US residents participate in US-only studies. By some estimates about 12% of the mTurk respondents are VPS users and the share of low-quality submissions among them about 8 times higher than among non-VPS users [12]. Even at the dawn of mTurk usage as a recruitment platform, researchers found that a small minority of workers were responsible for submitting most of the HITs among 132 academic studies [13]. Other estimates put a number of respondents realistically reachable in mTurk to less than 8 thousand (in stark contrast to the more than 500,000 formally registered mTurkers) [14]. The overexposure of this to the behavioral surveys and studies can bring in biased results. There has been proved effect of pool overexposure on trustfulness [15] and attention [16] of the respondents. In this paper we describe our experience of running interactive and non-interactive studies in a crowdsourcing platform Toloka, which obtains some features that may alleviate some problems with online research mentioned above.

2 General platform description

Toloka platform has gone relatively unnoticed in the academic community because its main focus is on providing workers for machine learning. However, its large and growing online population and a convenient programmatic access (API) to its resources make the integration with the existing experimental software, such as oTree [17], relatively straightforward. Toloka can be of interest to the academic community because it provides access to populations that have not been previously available via online crowdsourcing platforms. While the mTurk population is mostly based in the US and India [18], and Prolific is mostly in the UK and US [19], Toloka originally focused on Russia and other formerly Soviet countries. Although currently expanding to other countries, Russian-speaking participants still make up 75% of the active users depending on the time of the day.

Until Toloka, the post-Soviet space was unavailable for online studies. Furthermore, the lack of the behavioral labs there substantially limits the possibility for offline academic research. For example, we might ask whether the experience of living under state socialism make people from ex-socialist space behave differently as some behavioral studies suggest [20, 21]. But Toloka not only opens up ex-USSR for behaviorists, it also provides access to a large relatively inexperienced audience, with a limited knowledge about behavioral studies. The recent expansion of Toloka to other countries (such as India, Vietnam and Turkey) provides an additional incentive to use it for those looking beyond US and Western Europe population.

Another feature making Toloka distinguishable from its alternatives such as Prolific and mTurk is its vast online audience, ready to join a study in a matter of seconds. Although neither Prolific nor mTurk report the numbers of users currently active online, by some estimates the average online presence of mTurk population is about 2000 [18]. According to our measurements (Section 2.2), the Toloka interface provides access up to about 7 times more users than mTurk, varying from about 5,000 at night (in UTC timezone) to 22,000 at the peak of the working day. This impressive online presence let participants join studies fast, solving the issue of people stuck in the virtual waiting rooms while being matched with their partners.

2.1 Terminology

We intentionally avoid most of the technical details regarding the Toloka interface or its API functionality as the English documentation provided by Toloka is extensive. All the technical information regarding the connection of oTree and Toloka are available in the *readme* files in the Supplementary materials[1]. Here we provide only the crucial information about the main Toloka components that is required for general understanding of the recruitment and matching procedures.

[1] All the supplementary materials including raw data for the studies, and R code to replicate the graphs and tables are available online at https://osf.io/ye4w8/. oTree code and HTML/JavaScript/CSS code is also available at GitHub: https://github.com/chapkovski/toloka_games.

Similar to its competitors, mTurk and Prolific, Toloka is essentially a platform for recruiting participants for doing something, mostly online, and after the job is done, processing payments both in the form of a fixed fee, paid per task, and sometimes an additional bonus. Most crowdsourcing platforms do not limit on nature of jobs posted for their audience, with the obvious exceptions of illegal activities, and tasks that might reveal the identity of their workers. The tasks vary from image labelling and tagging texts for natural language processing to, as in our case, participating in surveys or behavioral games. A single task can be fulfilled by many workers, resulting in several assignments for each task. This number of workers who can fulfill the task (for instance, taking part in a survey) is set in Toloka, through the `task overlap` parameter.

To create a task, an experimenter must first create an interface through which Toloka users will communicate. Such an interface is called *project* and is a combination of code (in HTML or JSON format), and a set of input and output fields. The input fields are the variables that will be shown to each participant, and output fields are the responses provided by the participants within the Toloka system.

As soon as a project is created, participants can be invited to a specific study (or any task in general) though opening a *pool*. A *pool* is a combination of settings, such as a participation fee, number of participants (*task overlap*), and some filters that limit access to the study to a specific audience. These filters are either built-in, or provided by Toloka. These include a region by IP, a participant's country by their registered phone number, and their self-reported nationality, age, gender, educational level and knowledge of languages. Unlike Prolific, which provides an experimenter with an extensive list of dozens of screeners, Toloka's list of available filters is rather modest. Instead they follow the path of mTurk, allowing them to assign custom *skills* (in mTurk terminology, *qualifications*) to create population subsets that meet any requirements of the researcher. These skills serve as filters for follow-up studies where workers who own these skill can be invited or excluded.

As soon as the pool parameters are set, an experimenter needs to provide a *tasks file* with input fields to start the study: in the case of survey, this tasks file simply contains a link to a server where the study is hosted.

2.2 Size and characteristics of Toloka audience

The important feature that distinguishes Toloka from its competitors is that it provides, in real time, how many users are currently available online based on filters that you apply to its audience. This functionality allowed us to collect the size of the available audience in real time. For seven days in December 2021 (from 12/14 to 12/20) we monitored an online presence of Toloka users. Every quarter of an hour we requested the number of active users based on 32 different characteristics (see the full list in the Supplementary materials).

Here we report the average audience size by time of day, and day of the week. This information can be crucial for planning the best time for conducting experiments, since the population recruited at different times of the day shows different levels of experience with behavioral studies [22]. We report here the general audience size (without any restrictions), the Russian audience, measured by several methods (by current IP of the user, their registered phone number, self-reported nationality, and self-reported knowledge of the language), along with some other large post-soviet country/languages (Ukraine, Belarus, Kazakhstan plus India as one of the largest non-soviet country presented at that time in Toloka). We also evaluated the population size by their self-reported gender, educational levels, and fluency in English and Russian. All time and data values reported are in UTC. The code used for collecting these data, and the raw data are available in the Supplementary materials.

An average online presence across all weekdays remain roughly the same (about 15k total participants, of which about 10k had Russian IP addresses), with Sunday showing a slight drop in the online population size, and Thursday having the largest online presence. The graph in Figure 1, in addition to an average audience size, also reports the minimum and maximum values observed in that day, both for the general unrestricted audience and for those participants who were located in Russia (based on their IP address).

The size of the audience heavily fluctuates within a single day (Figure 2). It never falls below 6,000 active users, and during working hours (9 AM to 20 PM) it stays above 15,000 reaching its maximum of 22,000 at 3 PM. We also estimated the share of Russian-speaking audience: among those present, we counted those who said that they speak Russian, and those who marked in their Toloka profile that they speak English but not Russian. The size of Russian-speaking population drops at night, while English speaking and others remain the same, resulting in an increasing share of non-Russian speaking population during the night up to 45%. During the day the Russian-speaking share is about 72% (Figure 3). Across the week, the proportion of Russian speakers remains relatively the same (except on Saturday, when it drops to 70% (see Figure 4). The share of those who did not declare the knowledge of either English, or Russian never exceeds 5%.

We also estimated the online population sizes based on the country they are located.

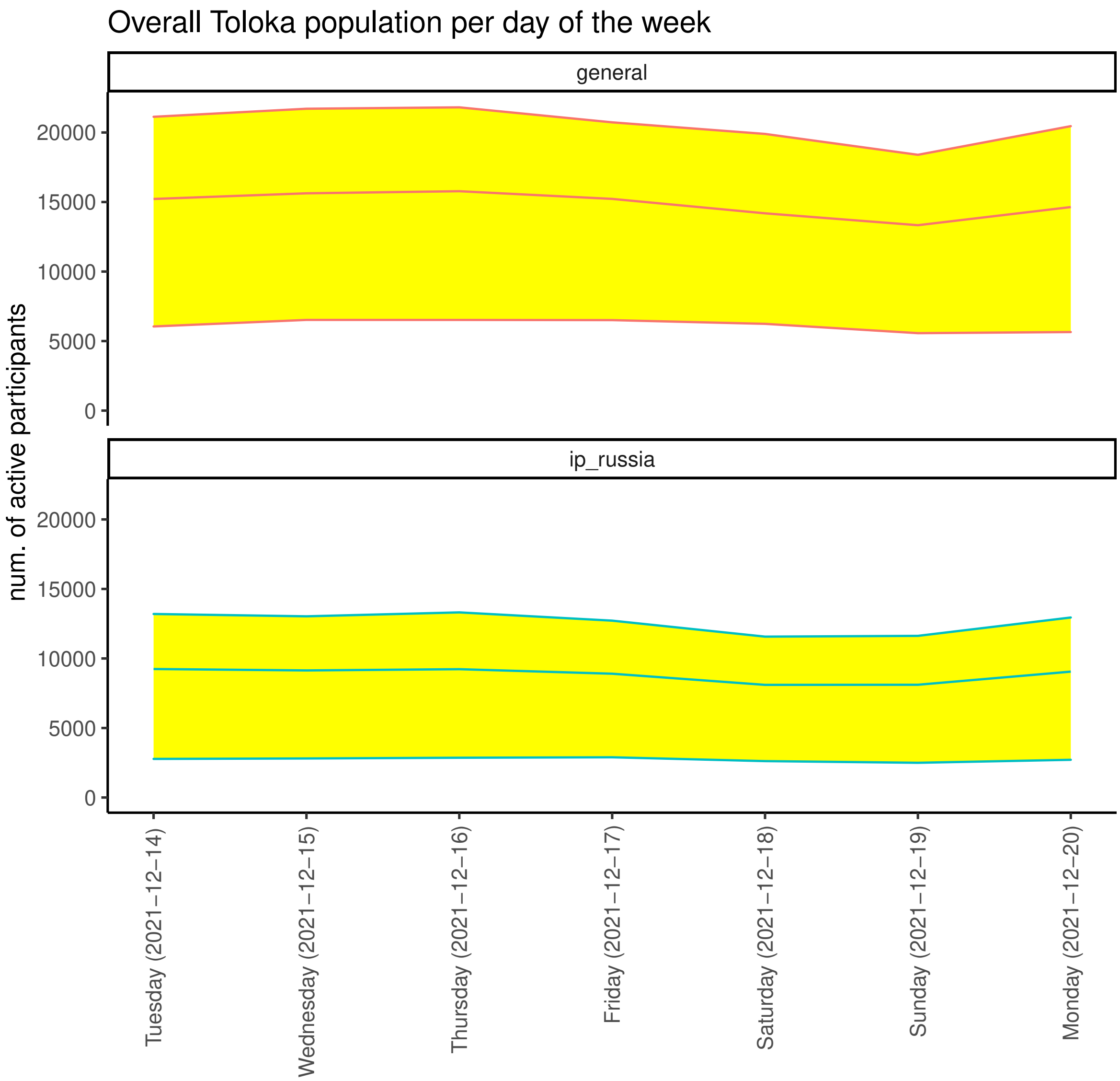

Figure 1: Toloka audience size per day of the week

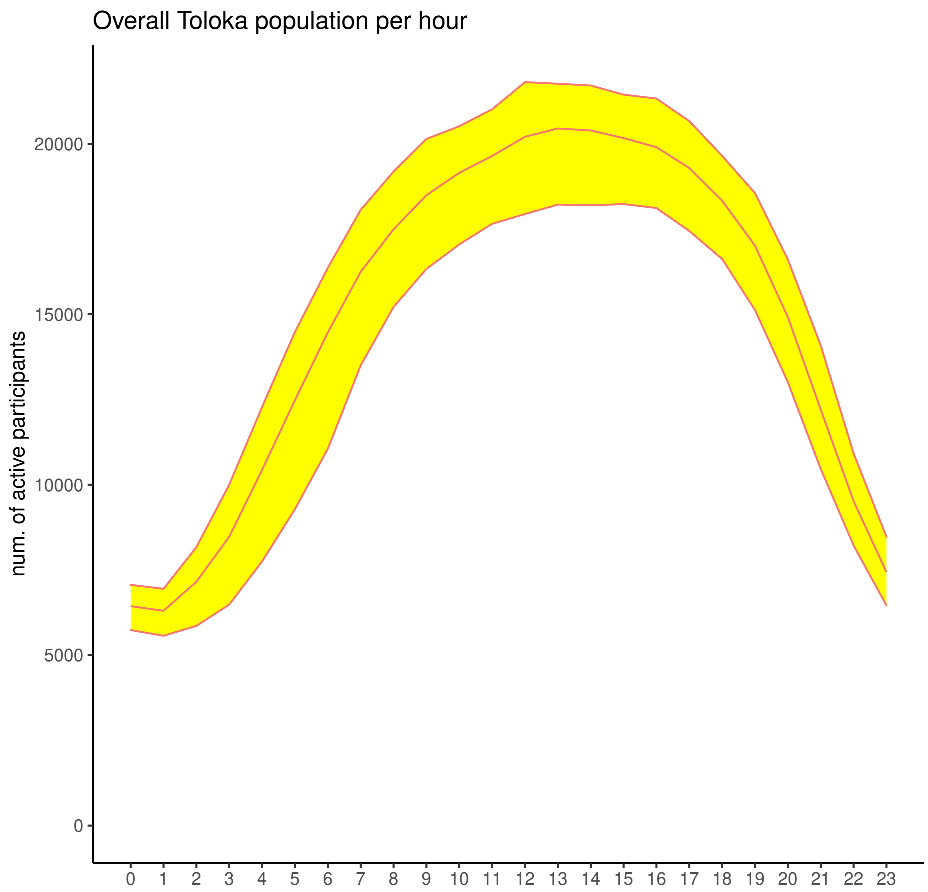

Figure 2: Overall Toloka population per hour

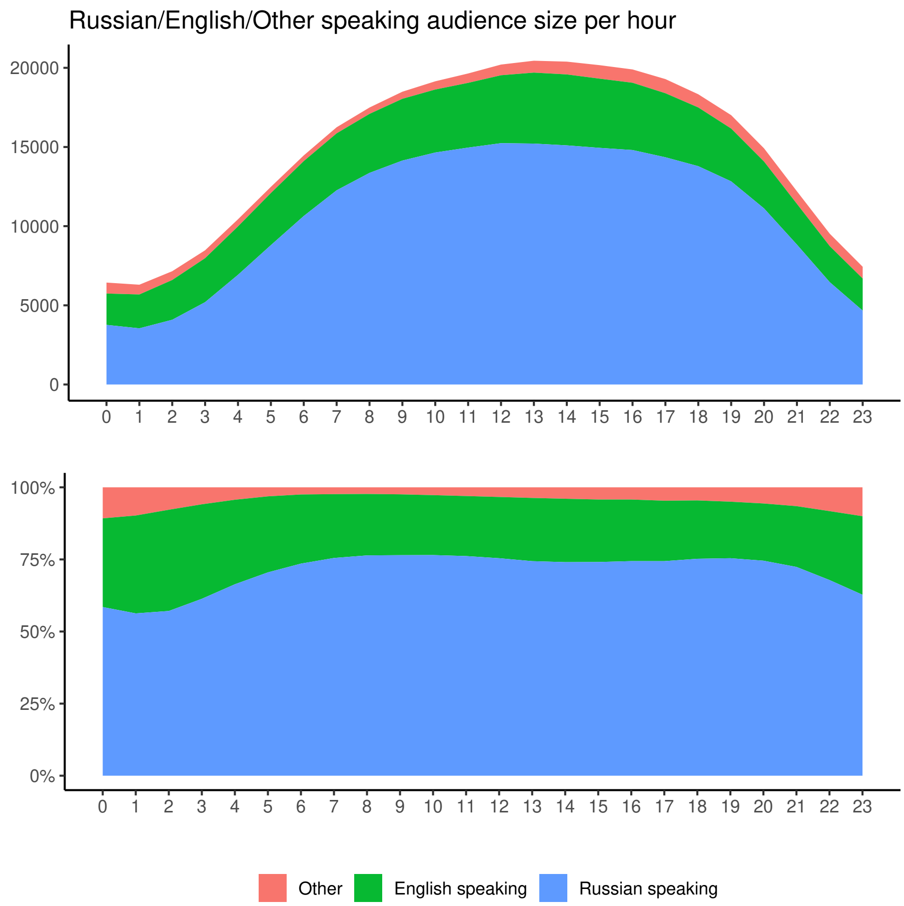

Figure 3: Russian/English/Other population per hour

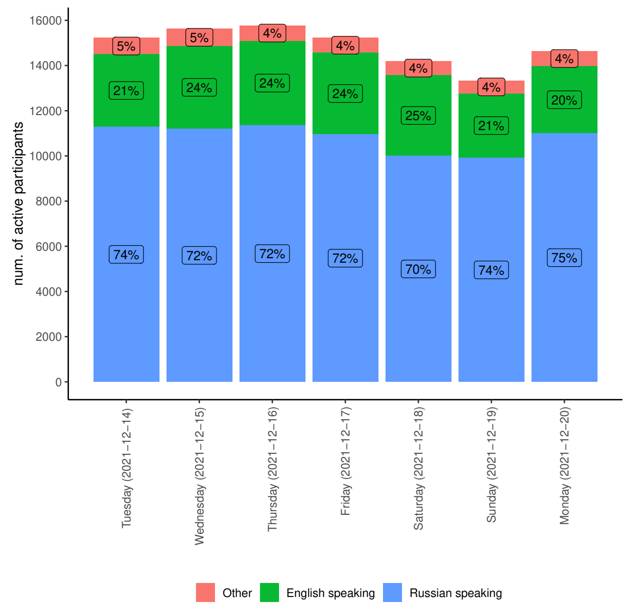

Figure 4: Russian and English speaking audience per day of the week

Measurement	Belarus	India	Kazakhstan	Other	Russia	Ukraine
By IP address	2.4%	2.8%	1.5%	30.6%	59.4%	3.2%
By phone number	2.4%	2.8%	1.5%	29.1%	59.0%	5.3%
By nationality	2.3%	2.4%	1.4%	33.4%	55.6%	4.9%

Table 1: Share of different countries in online audience of Toloka

There are three possible ways to do this in Toloka: using the current IP address, the phone number used at registration, and the self-reported nationality. All three methods can produce slightly different results: people can use VPN services to access the Toloka web-site (which is particularly relevant for Ukrainian users whose access to Russia-located resources can be limited), they can use mobile phones in different countries while roaming, or they may live in a country different from their nationality. However all three methods result in very similar estimates of population composition (see table 1). In general, the Russian audience (whatever method is used) contains slightly less than 60% of the population, with the other four largest countries (Ukraine, Belarus, India and Kazakhstan) are responsible for a combined 10% of the population (about 1500 of active participants available online in any given moment of time), and the rest of the world covers the remaining 30%. No other country comprised over 1.5% of users, although there is some anecdotal evidence that the population of users from Vietnam and Turkey is steadily growing.

There are three other parameters available for filtering: gender, age and educational level. We do not report age here (see more on age distribution for Russian-speaking population in the Section 3), but we collected data about educational level and gender. However due to the recent changes, Toloka no longer requires the gender and education fields be filled at registration. Therefore, about 30% of active users do not reveal their education and roughly 15% have no information about gender. In total, 44% marked that they have a "high education" (although Toloka does not specify what this means), for the Russian audience (by IP address) it is 41.4%.

On average gender composition is relatively stable across all hours of the day, with the share of 'other' or 'unknown' growing during the night: that is explained by the recent expansion of Toloka to other countries outside Russia, which coincided with turning the gender question to a non-mandatory field. The same is true for gender composition during the week: there it is also stable, with slightly more males than females (45% vs 37%) and about 17% who do not reveal their gender identity or selected the 'other' option. The situation among those located in Russia (by IP address) is similar but with the slightly lower share of those who skipped gender information: 45.7% of males, 42.8% of females and 11.4% of others or unknown.

3 Overview of Russian-speaking audience of Toloka

In addition to estimates from the online audience, we ran a study of 1000 participants to increase our understanding of the average Russian-speaking Toloka audience[2]. In total we collected 990 observations, as 10 participants dropped out. The only limiting filter for participation, which we applied was that the participants must have claimed to speak Russian in the Toloka profile. As shown in the section 2, that covers about 72% of the active Toloka population, and the average number of Russian speaking participants available online at any given moment is 10,810 participants (with a min value of 3,186 and max of 16,306, depending of the day of the week and hour). Thus, this survey covers a bit less than 10% of the active Russian-speaking Toloka participants.

This questionnaire investigates working patterns in Toloka, socio-demographic characteristics, values and political attitudes. The full questionnaire and detailed information about

[2]The study design was evaluated and approved by German association for Experimental Economic Research, GfeW e.V., certificate number bwcw68Gx, available at https://gfew.de/ethik/bwcw68Gx.

What share of your total income is earned on Toloka?	n	f
Non-significant	578	58.4%
A bit less than a half of my total income	187	18.9%
A bit more than a half of my total income	45	4.5%
Very significant	100	10.1%
All my income is generated by Toloka	80	8.1%

Table 2: Toloka earnings as total income share

How many surveys you participate in?	n	f
I do not participate in such studies	140	14.14%
1-2 per month	444	44.85%
3-5 per month	165	16.67%
More than 5 per month	139	14.04%
Other	24	2.42%
Hard to say	78	7.88%

Table 3: Participation in surveys

each variable and its values is available in Supplementary materials. Five questions were related to their work patterns in Toloka: "Is Toloka your main job?', "What share of your total income is earned on Toloka?" , "How many hours a week do you work on Toloka?", "How much do you on average earn per hour doing the tasks in Toloka", and "In an average month, in how many surveys do you participate?". About 31% said that Toloka is the main job for them and almost each 5th respondent claimed that Toloka provides either a very significant or the entire source of their income (Table 2). Their reported hourly income almost exactly coincides with the income reported by Toloka itself ($1.81): an average income reported in our survey is $1.78 (SD:6.40, CI:[1.38, 2.18], median: 1). The average member from our survey reported that they spend about 22 hours per week working for Toloka (SD: 42.8, CI:[19.9, 25.2], median: 15).

One of the main issues with mTurk is the "lab rats" issue: too many participants have too large exposure to all kinds of behavioral studies [22, 12]. It would be logical to expect that the Toloka audience is less experienced with these kinds of surveys, given the relative novelty of the platform. Indeed, about 60% participate in any kind of surveys twice a month or more rarely (Table 3).

Politically speaking Toloka represents the full spectre of Russian views from supporters of the governing party United Russia (18%) to Communists (7%), and from those who would vote for Vladimir Putin (30%) if the next presidential elections would happen next Sunday to

If presidential elections would happen next Sunday, whom you would vote for?	n	f
Vladimir Putin	297	30.00%
Gennady Zuganov	20	2.02%
Vladimir Zhirinovsky	27	2.73%
Sergey Shoigu	38	3.84%
Alexey Navalny	72	7.27%
Another candidate	130	13.13%
I would not vote	154	15.56%
I am not a Russian citizen	128	12.93%
Hard to say	124	12.53%

Table 4: Whom Toloka members would vote for on presidential elections

imprisoned opposition leader Alexey Navalny (7%) (Table 4). 363 of the 990 people answered that they voted on the last Duma elections: if we take into account that 148 (14.9%) of respondents said that they are not Russian citizens, that corresponds to 43.1% turnout rate, which is close to the officially reported turnout of 45.15%.

We also collected a set of answers regarding COVID-19. We asked four questions: whether they were/are sick with COVID-19, was someone in their family or close friends sick with COVID-19, what are their vaccination plans, and whether they think that vaccination should be mandatory. The results showed that at least in some dimensions, the Toloka audience is similar to the general Russian audience. For instance among Russian citizens participated in the survey, the share of vaccinated was 42.7% (N=368), which is similar to the 46% estimate from traditional pollsters in November 2021 [23].

The question about mandatory vaccination was not fully comparable with traditional pollsters data [23] because the Levada survey used a 4-point scale + 'Hard to say' choice while we used 5-point scale with the mid-point 'Neither agree nor disagree' without the 'Hard to say' choice. But still the number of those who either totally or somewhat agree in both surveys are almost exactly the same (42.5%, N=421) in the Toloka survey and 42% in the Levada survey (total N=1603).

We asked them a standard generalized trust question from the World Values Survey (WVS) [24], on whether people can be trusted or it is better to be careful with strangers. 78% (N=775) said that it is better to be careful: compare it with 74% (N=1358) of Russian audience who chose this option in the WVS survey. We asked them also to position themselves on the left-right wing political scale. While this question is not fully comparable with the WVS question, because in the WVS, there were both 'I do not know' and 'No answer' options,

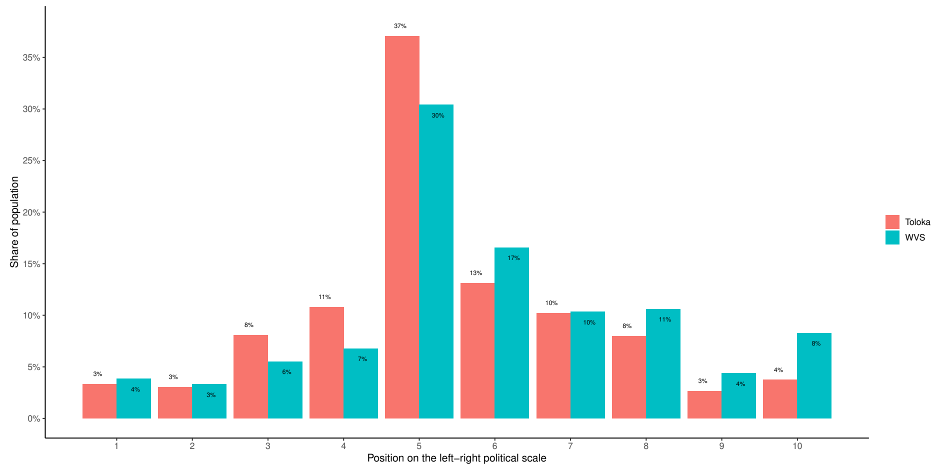

Figure 5: Position on the left-right political scale

which were chosen by 39% of the population. If we choose only those who made a choice on this question in the WVS, it becomes rather similar to what we observe among Toloka members, however Toloka members on average are slightly more left-leaning (mean 5.37 vs. 5.85 in the WVS) - see Figure 5.

In terms of socio-demographic characteristics we observed that the share of males were higher than what has been observed using Toloka's own data (58% of males). 17% of them were 24 years old or younger, and 31% were between 25 and 34 years old. 41% of respondents were single, and 37% were married. 43% had a higher education. 67% of them were employed either fully (27%), or part-time (12%) or reported themselves as self-employed (28%). 81% of the respondents reported that they were located in Russia, 7% were from Ukraine. Less than 4% were located outside the ex-USSR.

4 Behavioral experiments in Toloka

In this section we briefly present the results of several interactive studies we ran in Toloka[3]. The code used to create projects in the studies presented in this paper, and corresponding tasks files are provided in online supplementary materials.

[3]Studies 1 to 4 designs were evaluated and approved by German association for Experimental Economic Research, GfeW e.V., studies 1 to 3 are covered by the certificate number `bwcw68Gx`, available at https://gfew.de/ethik/bwcw68Gx. Study 4 is covered by the certificate number `ucfqCyFh`, available at https://gfew.de/ethik/ucfqCyFh

We described two types of projects: timed and untimed. Standard, untimed projects do not impose any starting time: a typical example is a survey, or an interactive game where the decisions can be made asynchronously even if the payoffs for different participants are interdependent (see Study 4, section 4.5). In a timed project, experimenters decide that all participants must join the study within a narrow time window, so they all can be matched into groups where they can interact. Studies 1 (Public Goods Game, section 4.2), 2, (Dictator game with the real effort task, section 4.3), and 3 (Rock-Scissors-Paper, section 4.4) are of this kind. But regardless of the kind of the study, it is necessary to link Toloka with an oTree-based server where the study data is collected. This task has two steps: first, we distribute the link to the study among Toloka participants; second, we provide an oTree server with Toloka user identification numbers so the payoffs can be paid later in the form of bonuses. Both aims were achieved by using the JavaScript code which appends to an oTree link a unique identification number that associates the oTree participant with a Toloka participant (the code is available in the Supplementary materials). This procedure is similar to an integration of external studies in Prolific via URL parameters [7]. To point participants to a specific study, a link to this study should be inserted in the *tasks file* as an input field. This link can either direct them to an active experimental session (in oTree terms: session-wide link) or to an existing waiting room. Figure 6 demonstrates a screenshot of a standard project in Toloka that points to an oTree study.

If we talk about standard asynchronous studies that do not involve real-time interactions of participants, the link to an existing session is the most logical option. However, to guarantee that participants will join the study in a dedicated time slot, an oTree *waiting room* can be used. A waiting room in oTree is a link that remains the same across different experimental sessions. When a participant accesses this URL in their browser, they are redirected to a page that announces that they should wait until the study will be initiated by an administrator. An administrator can monitor the number of people in the waiting room in real time, and choose to start a specific experimental session for those in the room. This scheme has been rarely used in online experiments because arrival times are rather unpredictable, which makes it impossible to open a session for a fixed number of participants. Some people who arrive early will leave the study after waiting too long in the waiting room, and some who arrive after the session is activated, will get the message that there are no more free slots. The structure and the size of Toloka audience mostly solves the issue of late arrivals. For studies 1 to 3 in order to make the arrival window more targeted, we gave participants access to the link to the study, which announced start times in the study description. We opened a Toloka

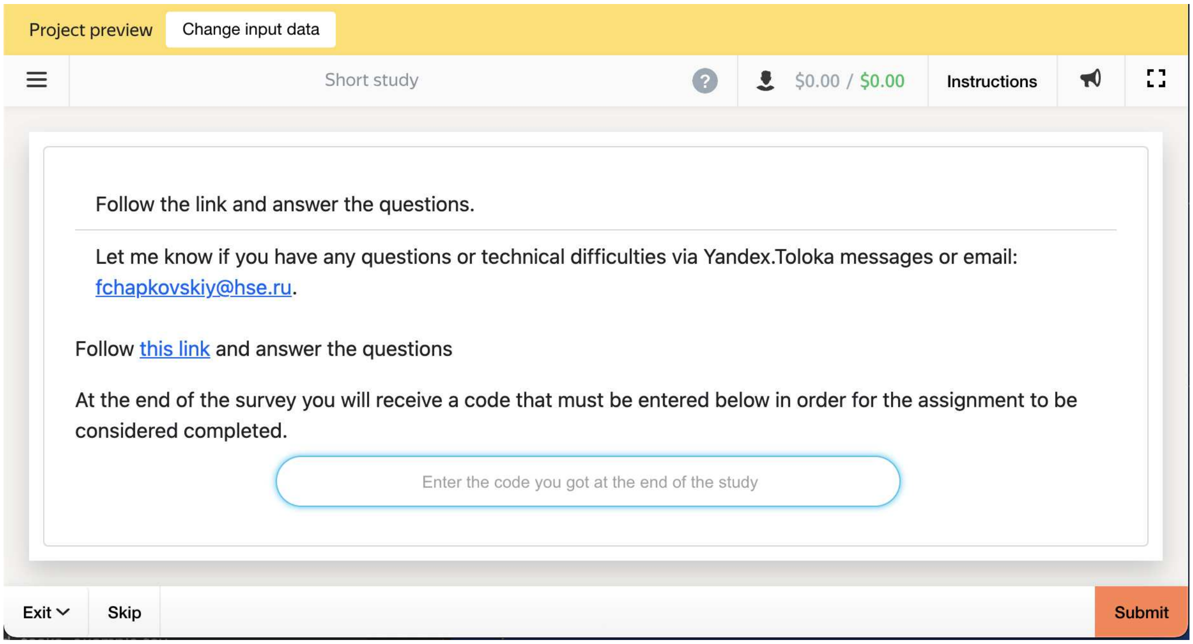

Figure 6: Screenshot of a standard (untimed) project shown to a Toloka participant

pool a few moments before the announced time, and those who joined the study before this time, could see a counter that calculated how in many seconds they would be able to access the link (see Figure 7 with a screenshot of a typical timed project interface as it is seen by a Toloka participant). The provided JavaScript code calculated the time until the link would be accessible no matter what time zone a participant was located.

When the counter is completed, the link to the study appears. When participants clicked the link they were redirected to the oTree waiting room. Within 60 seconds after the official time of the study the Toloka session was closed for new arrivals, and the participants waiting in the waiting room were redirected to a newly created experimental session. There, they accepted a consent form (if applicable), read some introductory instructions that included exchange rate, and were matched with other members of their group.

Those who were redirected to the waiting page for matching but were not able to find a partner after 60 (or 90 depending on settings) seconds of waiting, were redirected to the page where they were informed that we could not find a partner for them but would still be paid the participation fee. On average, people waited less than 5 seconds for their partners for the matching stage and the amount of drop-outs during the game unlike the Arechar et al. study was negligible. During the game we put some limits on the time given for

[3]Original screenshot was in Russian. Here we demonstrate an automatic translation to English.

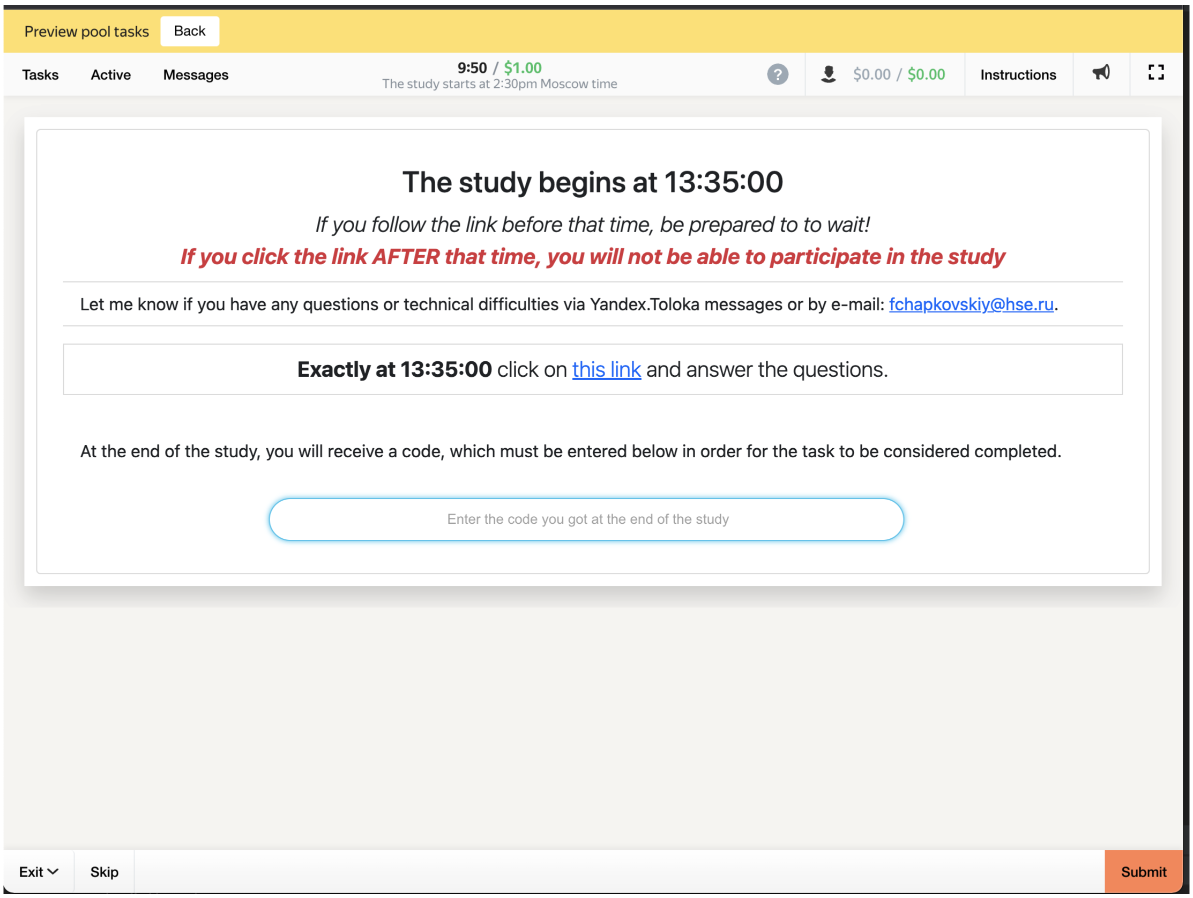

Figure 7: Screenshot of a timed project shown to a Toloka participant

decision-making (usually 60 seconds per decision). If a person did not deliver a decision within the limit, we counted him/her as a drop-out, s/he was redirected to the page where they were informed that they had been dropped from the study, while his/her partner was redirected to the 'Partner's drop out page' where we remind them that they were eligible for participation fee. The feedback provided by participants (available in the Supplementary materials) was positive: almost none of them encountered significant delays waiting for their partners' decisions.

The main reason that facilitated this matching was the short time required for an average participant to join the study, after the study was posted on the platform. A short arrival time distinguishes the Toloka platform from the other platforms (Prolific and mTurk), and opens up the possibility to conduct interactive experiments there.

4.1 Arrival times

To demonstrate the difference in arrival times across three platforms, we conducted three short small-scale (N=100 for each case) surveys tracking their time of acceptance of the task. We measured the time difference between the posting time, and an actual time when participants accepted the task in the platform interface. Since all three platforms have different audience composition country-wise, for making comparison possible we conducted all three studies on the same date (12/14/2021), but in different timeslots: we opened the Toloka study at 12:30pm Moscow time, the Prolific study at 12:30pm UTC, and the mTurk study at 12:30 EST.

The mean arriving time was 557 seconds for Prolific, 1171 seconds for mTurk and just 62 seconds for Toloka (Left panel of Figure 8). Median times were almost the same for mTurk and Prolific (516, and 597 correspondingly), while just 54 seconds for Toloka.

In less than 100 seconds after the start, all the Toloka participants had joined the study. It took 7 times longer for Prolific, whose speed of joining the study was much slower. After more than 20 minutes, less than 75% of requested slots in mTurk had been filled in (see Figure 9).

An additional issue with running interactive experiments online in crowdsourcing platforms such as Toloka or mTurk is that participants can accept the HIT (in mTurk) or assignment (in Toloka), but not start working on it for some time. They can reserve a task while, for instance, completing other, previously assigned tasks. The delay can substantially complicate

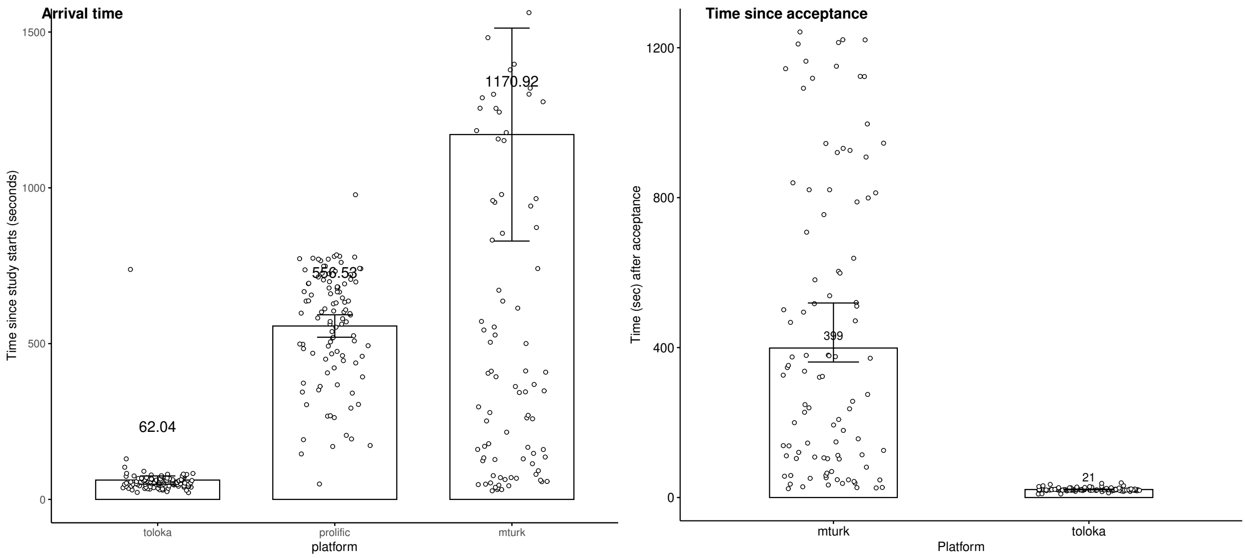

Figure 8: Mean time since the start of the study and acceptance time

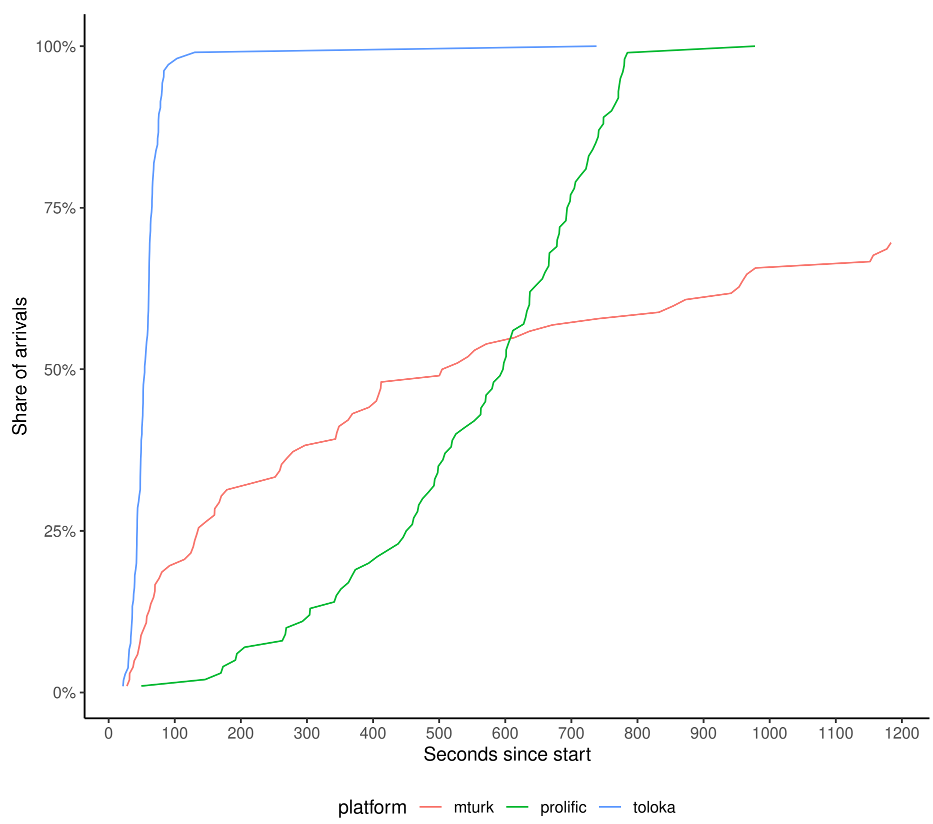

Figure 9: Cumulative share of participants who join the study by time (mTurk outliers are omitted)

the matching procedure, because some participants don't join the study on time, but they also prevent others from joining when the available slots are all booked. This problem does not occur on the Prolific platform, where people 'accept' the task at the moment they click on the study link. Since both Toloka and mTurk provide information on when a participant accepts the task, we can trace how long it takes for each participant to start working on the task after he or she accepts it. The average time-since-acceptance for mTurk was 399 seconds (SD:499, CI:[300, 498], median: 322) which was almost 20 times higher than for Toloka, where it took on average, 21 seconds (SD: 4.56, CI:[20.4, 22.2], median: 20.8), (right panel of Figure 8).

4.2 Study 1 - Public Goods Game

Public good game (PGG) is a standard tool of behavioral economics for measuring degree of cooperation within a group [25]. The typical PGG includes N subjects where each subject is provided with an endowment w. The subject cancan invest any amount g_i from 0 to w to the public good. Whatever is left from the endowment after the investment, remains in his/her private account. The total amount $G = \sum_{j=1}^{N} g_i$ contributed by all N members of the group into the public good project is multiplied by a positive coefficient k, and the resulting amount kG is distributed equally among all the team members no matter how large their contribution was. The individual payoff then is defined as follows:

$$p_i = w_i - g_i + k/N * G \tag{1}$$

Thus the return on investment to public and private accounts differ (it is 1 for the private fund, and k/N for the public fund) and as long as $k/N < 1$ it makes sense for a rational profit-maximiser to invest zero to the public good, regardless of the contributions of the other group members. If everyone in the group followed that logic, the contribution to public good would be 0, and the payoff for each member equals endowment w. However in case of $k > 1$, the Pareto-optimal solution for the group as whole is to invest the entire endowment to the public good, and get $kw > w$ as a payoff. This tension between individual profit-seeking and Pareto-optimal solution is what makes PGG such an attractive tool to measure an ability of a person to overcome their own self-interest for the sake of common good. Usually people invest more than predicted by rational choice theory, although over time the cooperation rate deteriorates so in the last periods people contribute much less than in the beginning [26].

Methods

We played a standard PGG without a a peer punishment stage. Participants were matched into groups of three, and they stayed in the same group during the entire study (so called *partner matching*). In each period participants were provided with an endowment of 100 US cents, and the study in total lasted for 10 periods. At the beginning of the study, participants were informed that only one period will be chosen randomly out of 10 and the payoff in this period will define their bonus. We do so to avoid wealth effects and hedging [27]. The public good investment coefficient k was 1.5 and the group size was 3, so MPCR (Marginal per-capital return) was 0.5. Thus in case of the full cooperation the maximum payoff in a single period was 150 US cents, and in case of 0 cooperation they got 100 US cents.

Results

Matching time The total number of participants who joined the study was 117. Ninty-two of them successfully completed the study, 7 people were blocked due to inactivity, 18 participants were blocked by the non-activity of their group members: 14 were blocked by 7 inactive participants mentioned above, and 4 participants who arrived to the matching page, after 90 seconds of waiting for a partner were redirected to the final page. Most importantly there were no participants who dropped out of the study in the middle of the game, all 7 drop-outs did this at the beginning of the first round. For those 92 participants who were successfully matched and completed the study the average matching time was just 1.63 seconds (SD: 3.92, SI: [0.83, 2.43]), with a median of 0.22, with 74 of them being matched with 1 second or less of waiting. Figure 10 shows the distribution of waiting times.

Mean time of the study for those who completed it was 673 seconds (SD: 146, SI:[644.04, 703.76], median: 675), distributed normally (Shapiro-Wilk test: $W = 0.98749$, p-value $= 0.5324$). The average contribution to public good was 45.94 (SD: 2.09, CI: [43.85, 48.03]). We observed a typical slow deterioration of cooperation rate that is typical for other VCM games (Figure 11) - while in the first 3 rounds the mean contribution was 49.6 (SD: 3.67, CI: [46.0, 53.3]), in the last three rounds it was 41.1 (SD: 3.88, CI:[37.3, 45.0]).

The amount of time needed for decisions dropped fast after the first few periods where an average time for decision was 21 seconds, reaching on average 9 seconds for the last rounds (Figure 12)

An important question that required some investigation was whether it makes sense to rely

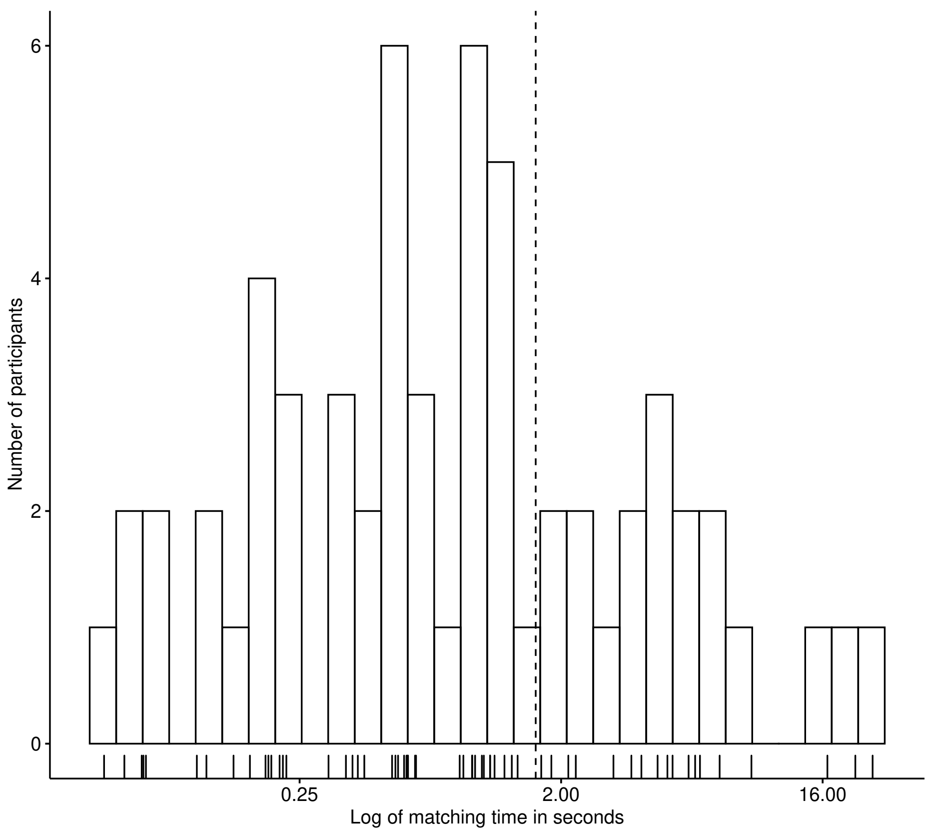

Figure 10: Matching time for PGG participants (log(seconds))

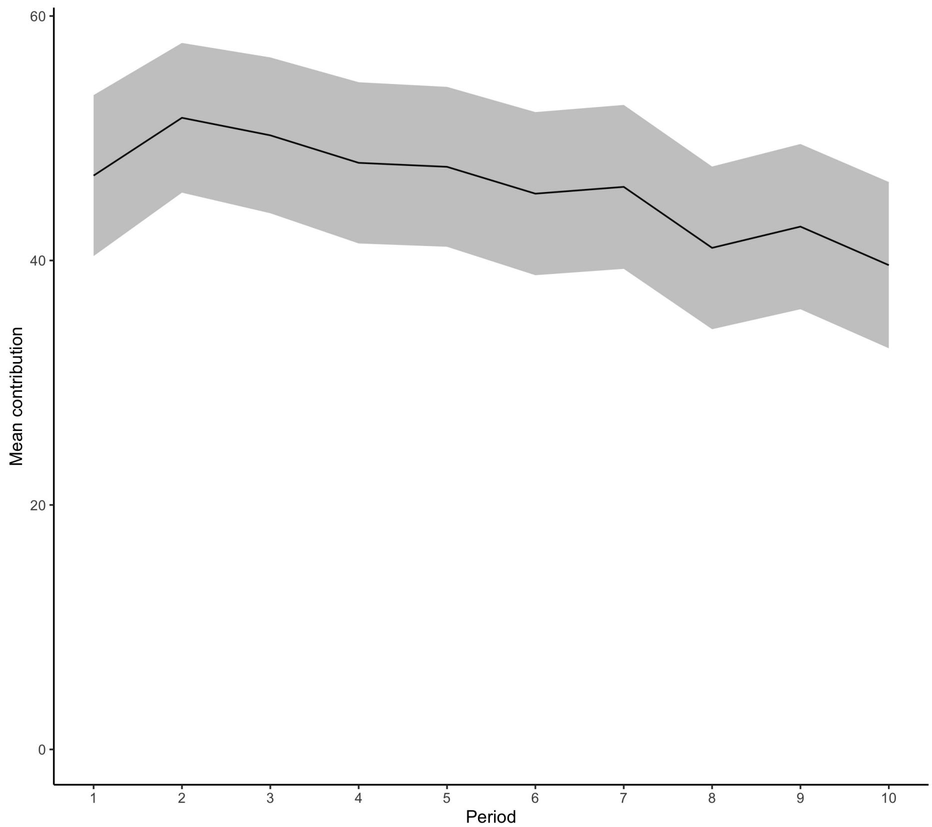

Figure 11: Mean contribution per round in PGG

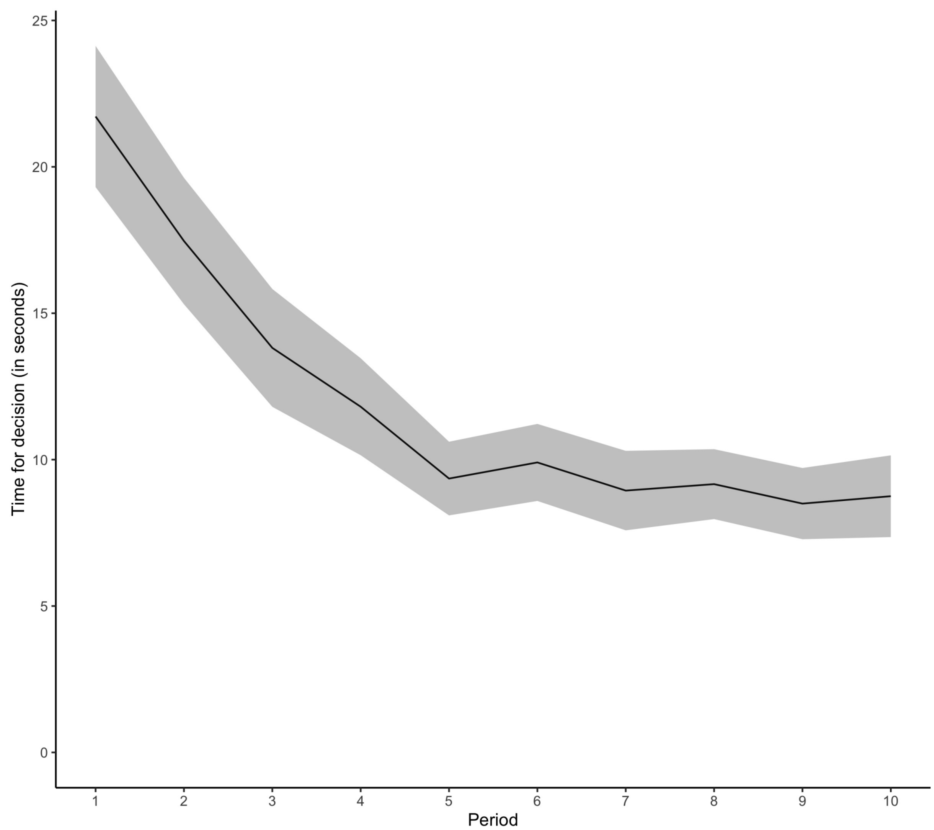

Figure 12: Time spent on decision stage per round in PGG

on the Toloka user rating as a filtering mechanism for increasing data quality. When a pool is created, it is possible to restrict availability to the top quality users (those online user who have the highest performer rating). Unlike mTurk which provides two major qualifications used to filter best users ('Number of HITs submitted' and 'Share of HITs accepted'), Toloka is sadly vague in documenting how the rating is calculated: "The performer's rating reflects the quality of the performer's responses to tasks. It takes into account responses to control tasks, bans and periods without work." However in studies reported below, Toloka user rating did not significantly affect neither decisions, nor time required to make a decision.

To measure a potential effect of Toloka rating on participants' decisions we downloaded the user profiles from Toloka (see the code in Supplementary materials). There were no correlations with contributions made in PGG nor with the time needed for the contribution decision (Pearson correlation coefficient for rating-decision time is -0.0749, and for rating-contribution is 0.0350).

4.3 Study 2 - Dictator game with the real effort task

The second study investigated how reliably Toloka can be used for behavioral games where the decision stage is preceded by some real effort task (RET) (a rather typical design for studies using modifications of Dictator game, see for instance [28, 29]). Usually the productivity of an actor in such RETs defines the choice set or endowment available for the decision stage. In our case, we used a standard RET of counting zeroes [30] in a large matrix of numbers. People worked in groups of two, and were matched to a role of dictator or recipient in a subsequent dictator game based on their productivity in the RET stage: the more productive participant becomes a dictator. In case of two players being equally productive, the dictator role was assigned randomly. This specific design was included into this set of studies to demonstrate how the matching can be done not only on the basis of random arrival to the waiting room, but also based on some measure (in this case performance) in the earlier stage of the study. It may also be informative to measure performance of the average Toloka user in this RET which has been widely used in similar studies.

Results

Out of 102 participants entered the study, 96 were successfully matched, 4 were blocked due to inactivity and 2 were blocked by inactivity of other members. The average matching time

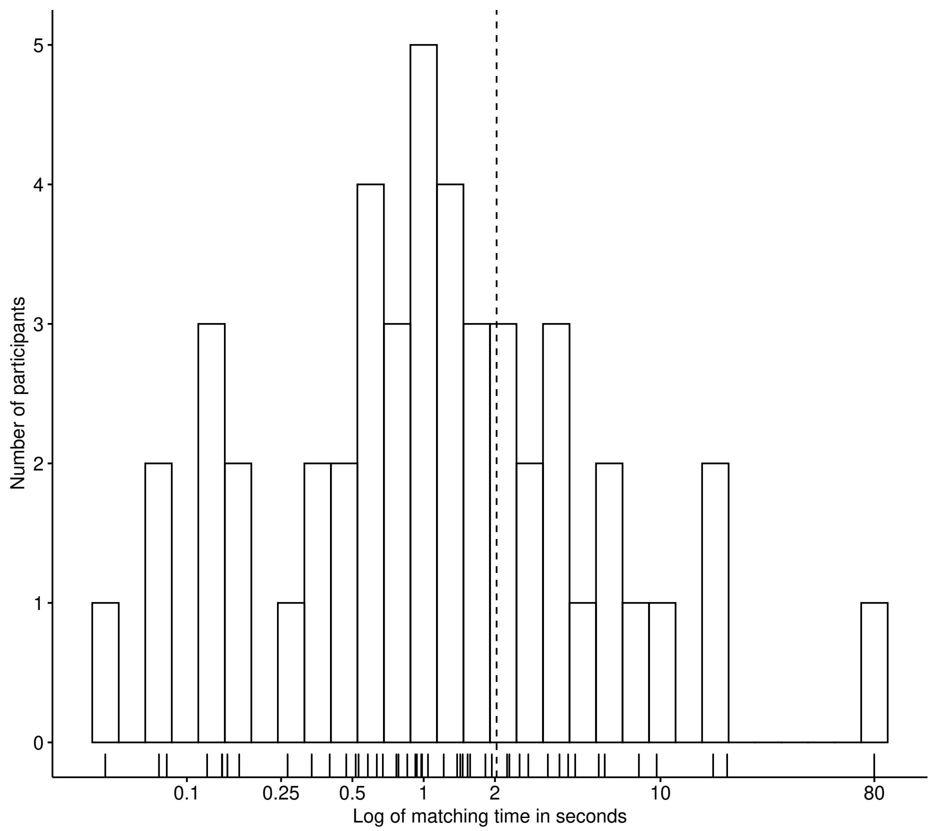

Figure 13: Matching time for RET-DG participants (log(seconds))

was 2.03 seconds (SD 8.57), with the maximum waiting time of 80 seconds (see Figure 13 for full distribution).

Participants were matched into pairs after demonstrating their productivity in the real effort task. Their productivity defined their role, with the more productive partner receiving the right to distribute a dictator's endowment between him/herself and another participant. The matching indeed resulted in a much higher number of correct tasks submitted by dictators (mean: 6.04, SD 1.86, N 48), than by recipients (mean: 3.69, SD 1.64, N 48) - see Figure 14.

Dictators contributed on average 30.8 cents out of 100 (SD 24.4, N 48, median 30), which is close to the average contribution in other dictator games of 28.35% [31]. The beliefs of recipients regarding the amount they expect their dictator would send them were a bit higher than the actual amount sent: 35.8 (SD 25.3, N 48, median 40). Detailed distribution of transfers and beliefs is shown at Figure 15.

There was no correlation between user rating (provided by Toloka) and their productivity in the first phase of the game (Figure 16).

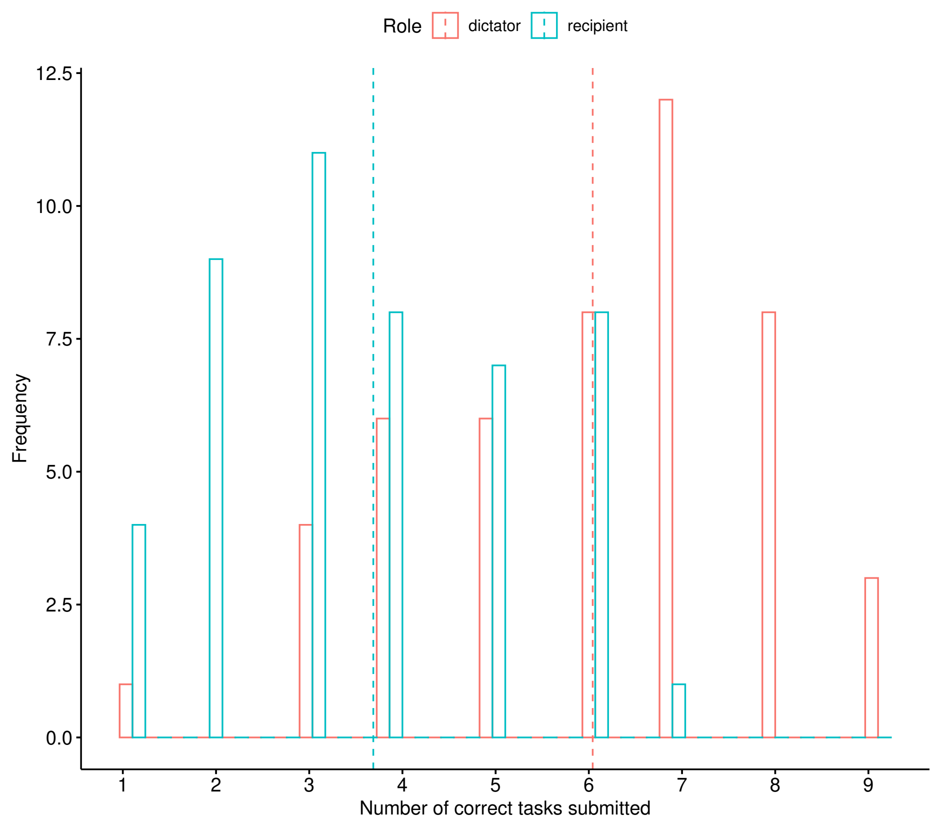

Figure 14: Correct tasks submitted by role in RET-DG

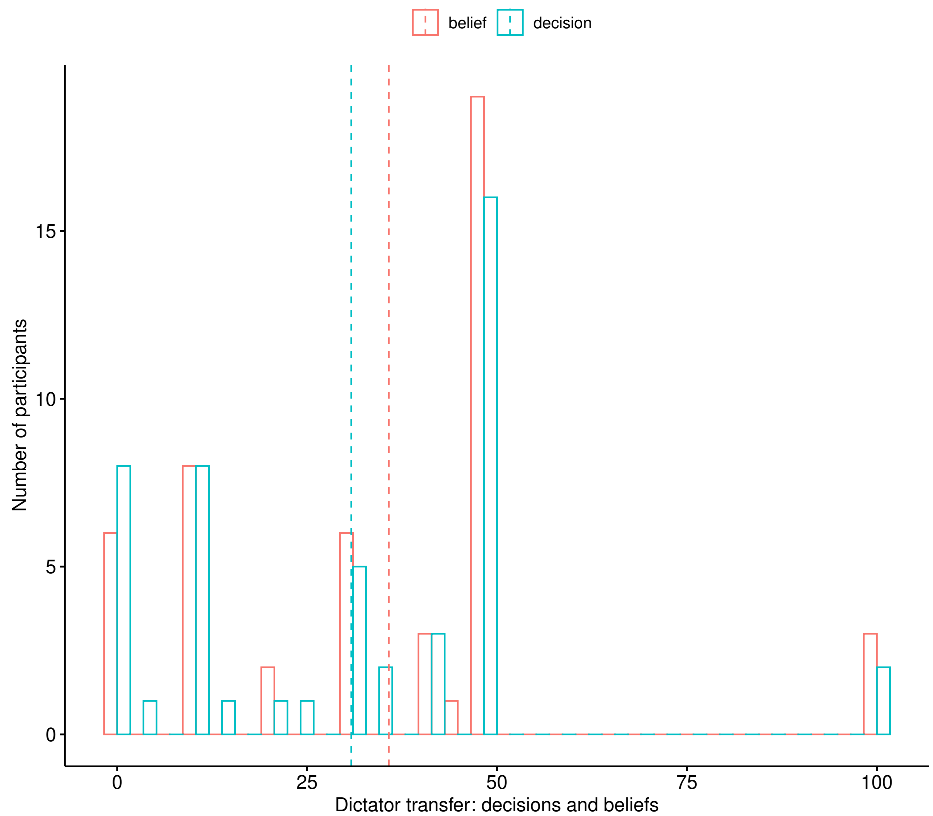

Figure 15: Decisions by dictators and beliefs by recipients in RET-DG

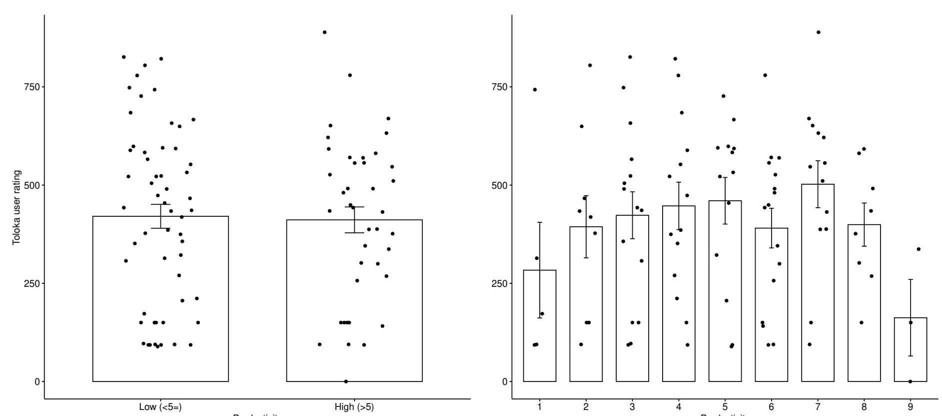

Figure 16: Productivity by user rating in RET-DG

4.4 Study 3 - Rock-Scissors-Paper

The Rock-Scissors-Paper (RSP) game is one of the simplest zero-sum games. In contrast to PGG, which is primarily used to study cooperative interactions, the RSP is the main workhorse to investigate how people make decisions in non-cooperative strategic interactions [32, 33, 34, 35]. During the game participants were matched into groups of two, and stayed in these fixed groups for 10 rounds. In each round they simultaneously chose one of three options (Rock, Scissors, or Paper) and then their decisions were matched with each other, and payoffs were calculated according to a simple rule: the Rock 'beats' Scissors, Scissors 'beats' Paper, and Paper 'beats' Rock. If both participants choose the same item there is a tie. In our game setup the winner gets $1, the bonus for a tie was $0.50, and the loser gets $0. One of 10 rounds was chosen randomly to define the final participant bonus.

Results

The total number of participants were 111, among whom 102 were successfully matched and completed the study, with 9 not matched after arrival. There were no dropouts during the game itself.

Time till matching: The matching time took even less than in the PGG: out of 102 matched participants 90 waited 1 second or less till matching. An average waiting time was 1.09 seconds (SD: 4.31, CI:[0.24,1.94]), see the full distribution of matching time in Figure 17. It took on average 542 seconds to complete (SD: 162, CI: [509.97, 573.75], median: 525). There were some substantial outliers on the right tail of the distribution thus the data of completion is not normally distributed, failing the Shapiro-Wilk test for normality (W=0.7146, p-value<0.0001).

Average wins, ties, and losses per round and in total: The amount of ties, losses and wins were distributed almost equally: in 1020 observations there were 335 losses, 335 wins, and 350 ties. In zero-sum games similar to RSP, the amount of ties may serve as a proxy to how well coordinated people are in their actions. The share of ties grew after the first round, and dropped down only at the very last round (Figure 18).

Decision time per round: An average decision time was 7.40 seconds (SD: 6.19, CI: [7.02, 7.78]), median 5.62. It went down from the first round of 13.1 seconds (SD: 9.08) to 5.90

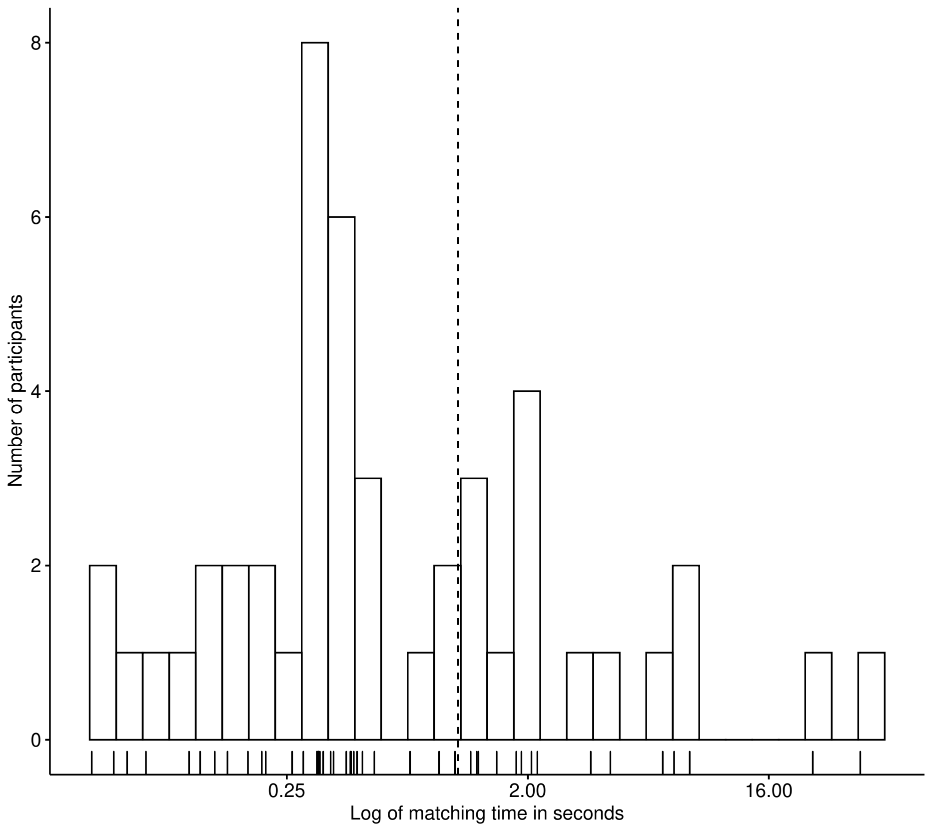

Figure 17: Matching time for RSP participants (log(seconds))

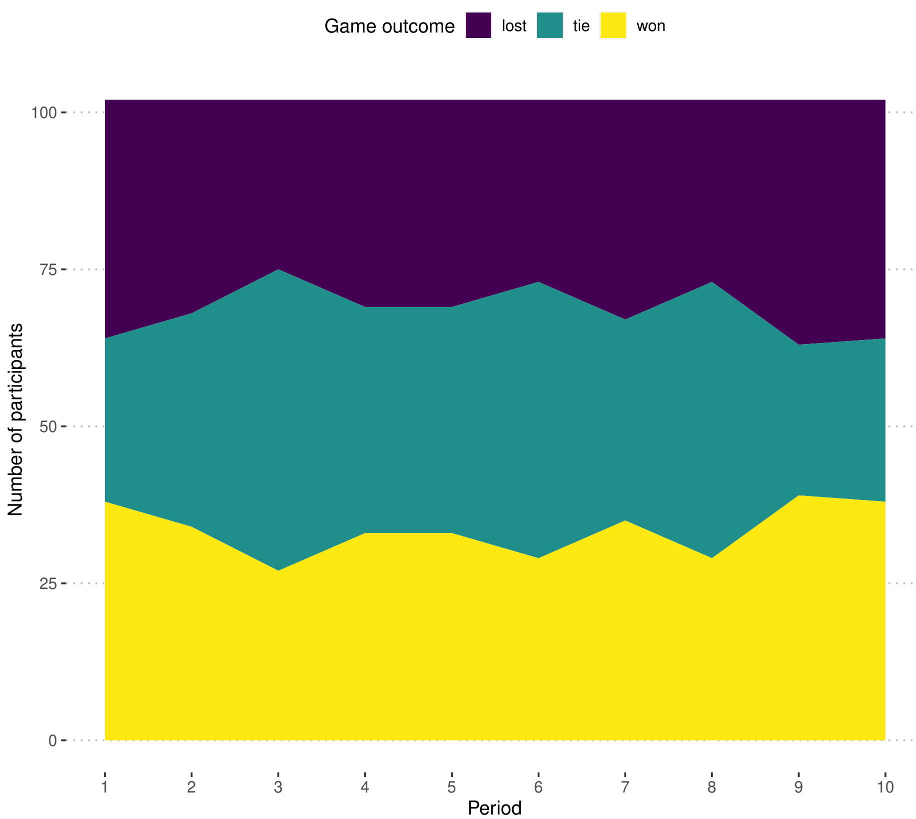

Figure 18: Game outcomes in RSP across rounds

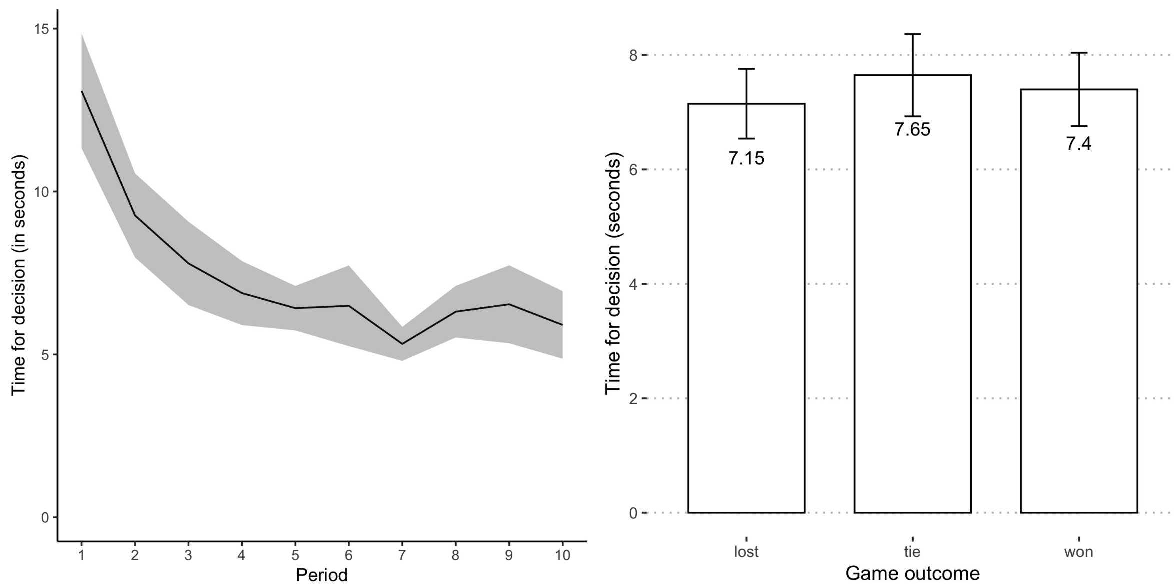

Figure 19: Time spent on decision in RSP

seconds (SD 5.34) in the last 10th round. There were no significant differences in decision time for different outcomes (see Figure 19).

As for Toloka rating, there is a very weak correlation between game outcome and Toloka rating, and no correlation between Toloka rating and time needed for decision (Figure 20).

4.5 Non-synchronous study 4: cheating game

We also report here the results of a study (N=296) that did not require real time interactions. In many cases real time interactions are not necessary, for instance in most of the Dictator games, where the decisions are taken by one of two participants, thus the game can be run asynchronously. In some cases, real time interactions are extremely hard to achieve due to the game design. For instance, any variation of a standard beauty contest game: individuals must make guesses about the most popular decision in population [36]. Without making participants wait for the entire pool of participants in the experimental session, it is impossible to calculate the correctness of their guesses. A standard tool to estimate honesty at the group level is to ask participants to report an outcome of a random event (flipping of a coin or a throw of a dice) where the outcome defines a participant payoff [2, 37]. In our study we first asked each participant to flip a coin and to report the result, which defined a person's payoff: $1 for reporting heads, and $0 for tails. These decisions and beliefs were

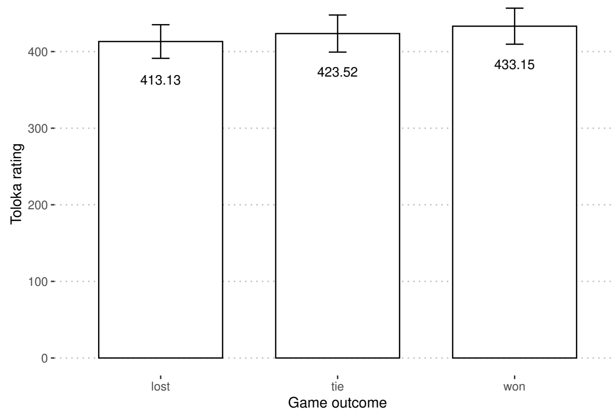

Figure 20: Average rating per outcome in RSP

collected in three separate experimental sessions for three different Russian regions (Moscow, Voronezh and Arghangelsk), using location filters by IP address provided by Toloka.

Results

Overall we observed a frequency of heads reported of 65% which did not vary much across three regions which participated in the study (Figure 21). This number (65%) is close to the number reported about Russian audience behavior in the similar study by Hugh-Jones, where data was collected online as well [38]: there Russians reported heads in 71% of the cases.

We also elicited their beliefs on what will be the average proportion of heads reported, according to a standard procedure of belief elicitation [39]. The belief elicitation was incentivised: if they guessed correctly (-/+ 10 percentage points) they received an additional bonus of $0.50. Overall beliefs about 'cheating' frequency are slightly higher than actual numbers (Figure 22) which also to a certain extent replicated the results of Hugh-Jones [38]: in his paper, average expectations of Russian audience was that Russians would report 82% of head, although in his case the number of observations was too small to draw any statistical conclusions (about 13 subjects).

5 Discussion

The Toloka platform looks like a promising tool for behavioral scientists in two aspects. First, the simultaneous presence of a large population available online (up to 20.000 participants

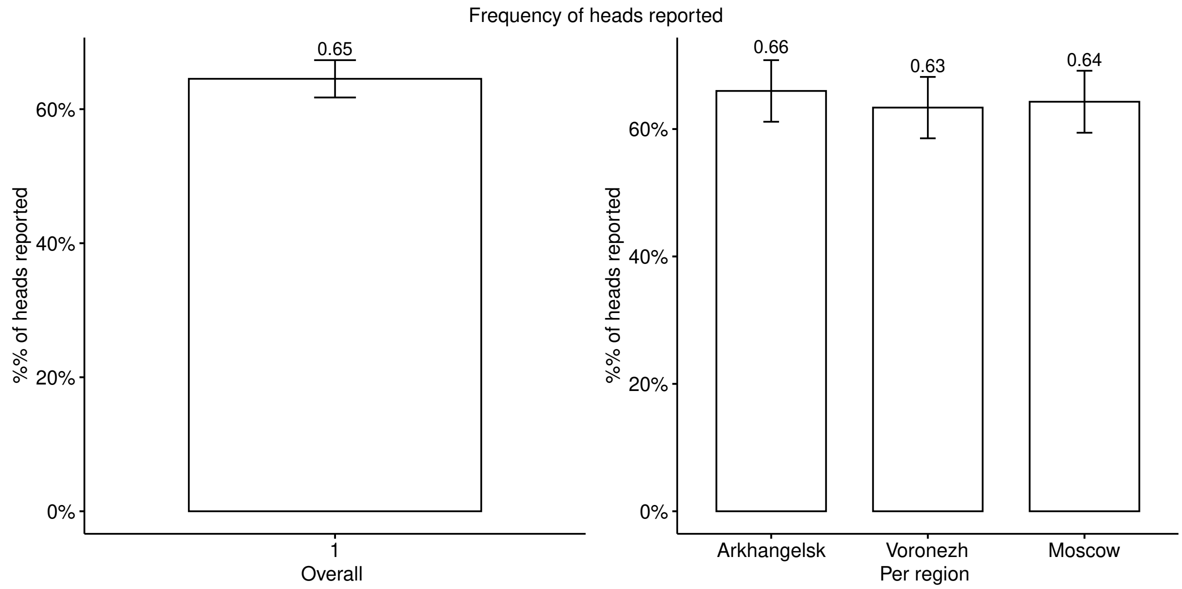

Figure 21: Average rating per outcome in RSP

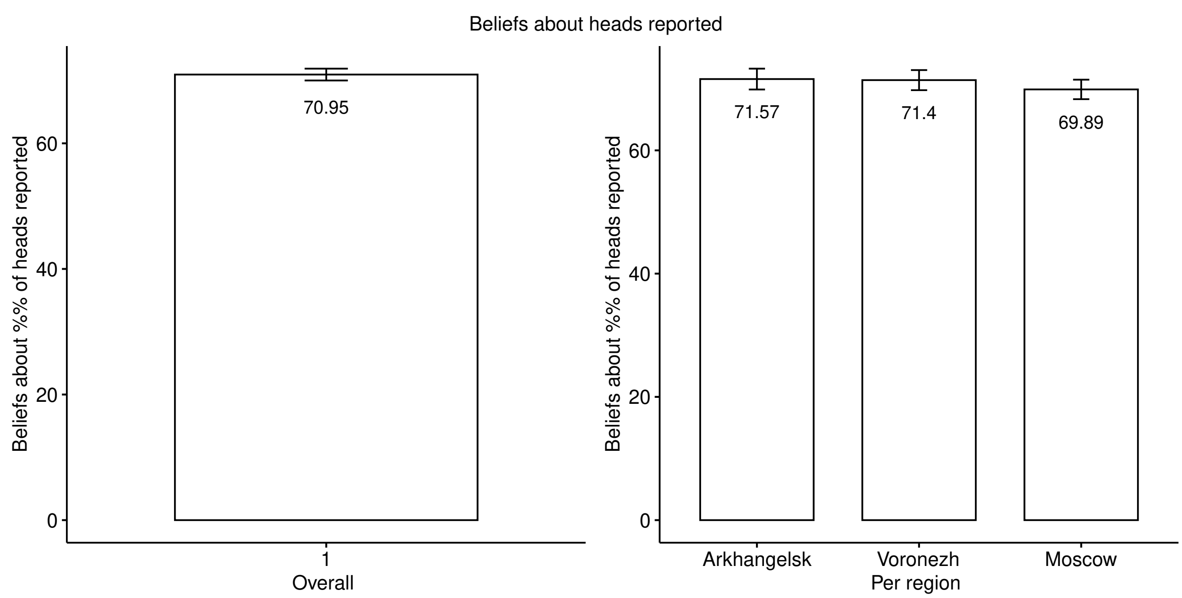

Figure 22: Average rating per outcome in RSP

presented online on the peak of the working day) and their eagerness to join the study in a matter of seconds finally solves the problem of real-time interactions. Second, it provides an easy access to the audience from the former Soviet Union that has been out of reach of experimentalists before.

But Toloka has some limitations that should be taken into account. First, unlike mTurk, the platform lacks the clear and transparent procedure to filter out participants with low quality of submissions. If in mTurk such qualifications as number of HITs submitted and approved are publicly available and can be used as filters before someone can participate in the study, it is not the case in Toloka. Instead, the platform offers a rating that as we have shown in this paper does not correlate with any behavior that is relevant for behavioral studies.

The second issue is that unlike Prolific, the number of available filters is very limited, and the recent procedural changes that made some fields (such as gender and education) optional limited it even further. Thus, a scholar who would like to target any specific audience most likely will have to conduct pre-filtering studies in order to select the audience for the main study. Third, the code that is used in this paper to send the assignment identification number back to oTree is custom-written by the authors. Prolific, for instance, provides a much friendlier way to connect to oTree (using url queries). Furthermore, unlike mTurk that allows to share qualification identification numbers across different accounts, and Prolific that lets users be blocked from participation by their personal identification numbers, there is no such an option in Toloka. On the practical level that means that there is no option for a scholar to block people who have participated in a study of another researcher or of another lab from participating in a similar study conducted in a different account.

But despite all these shortcomings, it seems that Toloka can be an important addendum to the tools available to the behavioral scientist. Its large and rather inexperienced audience, and convenient programmatic access to most of the features finally makes possible the development and conducting of interactive games online.

References

[1] Jordi Brandts and Gary Charness. “The strategy versus the direct-response method: a first survey of experimental comparisons”. In: *Experimental Economics* 14.3 (2011),

pp. 375–398. ISSN: 1573-6938. DOI: 10.1007/s10683-011-9272-x. URL: https://doi.org/10.1007/s10683-011-9272-x (visited on 02/02/2022).

[2] Urs Fischbacher, Simon Gächter, and Simone Quercia. “The behavioral validity of the strategy method in public good experiments”. In: *Journal of Economic Psychology* 33.4 (2012). Publisher: Elsevier, pp. 897–913.

[3] David J. Cooper and Hanming Fang. “Understanding overbidding in second price auctions: An experimental study”. In: *The Economic Journal* 118.532 (2008). Publisher: Oxford University Press Oxford, UK, pp. 1572–1595.

[4] Muriel Niederle and Lise Vesterlund. “Do women shy away from competition? Do men compete too much?” In: *The quarterly journal of economics* 122.3 (2007). Publisher: MIT Press, pp. 1067–1101.

[5] Marco Battaglini, Rebecca B. Morton, and Thomas R. Palfrey. “The swing voter’s curse in the laboratory”. In: *The Review of Economic Studies* 77.1 (2010). Publisher: Wiley-Blackwell, pp. 61–89.

[6] Ernst Fehr and Simon Gächter. “Cooperation and Punishment in Public Goods Experiments”. In: *American Economic Review* 90.4 (2000), pp. 980–994.

[7] Stefan Palan and Christian Schitter. “Prolific. ac—A subject pool for online experiments”. In: *Journal of Behavioral and Experimental Finance* 17 (2018). Publisher: Elsevier, pp. 22–27.

[8] Antonio A. Arechar, Simon Gächter, and Lucas Molleman. “Conducting interactive experiments online”. In: *Experimental economics* 21.1 (2018). Publisher: Springer, pp. 99–131.

[9] Yun Shin Lee, Yong Won Seo, and Enno Siemsen. “Running Behavioral Operations Experiments Using Amazon’s Mechanical Turk”. In: *Production and Operations Management* 27.5 (2018). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/poms.12841, pp. 973–989. ISSN: 1937-5956. DOI: 10.1111/poms.12841. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/poms.12841 (visited on 01/17/2022).

[10] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. “The weirdest people in the world?” In: *Behavioral and brain sciences* 33.2 (2010). Publisher: Cambridge University Press, pp. 61–83.

[11] Erik Snowberg and Leeat Yariv. "Testing the Waters: Behavior across Participant Pools". In: *American Economic Review* 111.2 (Feb. 2021), pp. 687–719. ISSN: 0002-8282. DOI: 10.1257/aer.20181065. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20181065 (visited on 01/17/2022).

[12] Ryan Kennedy et al. "The shape of and solutions to the MTurk quality crisis". In: *Political Science Research and Methods* 8.4 (Oct. 2020). Publisher: Cambridge University Press, pp. 614–629. ISSN: 2049-8470, 2049-8489. DOI: 10.1017/psrm.2020.6. URL: https://www.cambridge.org/core/journals/political-science-research-and-methods/article/shape-of-and-solutions-to-the-mturk-quality-crisis/521AEEB9A9753D5C6038440BD123826C (visited on 01/17/2022).

[13] Jesse Chandler, Pam Mueller, and Gabriele Paolacci. "Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers". In: *Behavior Research Methods* 46.1 (2014), pp. 112–130. ISSN: 1554-3528. DOI: 10.3758/s13428-013-0365-7. URL: https://doi.org/10.3758/s13428-013-0365-7 (visited on 01/17/2022).

[14] Neil Stewart et al. "The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers". In: *Judgment and Decision making* 10.5 (2015). Publisher: SOC JUDGMENT & DECISION MAKING, pp. 479–491.

[15] Volker Benndorf, Claudia Moellers, and Hans-Theo Normann. "Experienced vs. inexperienced participants in the lab: do they behave differently?" In: *Journal of the Economic Science Association* 3.1 (2017), pp. 12–25. ISSN: 2199-6784. DOI: 10.1007/s40881-017-0036-z. URL: https://doi.org/10.1007/s40881-017-0036-z (visited on 01/17/2022).

[16] Ard J. Barends and Reinout E. de Vries. "Noncompliant responding: Comparing exclusion criteria in MTurk personality research to improve data quality". In: *Personality and Individual Differences* 143 (2019), pp. 84–89. ISSN: 0191-8869. DOI: 10.1016/j.paid.2019.02.015. URL: https://www.sciencedirect.com/science/article/pii/S0191886919300972 (visited on 01/17/2022).

[17] Daniel L. Chen, Martin Schonger, and Chris Wickens. "oTree—An open-source platform for laboratory, online, and field experiments". In: *Journal of Behavioral and Experimental Finance* 9 (2016), pp. 88–97. URL: http://www.sciencedirect.com/science/article/pii/S2214635016000101 (visited on 12/10/2016).

[18] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. “Demographics and Dynamics of Mechanical Turk Workers”. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. WSDM ’18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 135–143. ISBN: 978-1-4503-5581-0. DOI: 10.1145/3159652.3159661. URL: https://doi.org/10.1145/3159652.3159661 (visited on 01/17/2022).

[19] Eyal Peer et al. “Beyond the Turk: Alternative platforms for crowdsourcing behavioral research”. In: *Journal of Experimental Social Psychology* 70 (2017). Publisher: Elsevier, pp. 153–163.

[20] Benedikt Herrmann, Christian Thöni, and Simon Gächter. “Antisocial punishment across societies”. In: *Science* 319.5868 (2008), pp. 1362–1367. URL: http://science.sciencemag.org/content/319/5868/1362.short (visited on 12/09/2016).

[21] Simon Gächter and Benedikt Herrmann. “The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural Russia”. In: *European Economic Review* 55.2 (2011). Publisher: Elsevier, pp. 193–210.

[22] Michael Chmielewski and Sarah C. Kucker. “An MTurk crisis? Shifts in data quality and the impact on study results”. In: *Social Psychological and Personality Science* 11.4 (2020). Publisher: SAGE Publications Sage CA: Los Angeles, CA, pp. 464–473.

[23] Levada. *Koronavirus, vakcinacija, QR-kody*. July 12, 2021. URL: https://www.levada.ru/2021/12/07/koronavirus-vaktsinatsiya-qr-kody/ (visited on 01/13/2022).

[24] Christian Haerpfer et al. “World values survey: round seven–country-pooled datafile”. In: *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat* (2020).

[25] R. M. Isaac and J. M. Walker. “Group-Size Effects In Public-Goods Provision - The Voluntary Contributions Mechanism”. In: *Quarterly Journal Of Economics* 103.1 (Feb. 1988), pp. 179–199.

[26] James Andreoni. “Why free ride?: Strategies and learning in public goods experiments”. In: *Journal of public Economics* 37.3 (1988), pp. 291–304.

[27] Gary Charness, Uri Gneezy, and Brianna Halladay. “Experimental methods: Pay one or pay all”. In: *Journal of Economic Behavior & Organization* 131 (2016), pp. 141–150. ISSN: 0167-2681. DOI: 10.1016/j.jebo.2016.08.010. URL: https://www.sciencedirect.com/science/article/pii/S0167268116301779 (visited on 01/17/2022).

[28] Todd L. Cherry, Peter Frykblom, and Jason F. Shogren. “Hardnose the dictator”. In: *American Economic Review* 92.4 (2002), pp. 1218–1221.

[29] Matthias Heinz, Steffen Juranek, and Holger A. Rau. “Do women behave more reciprocally than men? Gender differences in real effort dictator games”. In: *Journal of Economic Behavior & Organization*. Gender Differences in Risk Aversion and Competition 83.1 (2012), pp. 105–110. ISSN: 0167-2681. DOI: 10.1016/j.jebo.2011.06.015. URL: https://www.sciencedirect.com/science/article/pii/S0167268111001600 (visited on 01/30/2022).

[30] Johannes Abeler et al. “Reference points and effort provision”. In: *American Economic Review* 101.2 (2011), pp. 470–92.

[31] Christoph Engel. “Dictator games: A meta study”. In: *Experimental economics* 14.4 (2011). Publisher: Springer, pp. 583–610.

[32] Zhijian Wang, Bin Xu, and Hai-Jun Zhou. “Social cycling and conditional responses in the Rock-Paper-Scissors game”. In: *Scientific Reports* 4.1 (2014). Number: 1 Publisher: Nature Publishing Group, p. 5830. ISSN: 2045-2322. DOI: 10.1038/srep05830. URL: https://www.nature.com/articles/srep05830 (visited on 01/31/2022).

[33] Timothy N. Cason, Daniel Friedman, and Ed Hopkins. “Cycles and instability in a rock–paper–scissors population game: a continuous time experiment”. In: *Review of Economic Studies* 81.1 (2014). Publisher: Oxford University Press, pp. 112–136.

[34] Moshe Hoffman et al. “An experimental investigation of evolutionary dynamics in the Rock-Paper-Scissors game”. In: *Scientific reports* 5.1 (2015). Publisher: Nature Publishing Group, pp. 1–7.

[35] Benjamin James Dyson et al. “Negative outcomes evoke cyclic irrational decisions in Rock, Paper, Scissors”. In: *Scientific Reports* 6.1 (2016). Number: 1 Publisher: Nature Publishing Group, p. 20479. ISSN: 2045-2322. DOI: 10.1038/srep20479. URL: https://www.nature.com/articles/srep20479 (visited on 01/31/2022).

[36] Rosemarie Nagel. “Unraveling in guessing games: An experimental study”. In: *The American Economic Review* 85.5 (1995). Publisher: JSTOR, pp. 1313–1326.

[37] Johannes Abeler, Anke Becker, and Armin Falk. “Representative evidence on lying costs”. In: *Journal of Public Economics* 113 (2014). Publisher: Elsevier, pp. 96–104.

[38] David Hugh-Jones. “Honesty, beliefs about honesty, and economic growth in 15 countries”. In: *Journal of Economic Behavior & Organization* 127 (2016). Publisher: Elsevier, pp. 99–114.

[39] Erin L. Krupka and Roberto A. Weber. “Identifying social norms using coordination games: Why does dictator game sharing vary?” In: *Journal of the European Economic Association* 11.3 (2013). Publisher: Oxford University Press, pp. 495–524.