Interactive experiments in Toloka

Chapkovski, Philipp

University of Bonn

3 February 2022

Conducting interactive experiments on Toloka

Abstract

Even before the COVID-19 pandemic, the popularity of online behavioral experiments grew steadily. Due to lockdowns, online studies often became the only available option for behavioral economists, sociologists, and political scientists. The use of the most well-known platforms, such as mTurk, was so intensive that the quality of data was harmed. However, even before the pandemic-induced quality crisis, online studies were limited in scope; 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 relatively unknown crowdsourcing platform, Toloka, we successfully ran several multi-round interactive experiments. Toloka's sizeable online population, fairly low exposure of participants to sociological surveys and behavioral studies, and convenient application programming interface can make it a useful addition to the toolbox of an experimentalist who needs to run behavioral studies that require real-time interactions between participants.

Keywords: Crowdsourcing, mTurk, online research, survey research

JEL Classification: C90 , C92 , C81 , C88 , B41

1 Introduction

A multitude of experimental designs that assume the interdependence of decisions and payoffs allow for asynchronous decision-making. For example, in order to avoid synchronicity in one-shot games, experimentalists widely use the strategy method, in which responders makes conditional decisions for each possible information set (Brandts & Charness, 2011). This method has proven to be relatively reliable compared to the direct interaction of multiple players (Fischbacher, Gächter, & Quercia, 2012). And yet, numerous behavioral studies require interactions among their participants in real-time. Lab experiments on auctions (Cooper & Fang, 2008), wage negotiations (Niederle & Vesterlund, 2007), voting patterns (Battaglini, Morton, & Palfrey, 2010), and voluntary contributions with peer punishment (Fehr & Gächter, 2000) require the simultaneous presence of participants in a relatively stable group, wherein participants can receive real-time feedback about their partners' behavior across time.

Despite the recent rise in popularity of crowdsourcing platforms for conducting behavioral studies, so far, the obstacles to running experiments with real-time interactions outweighed their feasibility. In a review of Prolific, a crowdsourcing platform widely used for behavioral studies, the authors mention that "[s]imultaneous experiments are currently not a focus of Prolific (and neither of other crowdworking platforms, for that matter)" (Palan & Schitter, 2018).

In this paper, we share our experiences on the relatively new crowdsourcing platform, Toloka, which provides access to a large pool of experiment participants. Their quick response rate, together with a handful of mechanisms that we implemented during the data collection, provided us with an opportunity to successfully conduct interactive experiments online via Toloka. However, four standard behavioral games (public good game, dictator game, rock-scissors-paper, and a coin-flipping game) were conducted in Toloka among Russian-speaking audience only, so further investigations are necessary whether this platform can be a reliable supplier for recruiting non-Russian participants.

The primary concern with conducting interactive experiments using existing crowdsourcing platforms is the high drop-out rate. Arechar, Gächter, and Molleman (2018) describe their experience with replicating the study of a public good game with and without peer punishment conditions (Fehr & Gächter, 2000) using the crowdsourcing platform Amazon Mechanical Turk (mTurk). Over the course of their study, approximately 18% of the participants left the experiment before completion. In turn, each drop-out altered 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-out" problem commonly appears in the initial matching phase, during which individuals are matched with other participants. If this stage takes too long, participants may get tired of waiting and leave. This is particularly relevant when participants join a study at irregular intervals, in which

case early and late arriving participants are at a particular risk of waiting too long for a match. Furthermore, mTurk users self-reported that they participate more often in surveys while multitasking, and that they often leave a study page to return to it later (Necka, Cacioppo, Norman, & Cacioppo, 2016). In their study, Arechar and colleagues convincingly demonstrate that it is possible to utilize crowdsourcing platforms when the research design requires real-time interaction in a stable group. Other studies have been conducted since Arechar's, such as Lee, Seo, and Siemsen (2018), and confirm that the high drop-out rate for behavioral games that span several rounds can be a daunting, if not insurmountable, issue for researchers.

However, with the onset of the COVID-19 pandemic and the subsequent closing of most on-campus labs, interactive experiments online became essential. As well as this, an unmet demand for running interactive experiments using online populations predated the pandemic. Online studies provide access to online populations beyond the typically white, educated, industrialized, rich, and democrat-leaning population attainable in the university pool (Henrich, Heine, & Norenzayan, 2010). For instance, the online population of mTurk more accurately resembles the general US population than students from large US universities, with particularly salient differences noticed in behavioral measures of honesty and altruism (Snowberg & Yariv, 2021).

While the COVID-19 pandemic boosted interest and demand for interactive experiments, it simultaneously exacerbated the problem with existing crowdsourcing platforms due to the substantial increase of new studies on the limited pool of participants. This resulted in decreased quality of the data. The data quality crisis in mTurk began even before the pandemic, as a consequence of easy accessibility to VPS (virtual private servers) that allowed non-US residents to participate in US-only studies. By some estimates, approximately 12% of mTurk respondents are VPS users, and the share of low-quality submissions among them is about eight times higher than among non-VPS users (Kennedy et al., 2020). Even at the dawn of mTurk's use as a recruitment platform, researchers discovered that a small minority of workers were responsible for submitting most of the HITs among 132 academic studies (Chandler, Mueller, & Paolacci, 2014). Other estimates put the number of reachable respondents in mTurk at less than 8,000, in stark contrast to more than 500,000 formally registered users (Stewart et al., 2015). The overexposure of participants to behavioral surveys and studies make results liable to bias. Moreover, there has been a demonstrably negative effect upon pool overexposure on trustfulness (Benndorf, Moellers, & Normann, 2017) and attention (Barends & de Vries, 2019) of the respondents. In this paper, we describe our experience of running behavioral studies within Toloka, which utilizes features that may ameliorate the above-mentioned problems with online research. It should be noted, however, that this platform is far from being the only alternative platform for recruiting more "naive" participants. Prime Panels by Cloud Research is one

of such tools (Chandler, Rosenzweig, Moss, Robinson, & Litman, 2019). Prolific also provides a set of tools to recruit participants with different levels of exposure to surveys and experiments.

2 General platform description

Because Toloka's main focus is on providing workers for machine learning, the platform has gone relatively unnoticed in the academic community. However, its fairly large online population makes the process of finding partners in interactive games less tedious, and a convenient programmatic access (API) to its resources allows easily integrate it with the existing experimental software, such as oTree (Chen, Schonger, & Wickens, 2016).

The academic community might take interest in Toloka because it provides access to some populations that can be hard to access through Prolific or mTurk. While the mTurk population is mostly based in the US and India (Difallah, Filatova, & Ipeirotis, 2018), and Prolific users reside mostly in the UK and US (Peer, Brandimarte, Samat, & Acquisti, 2017), Toloka's users mostly reside in Russia and other former Soviet countries (Russian-speaking participants make up 60% of active users, depending on the time of the day). Although a number of companies provide access for doing survey-based research in the post-Soviet space, such as Lucid[1], the shortage of well-established behavioral labs in these countries substantially limits the possibility for conducting experimental behavioral studies. With access to these populations, researches might study, for instance, whether or not the experience of living under state socialism motivates people from ex-socialist spaces to behave differently as some behavioral studies suggested (Gächter & Herrmann, 2011; Herrmann, Thöni, & Gächter, 2008). Toloka may give the researchers one more channel to access the ex-communist countries, as well a large, relatively inexperienced online population with a so-far limited knowledge about behavioral studies.

Another feature that can make Toloka worth a consideration as a tool to recruit participants is its sizeable population of immediately available participants. Although neither Prolific nor mTurk report the number of users currently active online, by some estimates, the average online presence of the mTurk population is approximately 2,000 available participants at any given time (Difallah et al., 2018). According to our measurements (Section 2.2), the Toloka interface provides access to up to seven times more users than mTurk, varying from about 5,000 participants immediately available at night (in the UTC time zone) to 22,000 at the peak of the working day. This is comparable only to Prolific projects where a rate-limiting mechanism is deactivated (see details on this in Section 4.1).

[1] https://luc.id/

2.1 Terminology

In this paper, we intentionally avoid explanations of most of the technical details regarding the Toloka interface or its API functionality, as the English documentation provided by Toloka is detailed enough[2]. Instead, we provide only the crucial information about the main Toloka components required for general understanding of the recruitment and matching procedures.

Similar to its competitors Toloka recruits participants for a task, mostly online, and processes fixed-fee payments per task after the job is completed, occasionally providing an additional bonus. Most crowdsourcing platforms do not limit the nature of jobs posted, with the obvious exceptions of illegal activities and tasks that risk revealing the identity of their workers. The tasks vary from image labelling and tagging texts for natural language processing to, as in the present case, participating in surveys or behavioral games. A single task can be fulfilled by a large number of 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 through the *task overlap* parameter.

To create a task, a researcher 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 fields (variables shown to each participant) and output fields (responses provided by each participant).

As soon as a project is created, participants may be invited to a specific study (or any task in general) through opening a *pool*, which is a combination of settings including a participation fee, number of participants (*task overlap*), and some filters limiting access to a specific portion of the online population. These filters can be publicly available user characteristics collected by Toloka itself and include filtering a user's location by their IP address or by the country code of their registered phone number, as well as by information based on their self-reported nationality, age, gender, educational level, and knowledge of languages. Unlike Prolific, which provides a researcher with an extensive list of dozens of filters through which to select participants, Toloka's list of available filters is quite limited. In addition, similar to mTurk, Toloka allows experimenters to assign custom markers to specific users (*skills* in Toloka terminology, *qualifications* in mTurk terminology). These markers can be used to filter out participants who took part in previous studies, or they can serve as filters to re-invite participants for follow-up studies. Unfortunately, these markers are account-specific: unlike mTurk, Toloka does not allow to share them across different accounts, thus participants blocked in one account can not be blocked (or, vice versa specifically invited for a particular study) in an account of another researcher.

[2] All the Supplementary Materials including raw data for the studies, and R code to replicate the graphs and tables are available online at XXXXX. oTree code and HTML/JavaScript/CSS code is also available at GitHub: XXXXXX.

As soon as the pool parameters are set, the researcher must provide a *tasks file* with input fields to start the study: in the case of a survey, this file simply contains a link to a server which hosts the study.

2.2 Size and characteristics of Toloka audience

Toloka is primarily distinguished from its competitors by the information it provides concerning the number of users online in real time, based on a set of applied filters. We monitored the online presence of Toloka users for several days in November 2022, requesting the number of active users every 15 minutes, based on several different characteristics.

The average online population size by time of day and day of the week can be helpful to highlight the best time in which to conduct experiments, since the population recruited at different times of the day demonstrates different levels of experience with behavioral studies (Chmielewski & Kucker, 2020). The demographic composition of the available online audience also varies over time and day (Arechar, Kraft-Todd, & Rand, 2017; Casey, Chandler, Levine, Proctor, & Strolovitch, 2017). We report the general online population size (without any restrictions), as well as the Russian online population, measured by several methods (current IP address of the user, registered phone number, self-reported nationality, and self-reported knowledge of the language), along with other large post-Soviet countries/languages (e.g., Ukraine, Belarus, Kazakhstan, as well as India, as one of the largest, non-Soviet countries present at that time in Toloka). We also evaluated the population size by each individual's self-reported gender, educational levels, and fluency in English and Russian. All time and data values reported are in UTC.

During the observation, the average presence of immediately available participants remained roughly the same across all weekdays (about 15,000 total participants, of which about 7,000 had Russian IP addresses). There was a slight drop in the online population size on Sunday. The graph in Figure 1, in addition to an average online population size, reports the minimum and maximum values observed on that day, both for the general unrestricted online population and for those participants who self reported knowledge of English, or were not Russians (neither located in Russia nor self-reported knowledge of Russian language), as well self-reported non-Russian English speakers.

The size of the online population heavily fluctuated within a single day (Figure 2). It never fell below 6,000 active users, and during working hours (9 a.m. to 8 p.m.) it remained above 15,000, reaching its maximum of 22,000 at about 1 p.m.. The size of the Russian-speaking population dropped at night, while English speakers and others remained the same, resulting in an increasing percentage of more than 50% non-Russian speakers at night. During the day, the Russian-speaking percentage was approximately 60%. Across the week, the proportion of Russian speakers remained relatively constant. The percentage of those in population who declared knowledge or neither Russian nor English was about 6%, never exceeding 11%.

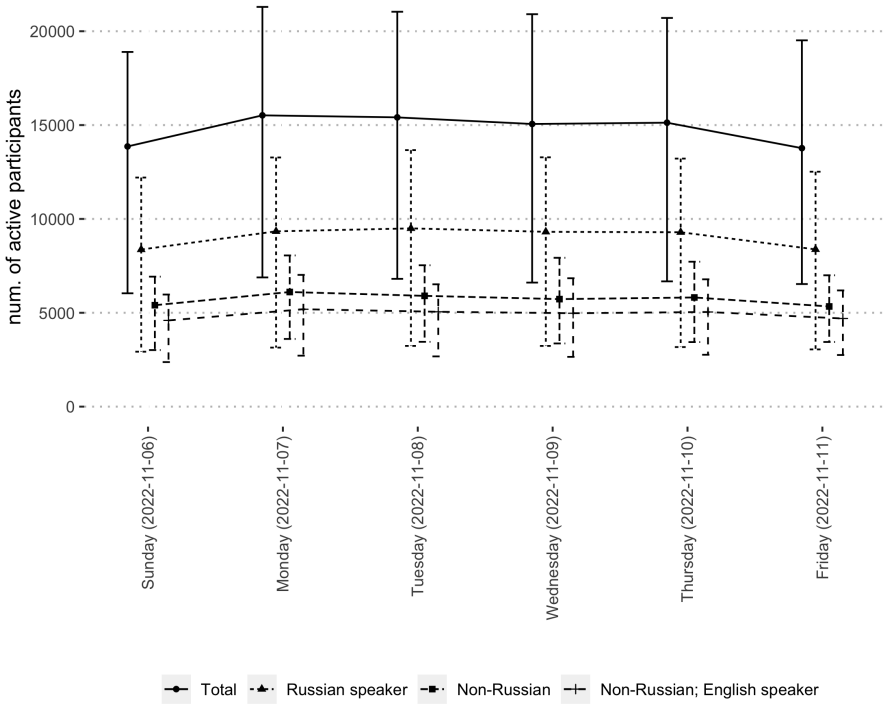

Fig. 1 Number of participants available at any given time by the day of the week. Whiskers show largest and smallest value observed that day. 'Russian/English speaker' is defined as any user who asserts the Russian/English language knowledge in their profile; 'Non-Russian' is defined as any user who was not located in Russia (by IP and self-registration) and was not self-registered as a Russian citizen.

We also estimated online population sizes based on the country in which participants were located. This can be done in three ways using the Toloka platform: by monitoring users' current IP address, phone number used at registration, or self-reported citizenship. All three methods produce different results: VPN services can be used to access the Toloka website; mobile phones can be used in different countries while roaming; or participants may live in a country which differs from their nationality. Location by IP or phone number resulted in similar estimates of population composition (see Table 2.2), while number of Russian citizens was substantially smaller (31% vs. 47% by other methods). Apart from Russian population the other four largest countries (Ukraine, Belarus, India, and Kazakhstan) were responsible for less than 10% of the population (about 1,000 active participants available online in any given moment of time).

Participants can also be filtered by three additional parameters: gender, age, and educational level. However Toloka does not require users to state their gender and education level at registration. Therefore, about 65% of active users did not reveal their education, and roughly 37% had no information

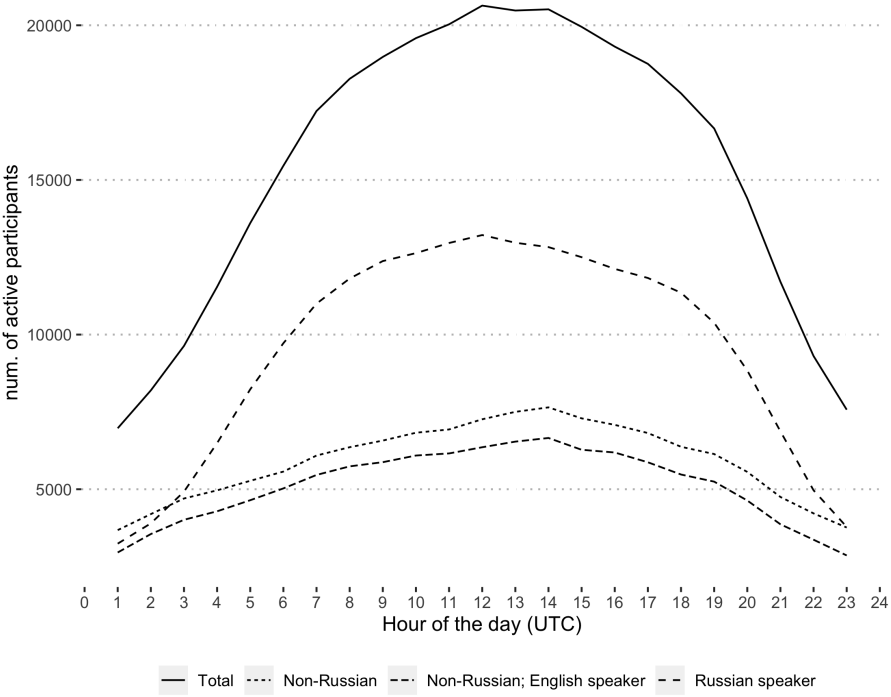

Fig. 2 Number of participants available at any given time by the hour of the day. 'Russian/English speaker' is defined as any user who asserts the Russian/English language knowledge in their profile; 'Non-Russian' is defined as any user who was not located in Russia (by IP and self-registration) and was not self-registered as a Russian citizen.

parameter	Belarus	India	Kazakhstan	Other	Russia	Ukraine
ip	1.8%	4.7%	1.7%	43.7%	47.2%	0.9%
phone	1.8%	4.6%	1.6%	42.1%	47.8%	2.1%
nationality	1.1%	0.7%	0.8%	64.6%	30.9%	1.9%

about gender. Among those 35% who self-reported their educational level, 61% marked that they had a "high education" level, and 23% that they have 'middle education" (although Toloka does not specify what these levels specifically mean).

On average, the gender composition was relatively stable across all days of the week, or hours of the day. Among those participants who self-reported their gender, there were marginally more men than women (54% to 46%).

3 Overview of Russian-speaking audience of Toloka

In addition to estimates from the online population, we conducted a study of 1,000 participants in order to increase our understanding of the average Russian-speaking Toloka population[3]. In total, we collected 990 observations, as 10 participants dropped out. We applied only one limiting filter for participation: the participants must have claimed to speak Russian in their Toloka profile. As demonstrated in Section 2, Russian speakers accounted for about 60% of the active Toloka population, and the average number of Russian-speaking participants available online at any given moment was 9,028 (with a minimum of 2,927 and maximum of 13,667, depending of the day and the hour). Thus, this survey covered slightly less than 10% of the active Russian-speaking Toloka participants.

A series of questions (Tables 2–4) investigated the working patterns, sociodemographic characteristics, values, and political attitudes of Toloka users during our research (see Supplementary Materials for the full questionnaire and detailed information about each variable).

In terms of socio-demographic characteristics, the proportion of male users was 58% (a bit higher than 54% that we observed in self-reported data in overall Toloka population reported in the previous section), 17% of male participants 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 higher education. 67% were fully employed (27%), working part-time (12%), or reported themselves as self-employed (28%). 81% of the respondents reported that they were located in Russia. Less than 4% were located outside the ex-USSR.

Five questions concerned participants' working patterns in Toloka, as follows: 1) Is Toloka your main job?; 2) What share of your total income is earned on Toloka?; 3) How many hours a week do you work on Toloka?; 4) How much do you earn on average per hour doing tasks on Toloka; and 5) In an average month, in how many surveys do you participate? Approximately 31% of participants responded that Toloka is their primary job, and nearly every fifth respondent claimed that Toloka provides either a very significant portion or the entire source of their income (Table 1). Their reported hourly income coincides almost exactly with the income reported by Toloka itself ($1.81): the average income reported in our survey was $1.78 (SD: 6.40, CI: [1.38, 2.18], median: 1). On average, our participants also 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 that mTurk faces is the "lab rats" issue: too many participants have too great an exposure to all kinds of behavioral studies (Chmielewski & Kucker, 2020; Kennedy et al., 2020). It would be logical to

[3]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 1 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 2 Participation in surveys

expect that the Toloka population is less experienced with these kinds of surveys, given the relative novelty of the platform. Indeed, about 60% of users participate in any kind of surveys twice a month or more rarely (Table 2).

Toloka users reflect the full Russian political spectrum, including supporters of the governing party, United Russia (18%), to Communists (7%), to those who would vote for Vladimir Putin (30%) if the next presidential elections would happen next Sunday, and supporters of imprisoned opposition leader Alexey Navalny (7%) (See Table A1 in Appendix). Of 990 participants, 363 responded that they voted in 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 a 43.1% turnout rate, which nearly parallels the officially reported turnout rate of 45.15%.

Additionally, we collected a set of answers regarding COVID-19. We asked four questions: 1) Are/were you sick with COVID-19?; 2) Was someone in your family or circle of close friends sick with COVID-19?; 3) What are your vaccination plans?; and 4) Do you think that vaccination should be mandatory? The results demonstrated that, at least in some dimensions, the Toloka population mirrored the general Russian population. For instance, among Russian citizens who participated in the survey, the share of vaccinated participants was 42.7% (N=368), which is similar to the 46% estimate from traditional pollsters in November 2021 (Levada, 2021).

The question concerning mandatory vaccination was not fully comparable with traditional polling data (Levada, 2021) because the Levada survey used a 4-point scale alongside a "Hard to say" choice, while we used a 5-point scale with the midpoint "Neither agree nor disagree" option, without the "Hard to say" choice. Nevertheless, the number of those who either totally or somewhat agreed in both surveys is almost exactly the same: 42.5% (N=421) in the Toloka survey and 42% in the Levada survey (total N=1603).

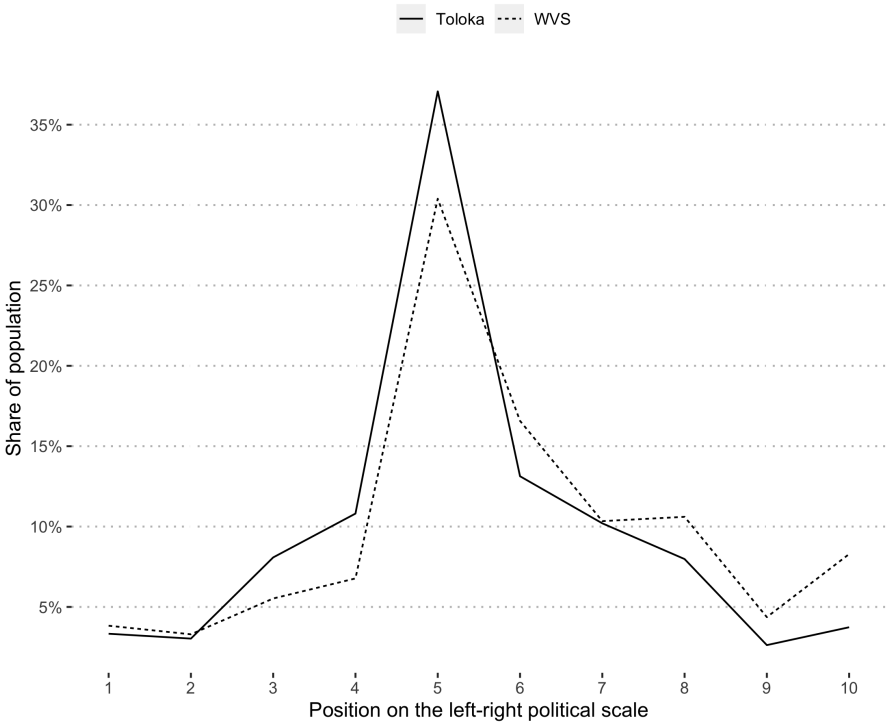

Fig. 3 Position on the left-right political scale. 'WVS' stays for World Values Survey (wave 2020).

In addition to the above questions, we surveyed participants regarding standard generalized trust questions from the World Values Survey (WVS) (Haerpfer et al., 2020), Specifically, we asked participants whether or not they believed that people can be trusted, or if, in their view, it would be better to be careful with strangers. 78% (N=775) chose the cautious response, compared with 74% (N=1358) of the Russian population who chose this same option in the WVS survey. We also asked participants to position themselves on a left–right political scale. It should be noted that this question is not fully comparable with its WVS counterpart since WVS included both "I do not know" and "No answer" options, which were chosen by 39% of the population, whereas our survey omitted these options. If we consider only those participants who made a choice between left or right on this question in the WVS, the results becomes rather similar to what we observed among Toloka members; however, Toloka members on average are slightly more left-leaning (with a mean average of 5.37 vs. 5.85 in the WVS, see Figure 3.

4 Behavioral experiments in Toloka

This section briefly presents the results of several interactive studies carried out on Toloka, as well as comparison of speed with which respondents accepted the task (arrival time) and started filling the questionnaire (acceptance time)[4].

We describe here two types of projects: timed and untimed. Standard, untimed projects do not impose any starting time, and include such projects as surveys and interactive games in which the decisions can be made asynchronously, even if the payoffs for different participants are interdependent. In a timed project, all participants need to join the study within a narrow time window, in order to be matched into groups in which they can interact. Studies 1 (Public Goods Game), 2 (Dictator game with the preceding real effort task), and 3 (Rock-Scissors-Paper) are of this kind. Study 4 ("Cheating" game in which people report the results of flipping a coin) is non-interactive, and thus does not require participants to join the study within a certain time period [5].

Regardless of the study type, Toloka should be linked with a server where the survey (or the game) is located. This task has two steps: first, we distribute the link to the study among Toloka participants, and second, we provide a server with Toloka user identification numbers so that participants can be identified and paid at a later stage in the form of bonuses. Both aims were achieved in our study by using the JavaScript code, which appends a unique identification number to a link that leads to the study. This identification number associates the participant of the game with a Toloka participant. This procedure is similar to an integration of external studies in Prolific via URL parameters (Palan & Schitter, 2018). To direct participants to a specific study, a link to this study should be inserted in the *tasks* file as an input field.

In order to make the time of arrival window more precise, we gave participants access to the study link, which announced start times in the study description for studies 1, 2, and 3. We opened a Toloka pool a few moments before the announced start time, and those who joined the study before this time were shown a counter that calculated the number of seconds they had before the link could be accessed. The provided JavaScript code calculated this time depending on the time zone in which a participant was located.

When participants clicked the link, they were redirected to the study, and within 60 seconds of the official start time, the Toloka session was closed to new arrivals. After the study began, participants accepted a consent form (if applicable), read introductory instructions that included exchange rate, and were matched with other members of their group.

Those who were not able to find a partner at the matching stage were redirected to a page informing them of this, but reassuring them that they

[4]The data on arrival and acceptance time is collected as a part of a different study, not connected to the studies 1 to 4. 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

[5]See Appendix for screenshots of a timed and untimed projects in Toloka.

would still be paid the participation fee. During the game, we limited participants' decision-making time to an average of 60 seconds. If a participant did not answer within this time frame, we counted them as a drop-out, and they were redirected to a page informing them that they had been dropped from the study. Their partner would then be redirected to the "Partner Drop-out Page", on which we reminded them that they were still eligible to receive the participation fee.

4.1 Arrival times

The main impetus that facilitated partner matching was the relatively short time required for an average participant to join the study after it was posted on the platform. To demonstrate the difference in arrival times across platforms, we conducted several short, small-scale (N=100 for each case) surveys[6], which were not connected to the four studies presented below. We measured the time difference between the arrival of each respondent compared to the arrival of the first participant. We collected the data for three of these surveys in December 2021, and did an additional data collection on Prolific in November 2022. The additional data collection was required because in 2022 Prolific introduced an option to remove so called 'throttling' mechanism that limited the rate, at which participants could join the study. This mechanism gives priority access to the studies for more naive participants[7]). The collected data are not fully comparable across these platforms for obvious reasons: the national composition is different (as we mentioned above the majority of Prolific participants are from the UK and USA, mTurk consists mostly of US and India participants, while in Toloka there are mostly participants from ex-USSR or Global South). Similarly, it is hard to calibrate the compensation in such a way that it has the same purchasing power across different online populations. Although, it may be noted that switching off the throttling mechanism in Prolific resulted in a very similar arrival time to one we observed in Toloka, despite all the demographic differences across these two platforms.

The mean arriving time was 562 seconds for Prolific with an activated throttling (rate-limiting) mechanism, which is the default option, 1143 seconds for mTurk, 50 seconds for Toloka and 40 seconds for Prolific pool with a deactivated throttling mechanism (panel **A** of Figure 4).

Less than 100 seconds after the start time of the study, all the Toloka participants had joined the study. With a deactivated throttling mechanism

[6]Upon the acceptance participants were redirected to a short (less than 2 minutes long) survey asking their socio-demographic characteristics, such as age, gender, and educational level. We adjusted the payments based on the purchasing power of the main target audience on each platform. mTurk population was recruited only among US residents, and were compensated $0.20 per survey; Prolific population was paid 0.23 UK Pounds; the wave 1 on Prolific collected in December 2021 used a screener of English language knowledge, while the Wave 2 in November 2022 was limited to the UK and US population only; Toloka participants were limited to Russian speaking population and received $0.10 per survey.

[7]See details at the Prolific support website at https://researcher-help.prolific.co/hc/en-gb/articles/360009223593-Dyadic-experiments. I am grateful to an anonymous reviewer for pointing out this novel feature of Prolific.

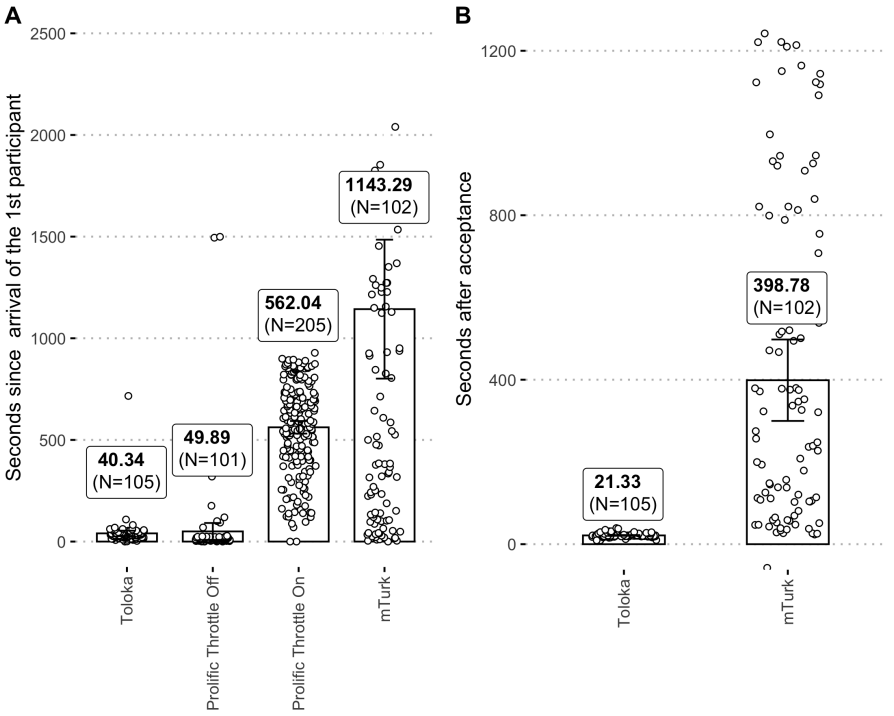

Fig. 4 Time since the arrival of the first participant and acceptance time. The whiskers show 95% confidence intervals, labels show the means and number of observations, dots show the individual observations. 'Prolific Throttle On' category includes both measurements done in December 2021 and November 2022.

Prolific was even more effective, but the with default option (throttling/rate-limiting mechanism on) it was substantially slower, as participants took up to 7 times longer to join. After more than 20 minutes from the start time of the study, still less than 75% of requested slots in mTurk had been filled (see Figure 5).

An issue that may arise while running interactive experiments online in crowdsourcing platforms is that participants can accept the HIT (in mTurk) or assignment (in Toloka), but may not start working on it for some time. For instance, participants may reserve a task while completing other, previously assigned tasks. This delay can substantially complicate the matching procedure because some participants do not join the study on time, and in turn, this prevents others from joining as the available slots are fully booked. This problem does not occur on the Prolific platform, where users accept the task only when they click on the study link. Since both Toloka and mTurk provide information on when a participant accepts a task, we traced the length of time it took for each participant to begin working on the task after they had accepted it. The average time-since-acceptance for mTurk was 399 seconds (SD: 499),

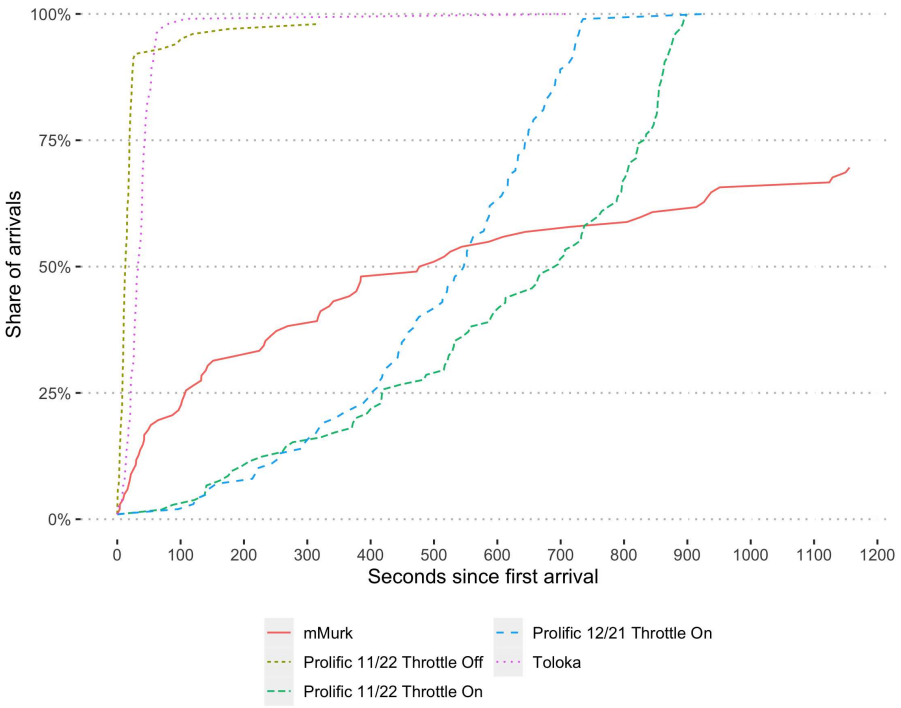

Fig. 5 Cumulative share of participants who join the study by time (mTurk outliers are omitted)

substantially longer than for Toloka, where it took an average of 21 seconds (SD: 4.56), see panel **B** of Figure 4).

4.2 Studies - short description

Using the Toloka platform, we ran a set of four simple standard games that are widely used by behavioral scientists. This included the Public Good game (PGG, Isaac and Walker (1988)), the Dictator game (DG) (Engel, 2011) preceded by the real-effort task (RET) Abeler, Falk, Goette, and Huffman (2011)), rock-scissors-paper (RSP, Wang, Xu, and Zhou (2014), and a coin-flipping game (Abeler, Becker, & Falk, 2014; Fischbacher et al., 2012).

This specific set of games was chosen to demonstrate different behavioral traits of Toloka's population, such as cooperation (PGG), altruism (DG), non-cooperative strategic interactions (RSP), and honesty (coin-flipping). In the list below, we provide a concise description for each game and its corresponding payoffs (all payoffs are in the US dollars because this currency is used on Toloka for paying the participants regardless the country they are located in).

PGG The typical public good game includes N subjects, and each subject is provided with an endowment w. The subject can invest any amount g_i from

0 to w to serve the public good. Whatever is left over from the endowment after the investment remains in the subject's private account. In our case, the public good investment coefficient k was 1.5, and the group size was 3, so the MPCR (Marginal Per Capital Return) was 0.5, similar to the MPCR in a majority of other linear PGG, as noted by Zelmer (2003). Thus, in the case of a subject's full cooperation, the maximum payoff in a single period was 150 US cents, and in the case of a subject's complete lack of cooperation, participants received 100 US cents.

DG+RET We used a standard real-effort task of counting zeroes (Abeler et al., 2011) in a large matrix of numbers. Participants worked in groups of two, and were matched to the role of either dictator or recipient in a subsequent dictator game, based on their productivity in the RET stage: the more productive participant became the dictator. If two paired players were equally productive in this first stage, the dictator role was assigned randomly provided with the endowment of 100 US cents that he or she can split between themselves and a recipient. We used this specific design in order to demonstrate how matching can be done not only on the somewhat random basis of arrival time, but also based on some specific measure (in this case performance) in the earlier stage of the study.

RSP During the 'Rock-Scissors-Paper' 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). Then, their decisions were matched with each other, and payoffs were calculated according to a simple rule: Rock "beats" Scissors, Scissors "beats" Paper, and Paper "beats" Rock. If both participants chose the same item, there was a tie. The winner received \$1, the bonus for a tie was \$0.50, and the loser received \$0. One of 10 rounds was chosen randomly to define the final participant bonus.

Coin-fliping Participants were asked to flip a coin and report the outcome (Abeler et al., 2014; Fischbacher et al., 2012). This report defined the payoff: \$1 for reporting heads, \$0 for tails. These decisions and beliefs were collected in three separate experimental sessions for three different Russian regions (Moscow, Voronezh, and Arghangelsk), using location filters from IP addresses provided by Toloka.

Below, we report the time participants had to wait before finding a partner (matching time), as well as a brief overview of the decisions participants made in each of the four studies.

4.3 Matching times and information on drop-outs

Figure 6 demonstrates the full distribution of matching times for those participants who had to wait for their partners across the three timed studies. Although participants in a DG with the RET had to wait slightly longer on average there were no substantial differences in matching times.

PGG

In the PGG study, the total number of participants who joined was 117. 92 of these participants successfully completed the study, 7 were blocked due to inactivity, and 18 were blocked by the non-activity of their group members. In addition, 14 participants were blocked by 7 inactive participants mentioned above, and 4 participants were directed to the final page after waiting 90 seconds for a partner in the matching stage. There were no participants who dropped out of the study in the middle of the game; all 7 drop-outs occurred at the beginning of the first round. Out of 92 participants, 60 had to wait for their partners, and their mean waiting time was 2.50 seconds (median: 0.9 seconds).

DG+RET

In the DG study, 96 out of 102 participants were successfully matched, 4 were blocked due to inactivity, and 2 were blocked by the inactivity of other members. 48 participants had to wait for their partners (the rest were immediately matched), and their mean matching time was 4.07 seconds (median: 1.01 seconds). The maximum waiting time was 80 seconds.

RSP

In the Rock-Scissors-Paper study, the total number of participants who initially joined the study was 111. 102 of these participants were successfully matched and completed the study, with 9 participants remaining unmatched after arrival. There were no drop-outs during the game itself. 47 participants had to wait for their partners for an average of 2.38 seconds (median: 0.43 seconds).

4.4 Results

We present here a brief overview of behavior across the four studies (PGG, DG, RSP, and the coin-flipping game). The distribution of the decisions is shown in Figure 7.

PGG

The average contribution to public good was 45.94 (SD: 2.09, CI: [43.85, 48.03]) which is slightly above the mean contribution of a meta study by Zelmer (2003). We observed a rate of deterioration of cooperation typical for other linear PGGs without peer-punishment stage (Figure 7, Panel A) - see, for instance, Fehr and Gächter (2000). 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]).

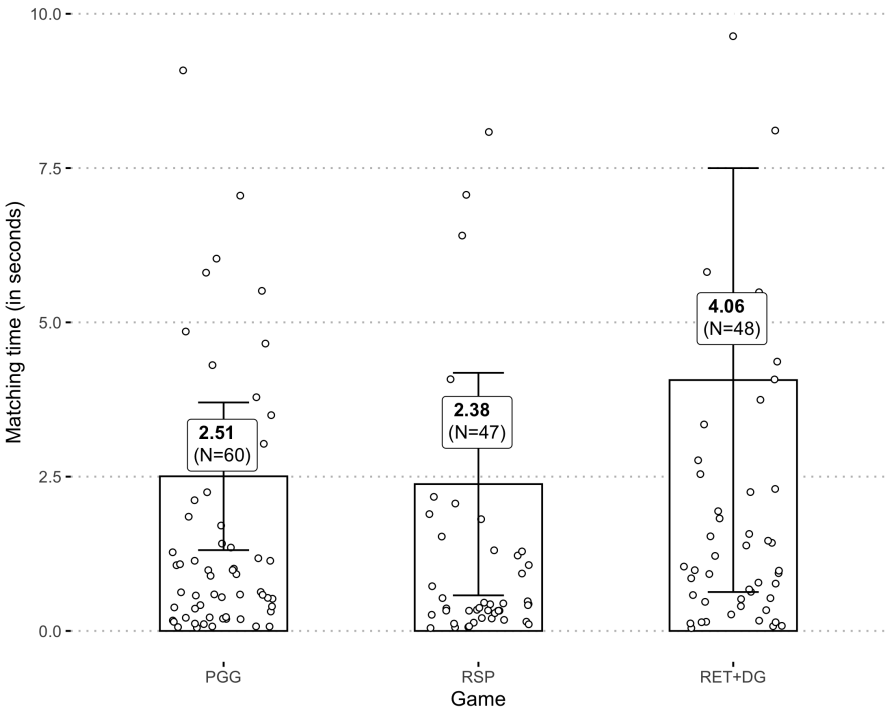

Fig. 6 Matching time (in seconds)for participants who had to wait for a partner: the whiskers show 95% confidence intervals, labels show the means and number of observations, dots show the individual observations. 8 outliers (matching time > 10 seconds) are not shown on the graph but used for the calculation of the means.

DG+RET

Participants were matched into pairs after demonstrating their productivity in the RET stage. Their productivity defined their role, with the more productive partner receiving the right to distribute a dictator's endowment between themselves and another participant.

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 DG of 28.35% (Engel, 2011) across the world - see Figure 7, Panel C. The amount that recipients believed their dictators would provide was slightly higher than the actual amount they received: 35.8 (SD 25.3, N 48, median 40).

RSP

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

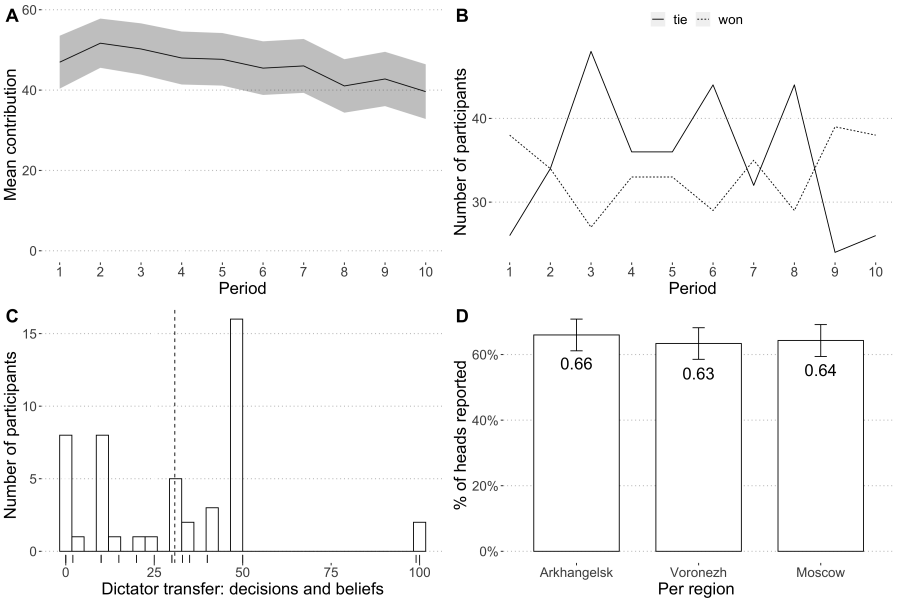

Fig. 7 Behavioral outcomes of studies 1-4: **A**: Mean contribution in Public good game (PGG) per round. Gray area is 95% confidence interval. **B**: Number of participants who won or played a draw per round in 'Rock-Scissors-Paper' (RSP) game. **C**: Frequency graph of dictator's decisions in a Dictator game (DG). Dashed line is a mean contribution. **D**: Share of participants who reported heads in a coin-flipping game in each region. The whiskers show 95% confidence intervals.

individuals are in their actions. The share of ties grew after the first round, and dropped only at the very last round - see Figure 7, Panel B.

Coin-flipping

Overall, we observed that the frequency of heads reported was 65%, which did not vary greatly across the three regions that participated in the study (see Figure 7, Panel D). This number is close to the number reported about Russian online-population behavior in a similar study, where data was also collected online. There Russians reported heads in 71% of the cases (Hugh-Jones, 2016).

We also collected data about participants' beliefs on the average proportion of heads reported would be, according to a standard procedure of belief elicitation (Krupka & Weber, 2013). The belief elicitation was incentivized: if a participant guessed correctly (-/+ 10 percentage points), they received an additional bonus of $0.50. Overall, beliefs about the frequency of "cheating" are slightly higher than the actual numbers, which replicated the results of the similar study that uses Russian population data (Hugh-Jones, 2016): reported the average expectations of the Russian participants was that Russians would report heads 82% of the time.

5 Discussion

The Toloka platform is a promising tool for behavioral scientists due to two aspects. First, the simultaneous presence of a fairly large population available online (up to 22,000 participants at the peak of the working day) and their eagerness to join a study in a matter of seconds may ease up the problem of real-time interactions. Second, it provides simple access to an online population from the former Soviet Union, which can be harder to access than more 'traditional' targets for behavioralists such as the USA or Western Europe.

However, Toloka's limitations should not be overlooked. First, the platform lacks a clear and transparent procedure to filter out participants with low-quality submissions. In mTurk, unlike Toloka, qualifications such as the number of HITs submitted and approved are publicly available and can be used as filters before an individual may participate in any given study.

Second, unlike Prolific, the number of available filters on Toloka is very limited, and the vast number of participants who do not self-report such basic characteristics as gender and education limit the filters available by default even further. Thus, a researcher hoping to target any specific online population would most likely have to conduct prescreening studies in order to select the desired pool of participants, facing additional financial and logistical burdens.

Third, the code used in this paper to send the assignment identification number back to oTree is custom-written by the authors. Prolific, for instance, provides a way user-friendlier way to connect to oTree (using URL queries). Furthermore, unlike mTurk, which allows the sharing of qualification identification numbers across different accounts, and Prolific, which allows the blocking of users from participation by use of their personal identification numbers, there is no such option in Toloka. On a practical level, that means that there is no blocking option for researchers to filter out participants who may have participated in other, similar studies, or using a different account.

Finally, the experience we describe in this paper should not be overstretched. The data for four behavioral studies were collected in December 2021. The war Russian has started against Ukraine in February 2022 may affect Toloka capacity to reliably recruit participants from the countries, which formerly made part of the Soviet Union. Furthermore, participants were recruited from Russian-speaking sub-population of Toloka, so additional studies are required to check the quality of submissions by English-speaking population of Toloka.

Despite these shortcomings, however, we believe that Toloka can be an important addendum to the tools available to behavioral scientists. Its large and rather inexperienced population of users and its convenient programmatic access to most of its features make the development and conduction of online interactive games easier for behavioral studies researchers.

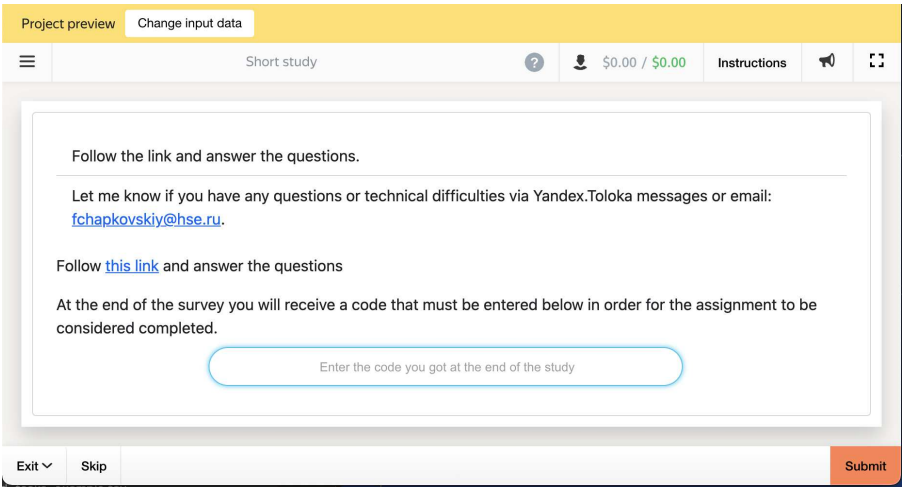

Fig. A1 Screenshot of a standard (untimed) project shown to a Toloka participant (automatic translation from Russian by Google Translate).

Appendix for "Conducting interactive experiments on Toloka"

A Screenshots of exemplary projects in Toloka

A.1 DECISION TIMES

A.2 Time for decision

We also collected the data on time per each decision.

PGG

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 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 A3)

RSP

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 seconds (SD 5.34) in the last 10th round. There were no significant differences in decision time for different outcomes (see Figure A4).

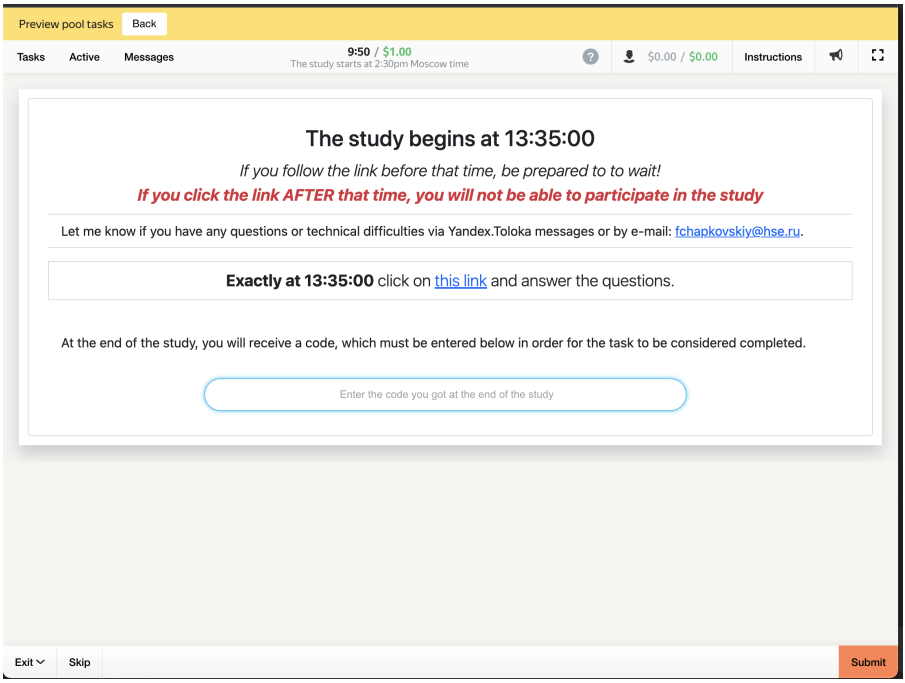

Fig. A2 Screenshot of a timed project shown to a Toloka participant. (automatic translation from Russian by Google Translate).

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 A1 Whom Toloka members would vote for on presidential elections

A.3 Additional information on decisions in Dictator game

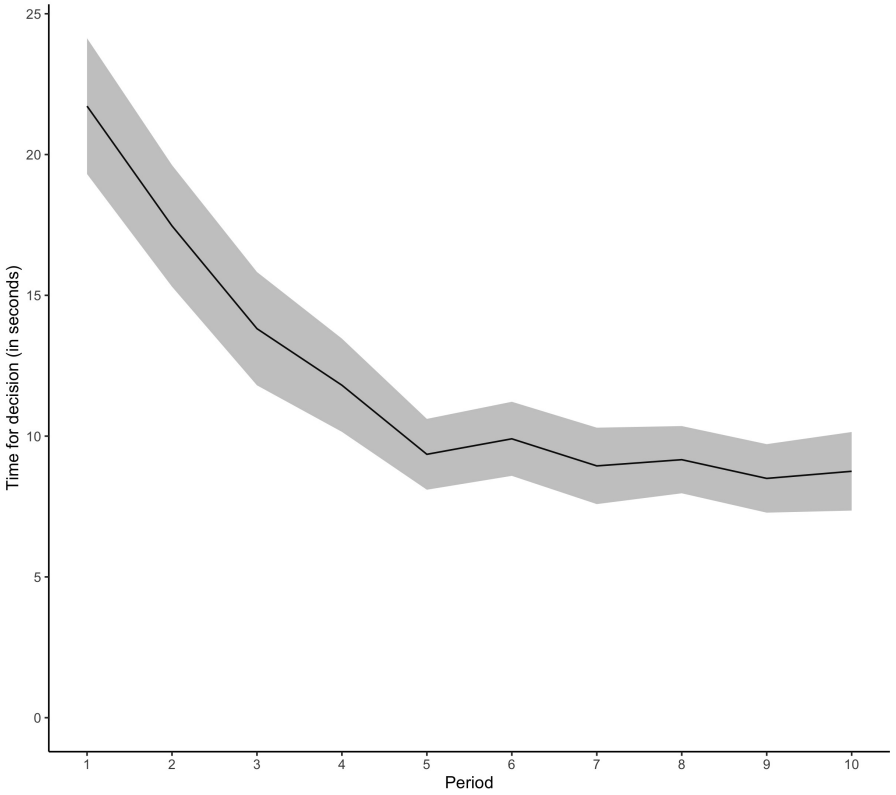

Fig. A3 Time spent on decision stage per round in PGG

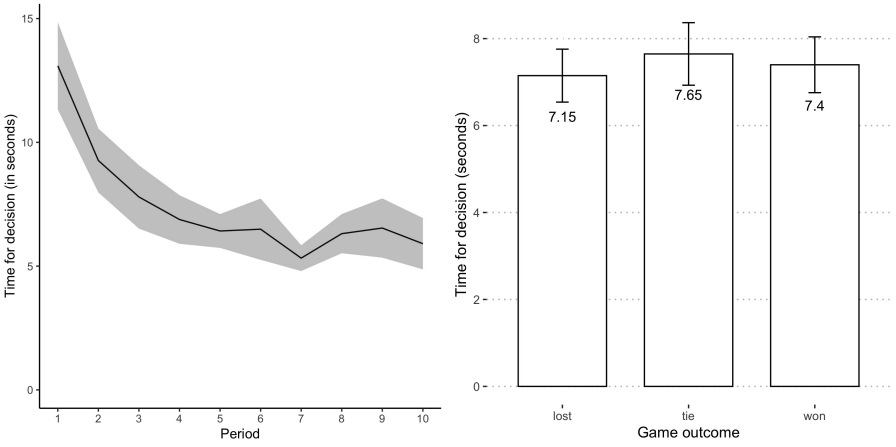

Fig. A4 Time spent on decision in RSP

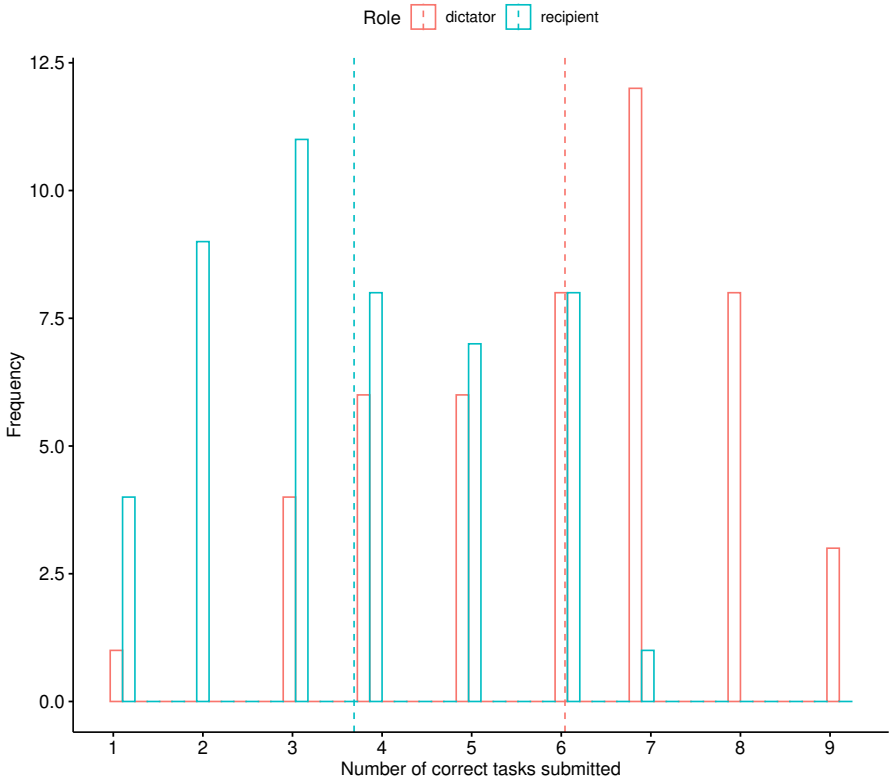

Fig. A5 Correct tasks submitted by role in RET-DG

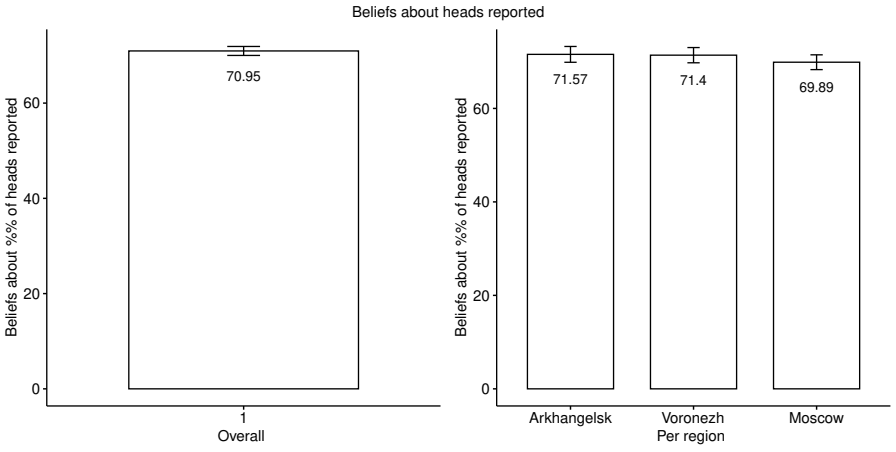

Fig. A6 Coin-flipping:user beliefs

References

Abeler, J., Becker, A., Falk, A. (2014). Representative evidence on lying costs. *Journal of Public Economics*, *113*, 96–104. (Publisher: Elsevier)

Abeler, J., Falk, A., Goette, L., Huffman, D. (2011). Reference points and effort provision. *American Economic Review*, *101*(2), 470–92.

Arechar, A.A., Gächter, S., Molleman, L. (2018). Conducting interactive experiments online. *Experimental economics*, *21*(1), 99–131. (Publisher: Springer)

Arechar, A.A., Kraft-Todd, G.T., Rand, D.G. (2017). Turking overtime: how participant characteristics and behavior vary over time and day on Amazon Mechanical Turk. *Journal of the Economic Science Association*, *3*(1), 1–11. Retrieved 2022-01-17, from https://doi.org/10.1007/s40881-017-0035-0

10.1007/s40881-017-0035-0

Barends, A.J., & de Vries, R.E. (2019). Noncompliant responding: Comparing exclusion criteria in MTurk personality research to improve data quality. *Personality and Individual Differences*, *143*, 84–89.

Battaglini, M., Morton, R.B., Palfrey, T.R. (2010). The swing voter's curse in the laboratory. *The Review of Economic Studies*, *77*(1), 61–89. (Publisher: Wiley-Blackwell)

Benndorf, V., Moellers, C., Normann, H.-T. (2017). Experienced vs. inexperienced participants in the lab: do they behave differently? *Journal of the Economic Science Association*, *3*(1), 12–25.

10.1007/s40881-017-0036-z

Brandts, J., & Charness, G. (2011). The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics*, *14*(3), 375–398. Retrieved 2022-02-02, from https://doi.org/10.1007/s10683-011-9272-x

10.1007/s10683-011-9272-x

Casey, L.S., Chandler, J., Levine, A.S., Proctor, A., Strolovitch, D.Z. (2017). Intertemporal differences among mturk workers: Time-based sample variations and implications for online data collection. *Sage Open*, *7*(2), 2158244017712774.

Chandler, J., Mueller, P., Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130.

10.3758/s13428-013-0365-7

Chandler, J., Rosenzweig, C., Moss, A.J., Robinson, J., Litman, L. (2019, October). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods*, *51*(5), 2022–2038. Retrieved 2022-11-01, from https://doi.org/10.3758/s13428-019-01273-7

10.3758/s13428-019-01273-7

Chen, D.L., Schonger, M., Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97. Retrieved 2016-12-10, from http://www.sciencedirect.com/science/article/pii/S2214635016000101

Chmielewski, M., & Kucker, S.C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, *11*(4), 464–473. (Publisher: SAGE Publications Sage CA: Los Angeles, CA)

Cooper, D.J., & Fang, H. (2008). Understanding overbidding in second price auctions: An experimental study. *The Economic Journal*, *118*(532), 1572–1595. (Publisher: Oxford University Press Oxford, UK)

Difallah, D., Filatova, E., Ipeirotis, P. (2018). Demographics and Dynamics of Mechanical Turk Workers. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 135–143). New York, NY, USA: Association for Computing Machinery. Retrieved 2022-01-17, from https://doi.org/10.1145/3159652.3159661 10.1145/3159652.3159661

Engel, C. (2011). Dictator games: A meta study. *Experimental economics*, *14*(4), 583–610. (Publisher: Springer)

Fehr, E., & Gächter, S. (2000). Cooperation and Punishment in Public Goods Experiments. *American Economic Review*, *90*(4), 980–994.

Fischbacher, U., Gächter, S., Quercia, S. (2012). The behavioral validity of the strategy method in public good experiments. *Journal of Economic Psychology*, *33*(4), 897–913. (Publisher: Elsevier)

Gächter, S., & Herrmann, B. (2011). The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural Russia. *European Economic Review*, *55*(2), 193–210. (Publisher: Elsevier)

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., ... Puranen, B. (2020). World values survey: round seven–country-pooled datafile. *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat*.

Henrich, J., Heine, S.J., Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and brain sciences*, *33*(2-3), 61–83. (Publisher: Cambridge University Press)

Herrmann, B., Thöni, C., Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362–1367. Retrieved 2016-12-09, from http://science.sciencemag.org/content/319/5868/1362.short

Hugh-Jones, D. (2016). Honesty, beliefs about honesty, and economic growth in 15 countries. *Journal of Economic Behavior & Organization*, *127*, 99–114. (Publisher: Elsevier)

Isaac, R.M., & Walker, J.M. (1988, February). Group-Size Effects In Public-Goods Provision - The Voluntary Contributions Mechanism. *Quarterly Journal Of Economics*, *103*(1), 179–199.

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P.D., Jewell, R., Winter, N.J.G. (2020, October). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods*, *8*(4), 614–629. Retrieved 2022-01-17, from https://www.cambridge.org/core/journals/political-science-research-and-methods/article/shape-of-and-solutions-to-the-mturk-quality-crisis/521AEEB9A9753D5C6038440BD123826C

(Publisher: Cambridge University Press)

10.1017/psrm.2020.6

Krupka, E.L., & Weber, R.A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, *11*(3), 495–524. (Publisher: Oxford University Press)

Lee, Y.S., Seo, Y.W., Siemsen, E. (2018). Running Behavioral Operations Experiments Using Amazon's Mechanical Turk. *Production and Operations Management*, *27*(5), 973–989. Retrieved 2022-01-17, from https://onlinelibrary.wiley.com/doi/abs/10.1111/poms.12841 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/poms.12841)

10.1111/poms.12841

Levada (2021, July). *Koronavirus, vakcinacija, QR-kody.* Retrieved 2022-01-13, from https://www.levada.ru/2021/12/07/koronavirus-vaktsinatsiya-qr-kody/

Necka, E.A., Cacioppo, S., Norman, G.J., Cacioppo, J.T. (2016). Measuring the Prevalence of Problematic Respondent Behaviors among MTurk, Campus, and Community Participants. *PLOS ONE*, *11*(6), e0157732. (Publisher: Public Library of Science)

10.1371/journal.pone.0157732

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The quarterly journal of economics*, *122*(3), 1067–1101. (Publisher: MIT Press)

Palan, S., & Schitter, C. (2018). Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. (Publisher: Elsevier)

Peer, E., Brandimarte, L., Samat, S., Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. (Publisher: Elsevier)

Snowberg, E., & Yariv, L. (2021, February). Testing the Waters: Behavior across Participant Pools. *American Economic Review*, *111*(2), 687–719. Retrieved 2022-01-17, from

https://www.aeaweb.org/articles?id=10.1257/aer.20181065

10.1257/aer.20181065

Stewart, N., Ungemach, C., Harris, A.J., Bartels, D.M., Newell, B.R., Paolacci, G., Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision making*, *10*(5), 479–491. (Publisher: SOC JUDGMENT & DECISION MAKING)

Wang, Z., Xu, B., Zhou, H.-J. (2014). Social cycling and conditional responses in the Rock-Paper-Scissors game. *Scientific Reports*, *4*(1), 5830. (Number: 1 Publisher: Nature Publishing Group)

10.1038/srep05830

Zelmer, J. (2003). Linear public goods experiments: A meta-analysis. *Experimental Economics*, *6*(3), 299–310.