Mission Possible: Data Quality in Online Surveys

Can Çelebi, Christine Exley, Sören Harrs,
Hannu Kivimaki, Marta Serra-Garcia, Jeffrey Yusof[*]

October 28, 2025

Abstract

High-quality data is essential to social science research. Online experiments and surveys are a central tool for data collection across many disciplines, but their data quality could be lacking, due to the presence of automated agents, participants who use LLMs without researcher knowledge, and inattentive participants. We identify behavioral patterns that can serve as data quality checks by collecting data from human subjects in the lab, automated agents, and online survey platforms. We further propose the *two-stage recruitment method* by which researchers first implement a short survey on their target sample and use checks to exclude plausibly low-quality responses. We test the method with a set of checks and demonstrate how data quality can improve with this method.

JEL Classification: C81, C83, C90, 033.

Keywords: Experiments, data quality, bots, AI

[*] Celebi: can.celebi@univie.ac.at, University of Vienna; Exley: exley@umich.edu, University of Michigan; Harrs: soeren.harrs@univie.ac.at, University of Vienna; Kivimaki: hkivimak@umich.edu, University of Michigan. Serra-Garcia: mserragarcia@ucsd.edu, University of California San Diego; Yusof: jeffrey.yusof@ivr.uni-stuttgart.de, University of Stuttgart.

1 Introduction

High-quality data is essential for the implementation of evidence-based policy and the development of empirically-valid theoretical frameworks. Increasingly, such data is collected through online platforms, where market researchers, pollsters, and social scientists use surveys to elicit preferences and choices across a wide range of domains. Yet, in online settings the responses collected can be subject to concerns. First, bots—particularly AI agents[1]—may emulate human participants. Second, human respondents may use large language models (LLMs) as assistants without researchers knowing about it. Third, participants may be inattentive. Lastly, participants may engage in account fraud, for instance by using multiple accounts to participate more than once.

Despite growing concerns, systematic evidence on how to secure high quality data online—in a way that also explicitly considers the threat of AI agents and LLM assistance—remains limited. Previously common attention checks may no longer be useful, and little is known about the prevalence of survey responses that were completed by AI agents or with the aid of LLMs. In addition, as AI agents continue to improve, developing methods to detect and mitigate their influence on survey data is an urgent challenge.

In this paper, we directly compare the behavior of AI agents, human participants in the laboratory, and respondents recruited online. We first identify behavioral patterns that can serve as checks for low quality data, including the identification of AI agents. Building on these insights, we then propose a two-stage recruitment method in which we first run our "baseline survey", and then only allow participants with high-quality submissions from the screening study to participate in our "main study". This recruitment method can be used to enhance data quality in a scalable and adaptive manner.

The baseline survey includes a dictator game, an open-ended text question, standard sociodemographic questions, and two types of attention checks (classic textual questions and a novel video-based question). We collect data from three groups: 314 human subjects in a controlled laboratory setting, 200 responses from AI agents (GPT Agent mode, Perplexity Comet, and Browser Use), and approximately 2,400 online participants recruited on MTurk and Prolific Academic. We compare four types of measures within the baseline survey. First, we check the performance on both the classic and video attention check questions. Second, we examine interaction traces, such as copy–paste–blur–tab events, keystrokes, and mouse movements. Third, we analyze the response data, including the open-text answer, the dictator game decision, and reported demographics. Finally, we incorporate external validation metrics from sources including reCAPTCHA scores, device fingerprints, and Pangram AI likelihood scores.

[1] We specifically refer to AI agents that are powered by large multimodal models (LMMs) such as GPT Agent Mode and Perplexity Comet.

The baseline survey data reveal several measures along which human participants in the lab, AI agents and online respondents differ. Nearly all AI agents (98%) pass classic text-based attention checks (e.g., a multiple-part question that only asks the respondent to check the box on the left-hand side of the screen). But, only 19% pass our novel video attention check. AI agents also frequently paste their responses in open-ended text responses or type their responses at speeds that are atypically fast.

Human participants in the laboratory perform well on both types of attention check questions; 84% pass the classic attention check questions, and 97% pass the novel video-based attention check question. For the open-ended text response, all lab participants typed their answer, and almost none typed at speeds that approached those of most AI agents.

Prolific participants perform, if anything, better than human participants in the lab on the attention checks; 98% pass the classic attention check questions, and 97% pass the novel video-based attention check question. For the open-ended text response, nearly all Prolific participants typed their responses in open-ended text responses and had typical typing speeds for humans.

MTurk participants are more nuanced. While 98% pass the novel video-based attention check (perhaps due to the high salience of videos), only 56% pass attention checks that require reading, showing clear inattention. In addition, the vast majority of MTurk participants did not type their responses into the open-ended text or had atypical typing speeds for humans, suggesting they are using AI assistance.

Combining these findings from the baseline survey, we propose and implement a set of checks to identify high-attention human participants. These checks require the respondent to successfully complete the text- and video-based attention check questions, and to satisfy two requirements with regard to their open text response (i.e., no pasting, typing with typical speed). Among participants in the lab, a large majority (80%) satisfy all data quality checks. Data quality is even better on Prolific, where 89% of Prolific respondents from the baseline survey satisfy these data quality checks. By contrast, only 9% of MTurk respondents from the baseline survey satisfy them, indicating very poor data quality. Additionally, no bots (0%) satisfy the checks, providing a strong exclusion restriction for AI agents.

Next, we implement the two-stage recruitment method, using these data-driven quality checks. Specifically, for our “main study," we recruit respondents on Prolific—restricting to the set of Prolific participants who passed the checks in the baseline survey. While our main study simply asks participants to again complete the same set of questions as those included in the baseline survey, the main study could be—as suggested by the name—the study that elicits the main outcomes of interest for the researchers in future work. In our main study, we find that 93% of Prolific respondents now satisfy all checks, reaching the highest data quality levels, which directionally exceed those of the lab.

The two-stage recruitment method provides a helpful tool for researchers using online platforms to collect data, for a variety of platforms. Online platforms allow researchers to restrict who is eligible to participate in their studies based off of prior study participation. Thus, researchers can use a baseline survey to find high-quality respondents and then only recruit from the high-quality respondents for their main study. One can consider this approach as akin to the approach that is commonly used for laboratory studies: develop and maintain a subject pool and then recruit from that subject pool for one's study. For researchers or teams of researchers who conduct many studies online, this pool of high quality respondents can be much larger than needed for one specific study and instead serve as the study recruitment pool for many studies. In this way, the expansion of AI agents poses challenges, but—just as in-person laboratory studies recruit and maintain high-quality subjects—the two stage recruitment method can allow researchers to recruit and maintain high-quality subjects.

The two-stage recruitment method is intently simple, such that it can be used by researchers across disciplines, and address concerns of data quality that could lead to calls to conduct experiments only in laboratory (in-person) settings. This is important because a wide range of tools are needed to answer important questions in social science, including both laboratory and online data.

In addition, we emphasize that carefully considering data quality is *always* important for researchers, including researchers who collect from online platforms, laboratory settings, and representative samples.[2] Data quality issues can exist within all of these settings; quality and attention checks are useful for all studies. As with other important topics in research, our concerns about data quality should be, we believe, driven by empirical evidence, and *not* our own impressions, which may or may not be accurate. Some data sources may prove more problematic than one's prior, while other data sources may prove more promising than one's prior.

This paper provides evidence on the current data quality of Prolific, a commonly used online platform. While our project was originally motivated by a much-discussed drop in data quality on Prolific during the earlier part of 2025, it is reassuring that this data quality issue did not appear when we turned to design and run this study in the Summer and Fall of 2025. This also speaks to the need for researchers who collect data to, always, collect and pay careful attention to trends in data quality measures.

Our work contributes to a body of literature examining data quality across a variety of fields (e.g., **?**; **?**; **?**; **?**; **?**; **?**; **?**; **?**;**?**; **?**). As technology such as AI improves, there might be two parallel developments: AI agents may become better at mimicking humans, and at the same time, AI-detection software may improve its detection of these agents (e.g., **?**, **?**). That is, while specific attention checks may not be valid for longer periods of time, improved checks may appear over

[2] Data quality is of course also an important topic for observational and administrative data, as often well-understood by various subfields.

time. For example, one of the most common checks pertains to reCAPTCHA scores, which are an adaptive measure coded by Google in response to how an individual or bot interacts with a website.[3] Researchers need not be able to "out smart" AI; they only need to rely on technology that can do so. This is why—with this paper—we propose the two-stage recruitment method, which can be adjusted to new checks over time, as bots and tools to detect them necessarily evolve.

2 Experimental Design

In this section, we first describe the survey instrument used throughout all samples.[4] We then describe the implementation of the experiment in each sample.

2.1 Survey Instrument

The two-stage recruitment method consists of two surveys: the baseline survey and the main survey. The baseline survey is a brief survey that recruits from a broad sample of participants and is used to identify high-quality respondents, by collecting several measures of data quality. The main survey collects the main outcomes of interest for the researcher but only recruits from high-quality respondents who have been previously identified in the baseline survey. Ensuring a sufficient number of high-quality respondents complete the baseline survey is thus a necessary prerequisite to conducting the main survey, which may or may not be possible with all subject populations.

To examine the feasibility of the two-stage recruitment method via common online platform, we largely focus on the data quality measures in both the baseline survey and main survey. Thus, while the main survey would likely differ from the baseline survey for many researchers, this paper uses, for simplicity, the same survey instrument as the baseline survey and the main survey. For simplicity, we also refer to each submission as being provided by a "respondent," which implies that individuals or automated agents may complete the survey.

The baseline and main surveys start with a consent form and an information page describing that the study consists of one decision and a series of survey questions. One out of every 100 respondents is randomly selected for their decision to be implemented. Prior to beginning, they are asked to answer one understanding question before proceeding with the survey, and they complete a Captcha verification question.

The baseline and main surveys then proceeds as follows: respondents complete one main economic decision, one open-ended survey question that asks about their decision, and a final questionnaire that includes questions about their socio-demographic characteristics and attention checks.

The economic decision that respondents make is a standard dictator game, in which they decide

[3] For more details, see Appendix B.1.

[4] See Appendix F for screenshots of the survey instructions.

how to split $10 with another respondent in the study.

The open-ended survey question asks respondents to describe how they made their decisions. Specifically, respondents are asked the following:

> *Please consider both the decisions that YOU made and the decisions that OTHER PARTICIPANTS may have made in this study.*
>
> - *How would you describe your decisions?*
> - *What factors and considerations influenced your decisions?*
> - *How do you think other participants might have approached these decisions?*
> - *Are there any reasons why others might have made different choices?*
>
> *Please write at least 3-4 full sentences.*

The final questionnaire starts by asking respondents to indicate their agreement (on a 5-point scale, ranging from "Strongly Disagree" to "Strongly Agree") with four statements. The first two statements are "I made each decision in this study carefully" and "I understood how my decisions would affect my earnings in this study." The second two statements serve as *classic attention checks* in which participants are asked to "Select the button that is furthest to the right" and " to the left", respectively.

The respondents then indicate their race or ethnicity, their gender identity, the US state in which they are located, the type of community they currently live in (urban, suburban, or rural), their political identity (Republican or Democrat), and their age.

They end the study by completing a video attention check, which consists of a short animation in which four numbers appear sequentially and respondents are asked to write them in a dedicated answer box.[5]

2.2 Data Quality Measures

We have two data quality checks that are immediately implied by the answers respondents provide to the questions on the survey instrument:

- *Passed classic checks* is an indicator variable equal to 1 if respondents correctly select the option furthest to the "left" and "right" when asked to do so in the 5-point scale question in the follow-up questionnaire.

[5] The video attention question leverages a technical limitation of AI agents. AI agents sample visual information infrequently via repeated screenshots. As a result, they may either not see all numbers shown in our video or the correct sequence of numbers.

- *Passed video check* is an indicator variable equal to 1 if respondents correctly type in the four numbers when asked to do so after viewing a video that displays these four numbers.

We then track measures of typing based on responses to the open-ended survey question:[6]

- *Typed text* is an indicator variable for whether the respondent typed text in this open-ended question. It takes value 1 if (i) respondents do not trigger a paste event, (ii) there is no large increase in the amount of text without corresponding keystrokes (no input jump event of more than 50 characters), and (iii) at least one keystroke is observed.
- *Typed with typical speed* is an indicator variable equal to 1 if the respondent typed the text and the median typing speed is slower than 75 milliseconds per keystroke.[7]

On the questionnaire page, we track mouse movements and clicks:

- *Mouse clicks >0* is an indicator variable that takes value 1 if the respondent clicked on the screen at least once.[8]
- *Mouse movements >0* is an indicator variable that takes value 1 if the respondent moved the mouse at least one time within the questionnaire page.[9]

We use reCAPTCHA scores by Google and AI likelihood scores by Pangram to identify potential AI agents and AI assistance using external measures:

- *ReCAPTCHA score $=1$ or ≥ 0.9* is an indicator variable equal to 1 if Google's reCAPTCHA score is equal to 1 or higher than 0.9, respectively. The reCAPTCHA algorithm evaluates a range of behavioral and contextual signals during a user's interaction with the survey interface and assigns each response a value between 0 and 1, with lower scores indicating a higher likelihood that the submission originated from a bot.[10]
- *Pangram AI likelihood <1 or <0.5* is an indicator that takes value 1 if the Pangram AI likelihood score is less than 1 or 0.5, respectively. The Pangram score indicates the probability that an LLM was used to generate the response to the open-ended question.[11]

[6] We present representative keylog data in Appendix E.2 to illustrate how AI agents and humans exhibit distinct typing behavior.

[7] As shown in Appendix Figure A.1, which plots the distribution of typing speeds for the different samples, lab participants almost always type slower than 75 milliseconds while bots frequently type faster than 75 milliseconds.

[8] Appendix Figure A.2 shows the distribution of mouse click counts for the different samples.

[9] Appendix Figure A.3 shows the distribution of mouse movements for the different samples.

[10] See Appendix B.1 for a more detailed discussion of reCAPTCHA scores.

[11] See Appendix B.2 for a more detailed discussion of Pangram scores. Appendix Figure A.5 further shows the distribution of Pangram scores for the different samples.

To identify potential fraudulent accounts, we use IP addresses, geolocations, cookies, and device fingerprints. Note that for our bot and lab samples, these measures are recoded as missing, as shared IP addresses and device fingerprints are not indicative of account fraud in these samples.[12]

- *Unique IP address* is an indicator variable equal to 1 if the IP address is unique within each sample (for Prolific and MTurk respondents).
- *US IP address* is an indicator variable equal to 1 if the IP address is not located outside the United States (provided by Fingerprint.com).
- *Not in a geolocation cluster* is an indicator variable that takes value 1 if fewer than five responses have a geographic latitude–longitude combination (provided by Qualtrics) that is the same.
- *Unique Qualtrics submission* is an indicator variable that takes value 1 if it is not classified as a duplicate submission by Qualtrics. To prevent respondents from making repeated submissions from the same browser and device, Qualtrics flags submissions as duplicates using cookies.
- *Unique device fingerprint* is an indicator variable that takes value 1 if the submission has a unique device fingerprint (provided by Fingerprint.com). Device fingerprinting assigns a unique and persistent identifier to any device (visitor ID). This device fingerprint allows to track devices even when IPs change (e.g. through VPNs) or cookies are deleted.[13]

2.3 Samples

We conducted the experiment in the lab, with undergraduate students, on two widely used platforms, Prolific and MTurk, and with automated agents (or bots). We pre-registered the lab and Prolific data collection under pre-registration #242081, and added the data collection on MTurk under pre-registration #247939 on AsPredicted.org.

2.3.1 Lab Sample

We recruited students at UC San Diego to participate in a 5-minute study. The survey is completed in-person, and thus provides a benchmark of how a human would complete the survey. Students complete the survey in a room with 40 computer terminals that provide privacy, despite the presence of research assistants at the front of the room. The use of an LLM to provide responses is not expected and not allowed. During the Summer and Fall 2025 quarters, 314 students came to the lab as part of their class credit, which implied that there was no fixed fee for participation but the incentive-compatible dictator game decision.

[12] AI agents would in general not share the same IP addresses or device fingerprints, if they are run on different devices and networks. Thus, these two measures should not be viewed as a reliable way of identifying AI agents.

[13] See Appendix B.3 for a more detailed discussion of device fingerprinting scores.

2.3.2 AI Agent (Bot) Sample

AI agents are a recent development in LLM-assisted tools that can interact with external data sources and autonomously perform specific tasks, such as filling out an online survey. Since different AI agents are built on different LLMs and may exhibit varying behaviors, we collected data from several state-of-the-art AI agents available at the time of the study: ChatGPT Agent, Perplexity Comet, and BrowserUse. ChatGPT Agent and Perplexity Comet are commercial products, while BrowserUse is an open-source web automation framework that allows various LLMs to operate as agents. For BrowserUse, we implemented agents using OpenAI's O3 and Google's Gemini 2.5 Flash.[14]

We instructed the AI agents to complete the survey under two prompt conditions. The *Simple* prompt provided only the survey link and directed the agent to act as a human participant. The *Complex* prompt added instructions on persona and answer behavior to imitate human responses. This design allows us to test the sensitivity of our checks to prompt complexity.[15] We restrict the sample to attempts in which the agents successfully completed the survey. For the ReCAPTCHA verification question and the video attention check, we recorded each instance in which the agent requested human assistance by the researcher. In total, the AI agent sample consists of 200 observations.[16] Comparing this sample with the lab sample provides an empirical benchmark to evaluate which measures effectively distinguish human from AI-generated responses.

2.3.3 Prolific Sample

We recruited 1200 participants on Prolific Academic. Specifically, we recruit 300 participants for each of our four treatments run on Prolific: the Prolific (All) treatment, the Prolific (T95) treatment, the Prolific (T99) treatment, and the Prolific (two-stage) treatment.

In the Prolific (All) treatment, the Prolific (T95) treatment, and the Prolific (T99) treatment, participants complete the baseline survey, and this survey does not vary across treatments. All that varies are the eligibility screeners applied to participants, according to their approval rating and prior experience on Prolific, to test whether higher approval rating and high prior experience result in respondents with responses and traces consistent with higher data quality. In the first treatment (*All*), Prolific participants based in the US, with any approval rating and any prior experience, complete the stage-one survey. In *T95*, we introduce the commonly used requirements for participants to have at least a 95% approval rating in prior submissions and more than 100 prior submissions. In *T99*, we introduce more stringent requirements, requiring that participants have at least a 99% approval rating and more than 1000 prior submissions.

[14] See Appendix C.4 for more technical information about the AI agents.

[15] See Appendix C.2 for a more detailed discussion of the used prompts.

[16] See Appendix C.1 for details on the implementation of the AI agent data collection and Appendix C.3 for additional analyses of AI agent survey completions.

For the two-stage treatment (2S), we use data from these participants and all other samples to determine our main set of quality checks that would exclude bots in the bot sample and include a large fraction of human participants based on their lab behavior.[17] These main quality checks require participants to have correctly answered the three attention checks (*Passed classic checks* and *Passed video check*), to have typed text at typical speeds (*Typed text* and *Typical typing speed*) and to have a *unique IP address*.

Then, we only invite participants who pass all of our main quality checks to complete the main survey in our *two-stage* treatment.

2.3.4 MTurk Sample

We recruited 900 participants on MTurk. Specifically, we recruited 300 participants for each of the following three treatments: the MTurk (All) treatment, the MTurk (T95) treatment, and the MTurk (T99) treatment. These treatments mirror the corresponding three Prolific treatments; for example, the *T95* treatments restricts to MTurk participants with have a 95% or greater approval rating on their previous HITs and more than 100 HITs approved.

Because the data quality was extremely low on MTurk (9% of respondents passed the quality checks that would make them eligible for the two-stage treatment), only 79 of 881 respondents would have been available to conduct the two-stage recruitment treatment, as planned. Given this limited sample size and severe concerns about data quality, we did not conduct the two-stage treatment, as planned.

3 Results

We start by describing our main data quality checks. Then, we provide detailed results for each of the data quality measures collected for each sample and treatment. Finally, we describe the economic decisions and sociodemographic characteristics of each sample.

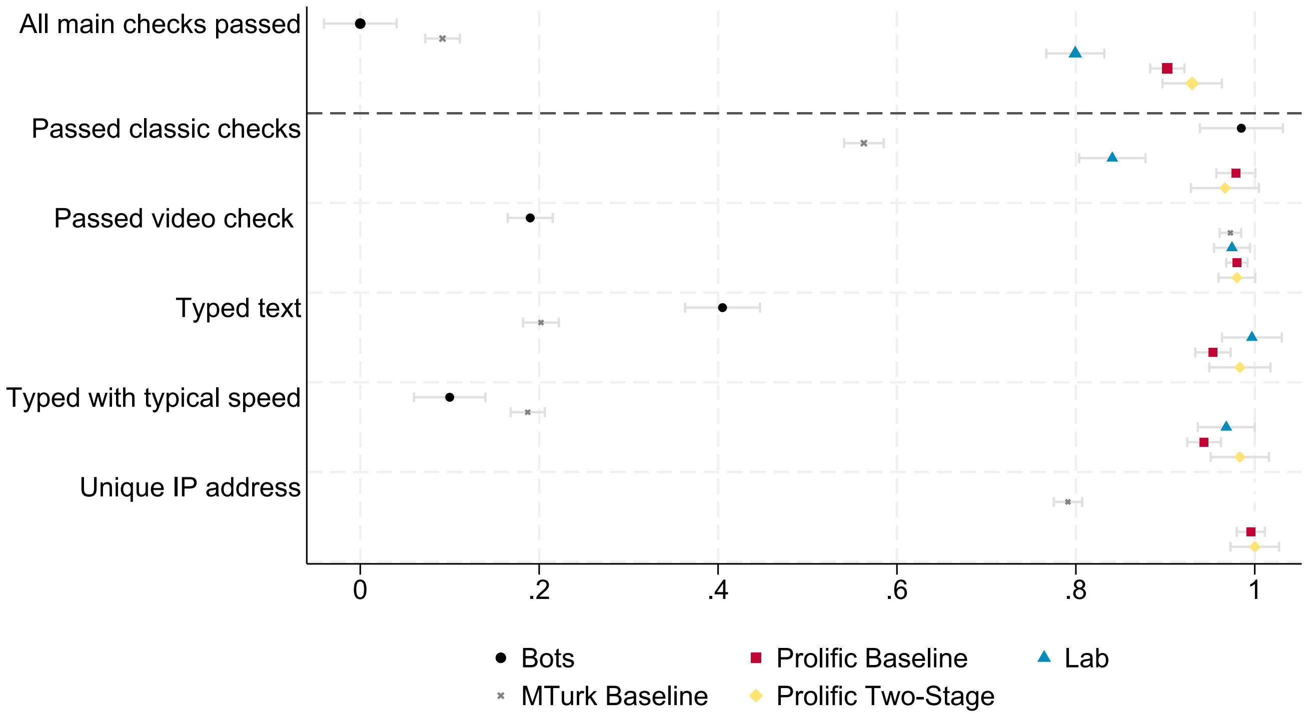
3.1 Main Data Quality Checks

Figure 1 shows the fraction of respondents that passed the main data quality checks simultaneously and individually, by sample. Two samples exhibit low data quality: 0% of bots and 9% of MTurk respondents pass all 5 checks. MTurk respondents perform better in these checks than the bot data, but substantially worse than lab and Prolific participants.

There are large data quality improvements when we turn to Prolific respondents, who pass the 5 checks 89% of the time. There is even slightly better data quality when we turn to Prolific data

[17] It included 87% of respondents in the lab, based on 60 subjects who participated in the summer quarter, and 80% based on all 314 included through the Summer and Fall quarters.

Figure 1: Data Quality - Main Checks



Notes: This figure shows the average average rate with which each sample passed the data quality checks. Whisked bars indicate 95% confidence intervals. See Section 2.2 for detailed definitions.

that is collected via our two-stage recruitment method—93% of respondents pass all 5 data quality checks.

Can data quality be improved by turning to the lab? When considering whether all checks are passed, the Prolific data performs—if anything—slightly better than the lab data. While 89-93% of Prolific participants pass all 5 quality checks, only 80% of lab participants pass all 5 quality checks. Prolific respondents are particularly better than lab participants when considering whether responses passed the classic attention checks. Lab participants only passed the classic attention checks 84% of the time, which likely captures lab participants who are inattentive when reading in the lab.

In addition to the attention checks, bots and respondents in the MTurk sample differ from lab and Prolific respondents in how they answer the open-ended survey question. Bots and MTurk respondents are less likely to have typed into the text box (in 80% of the cases for MTurk and 58% of the cases for bots). In addition, most bots don't type the text at speeds typical for humans.

Some account fraud may also be happening on MTurk, where we observe that 79% of respondents submit their surveys from a unique IP address, while 21% submit their surveys from a non-unique IP addresses. These repeat submissions could stem from users that submit the same HIT multiple times, under different accounts.

Table 1 shows the frequency with which each data quality check is passed, separating the groups in Prolific and MTurk according to the their screeners that are determined by the treatment in which they were recruited (All, T95, and T99).[18] The data reveal that similar results hold across the three screeners used on MTurk as well as the three screeners used on Prolific. Hence, prior experience and approval ratings do not seem to substantially interact with the data quality measures we collect.

Table 1: Main Data Quality Checks

	Bot	**MTurk**			**Prolific**				**Lab**
		All	T95	T99	All	T95	T99	2S	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
All main checks passed	0.00	0.09	0.11	0.07	0.89	0.91	0.91	0.93	0.80
Passed classic checks	0.98	0.66	0.52	0.52	0.98	0.98	0.98	0.97	0.84
Passed video check	0.19	0.98	0.96	0.98	0.99	0.98	0.98	0.98	0.97
Typed text	0.41	0.23	0.25	0.13	0.94	0.96	0.96	0.98	1.00
Typed with typical speed	0.10	0.22	0.23	0.11	0.93	0.96	0.94	0.98	0.97
Unique IP address	.	0.74	0.83	0.80	0.99	1.00	1.00	1.00	.
Observations	200	287	299	295	300	300	300	300	314

This table presents the fraction of participants passing each data quality check from each response type in this paper. Column (1) presents Bot respondents. Columns (2) - (4) present MTurk respondents recruited via the All treatment, T95 treatment, and T99 treatment, respectively. Columns (5) - (8) present Prolific respondents recruited via the All treatment, T95 treatment, T99 treatment, and the two-stage 2S treatment, respectively. Column (9) presents lab respondents. *All main checks checks passed* shows the fraction passing all data quality checks; The remaining variables show the fraction passing the noted check (see Section 2.2 for specific definitions.)

3.2 Robustness Measures of Data Quality

Table 2 shows additional data quality measures.[19] When differences emerge across the data sources, these data quality measures show similar patterns to the main quality checks analyzed so far.

We obtain three main insights. First, using these additional measures does not improve the detection of bots, which were already detected 100% of the time with the main quality checks. Second, these additional measures are unlikely, relative to the main quality checks that pertain to the classic and video checks, to capture inattention.

Third, we find no evidence that these additional measures meaningfully improve data quality within our two-stage recruitment framework. As shown in Column (8), 100% of the Prolific participants recruited via our two-stage recruitment method pass our additional checks for almost all additional checks. In addition, close to 100% of these Prolific participants the remaining additional checks.

[18] Appendix Table A.2 shows the results separately for each AI agent model and prompt type.

[19] Appendix Table A.3 shows the results separately for each AI agent model and prompt type.

98% of these Prolific participants have a Pangram AI likelihood that suggests LLM was unlikely to generate the response to the open-ended question, and 98% also have a reCAPTCHA score that is 0.90 or higher. Even 93% of these Prolific participants have the maximum reCAPTCHA score of 1—a threshold that is only met by 69% of human participants in the lab and hence seems likely to inaccurately flag some (clearly) human behavior.

Table 2: Additional Data Quality Checks

	Bot	**MTurk**			**Prolific**				**Lab**
		All	T95	T99	All	T95	T99	2S	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
All main checks passed	0.00	0.09	0.11	0.07	0.89	0.91	0.91	0.93	0.80
Mouse clicks > 0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Mouse movements > 0	0.85	0.99	0.91	0.83	1.00	0.99	1.00	1.00	1.00
ReCAPTCHA challenge: passed	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ReCAPTCHA score $\geq$ 0.9	0.60	0.54	0.65	0.76	0.91	0.92	0.97	0.98	0.93
ReCAPTCHA score == 1	0.23	0.37	0.50	0.61	0.80	0.81	0.87	0.93	0.69
Pangram AI likelihood < 0.5	0.07	0.29	0.30	0.25	0.97	0.98	0.96	0.98	1.00
Pangram AI likelihood < 1	0.19	0.42	0.40	0.34	0.98	0.99	0.97	0.99	1.00
US IP address	.	0.93	0.96	0.98	1.00	1.00	1.00	1.00	.
Not in a geolocation cluster	.	0.70	0.81	0.76	1.00	1.00	1.00	1.00	.
Unique Qualtrics submission	.	0.91	0.93	0.94	1.00	1.00	1.00	1.00	.
Unique device fingerprint	.	0.62	0.84	0.55	0.99	1.00	0.99	1.00	.
Observations	200	287	299	295	300	300	300	300	314

This table presents the fraction of participants passing each data quality check from each response type in this paper. Column (1) presents Bot respondents. Columns (2) - (4) present MTurk respondents recruited via the All treatment, T95 treatment, and T99 treatment, respectively. Columns (5) - (8) present Prolific respondents recruited via the All treatment, T95 treatment, T99 treatment, and the two-stage 2S treatment, respectively. Column (9) presents lab respondents. *All main checks passed* shows the fraction pass all the main checks (shown in Table 1). The remaining variables show the fraction passing the additional noted checks (see Section 2.2 for specific definitions.)

3.3 Donation and Demographics

We also explore how respondents, across all samples, made decisions in the dictator game and their demographic information (see Appendix Table A.1).[20]

While several differences emerge, the only clearly expected difference is that lab participants are much younger, which corresponds with them being undergraduate students.

[20] Appendix Figure A.6 shows the distribution of the dictator game decisions for the different samples and Appendix Table A.2 shows the demographic information separately for each AI agent model and prompt type.

4 Conclusion

This paper develops and tests several checks of data quality to mitigate concerns about the quality of responses obtained in online experiments, given the rapid expansion of AI assistants and AI agents. Although AI is constantly evolving, it is important to develop methodologies that ensure high data quality over time. For that reason, we propose a simple method, the two-stage recruitment method, which can help researchers ensure their respondents provide high quality data that can be relied on to answer their research questions.

The main data quality checks we use rely on a combination of several simple measures based on: (1) attention check questions, which can be added to any survey, (2) typing patterns, which can be tracked in any open-ended text question, and (3) IP address checks, which are automatically collected with the default settings in survey platforms like Qualtrics. The attention checks serve to detect participants who do not pay close attention to questions and bots that cannot (yet) assess video content. The typing patterns serve to detect text stemming from other sources, such as AI assistants used outside of the survey context, or text typed at speeds that are unlikely to be humans. Finally, IP address checks are helpful in detecting potentially fraudulent submissions on online platforms.

When considering these data quality checks, we observe that Prolific respondents are of very high data quality, even slightly higher than participants in the lab. By contrast, no bots pass the data quality checks and only a small fraction of MTurk responses do. We then apply the two-stage recruitment method, and only invite Prolific participants who pass all data quality checks to a second survey. The data reveal that data quality measures in this selected sample are even higher, though only directionally so because of the high levels of data quality of Prolific to begin with.

We hope this paper serves other researchers in evaluating the data quality of experiments they conduct online and provides a methodology to ensure that data quality remains high, as AI evolves in the ways it can assist (or impersonate) humans.

Online Appendix (For Online Publication Only)

Mission Possible: Data Quality in Online Survey

October 28, 2025

A Additional Results

Table A.1: Demographics Data

		MTurk			Prolific				
	Bot	All	T95	T99	All	T95	T99	Two Stage	Lab
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Give $0	0.04	0.04	0.03	0.03	0.15	0.18	0.18	0.18	0.07
Give $1-4	0.42	0.20	0.15	0.26	0.26	0.25	0.24	0.26	0.12
Give $5	0.54	0.41	0.40	0.49	0.56	0.55	0.56	0.54	0.71
Give $6-10	0.00	0.36	0.42	0.22	0.03	0.02	0.02	0.01	0.10
Age: 18-25	0.20	0.14	0.08	0.07	0.08	0.08	0.06	0.06	0.97
Age: 26-45	0.46	0.81	0.87	0.90	0.59	0.50	0.55	0.54	0.03
Age: 46-65	0.34	0.05	0.05	0.02	0.30	0.38	0.32	0.34	0.00
Age: 65+	0.00	0.00	0.00	0.01	0.03	0.04	0.07	0.07	0.00
Identifies as a man	0.34	0.72	0.70	0.56	0.48	0.49	0.48	0.49	0.34
Identifies as a woman	0.44	0.29	0.30	0.47	0.49	0.48	0.50	0.47	0.63
Identifies as gender diverse	0.03	0.00	0.01	0.04	0.03	0.03	0.04	0.04	0.06
Gender: prefer not to say	0.20	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01
Identifies as White	0.70	0.86	0.89	0.93	0.79	0.74	0.76	0.75	0.18
Identifies as Black	0.01	0.02	0.01	0.06	0.08	0.12	0.10	0.09	0.04
Identifies as Hispanic	0.01	0.02	0.00	0.07	0.08	0.08	0.08	0.09	0.17
Identifies as Asian	0.10	0.05	0.02	0.12	0.09	0.10	0.11	0.10	0.64
Identifies as Native Hawaiian	0.00	0.00	0.00	0.06	0.00	0.00	0.01	0.00	0.01
Identifies as Middle-eastern	0.01	0.01	0.00	0.07	0.01	0.00	0.01	0.02	0.04
Identifies as Native American	0.01	0.09	0.08	0.09	0.01	0.00	0.01	0.00	0.00
Race: prefer not to say	0.17	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01
Republican	0.36	0.44	0.49	0.43	0.37	0.34	0.32	0.34	0.11
Democrat	0.16	0.55	0.51	0.53	0.52	0.57	0.55	0.55	0.46
Political affiliation: prefer not to say	0.47	0.01	0.00	0.04	0.11	0.10	0.13	0.12	0.43
Rural	0.07	0.10	0.09	0.04	0.17	0.22	0.18	0.19	0.03
Suburban	0.24	0.19	0.35	0.49	0.58	0.52	0.53	0.57	0.48
Urban	0.69	0.71	0.56	0.47	0.26	0.26	0.29	0.24	0.49
Observations	200	287	299	295	300	300	300	300	314

This table presents information on the fraction of participants with each demographic characteristic. Column (1) presents Bot respondents. Columns (2) - (4) present MTurk respondents recruited via the All treatment, T95 treatment, and T99 treatment, respectively. Columns (5) - (8) present Prolific respondents recruited via the All treatment, T95 treatment, T99 treatment, and the two-stage treatment, respectively. Column (9) presents lab respondents.

Table A.2: AI Agents: Main Data Quality Checks

	Bots (1)	GPT-S (2)	GPT-C (3)	PERP-S (4)	PERP-C (5)	BU-O3-S (6)	BU-O3-C (7)	BU-G25F-S (8)	BU-G25F-C (9)
All main checks passed	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Passed classic check	0.98	1.00	0.93	1.00	1.00	1.00	1.00	1.00	1.00
Passed video check	0.19	0.03	0.03	0.90	0.85	0.00	0.00	0.00	0.05
Typed text	0.41	0.00	0.00	0.00	0.05	1.00	1.00	1.00	1.00
Typed with typical speed	0.10	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Unique IP address	.	.	.	.	.	.	.	.	.
Observations	200	40	40	20	20	20	20	20	20

This table presents results for survey responses generated by different AI agents and prompt types. Column (1) presents all Bot respondents. Columns (2)–(3) present responses from the GPT agent under the simple (-S) and complex (-C) prompt conditions, respectively. Columns (4)–(5) present responses from the Perplexity agent under the simple and complex prompt conditions. Columns (6)–(7) present responses from the BrowserUse agent (BU-O3, based on GPTO3) under the simple and complex prompt conditions. Columns (8)–(9) present responses from the BrowserUse agent (BU-G25F, based on Gemini 2.5 Flash) under the simple and complex prompt conditions.

Table A.3: AI Agents: Additional Data Quality Checks

	Bots (1)	GPT-S (2)	GPT-C (3)	PERP-S (4)	PERP-C (5)	BU-O3-S (6)	BU-O3-C (7)	BU-G25F-S (8)	BU-G25F-C (9)
All main checks passed	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mouse clicks > 0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Mouse movements > 0	0.85	1.00	1.00	0.30	0.20	1.00	1.00	1.00	1.00
ReCAPTCHA challenge: passed	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ReCAPTCHA score ≥ 0.9	0.60	0.78	0.82	0.55	0.25	0.90	0.70	0.00	0.45
ReCAPTCHA score $==$ 1	0.23	0.20	0.35	0.10	0.05	0.75	0.25	0.00	0.00
Pangram AI likelihood < 0.5	0.07	0.03	0.25	0.00	0.10	0.00	0.05	0.00	0.00
Pangram AI likelihood < 1	0.19	0.10	0.47	0.05	0.25	0.00	0.10	0.35	0.00
US IP address	.	.	.	.	.	.	.	.	.
Not in a geolocation cluster	.	.	.	.	.	.	.	.	.
Unique Qualtrics submission	.	.	.	.	.	.	.	.	.
Unique device fingerprint	.	.	.	.	.	.	.	.	.

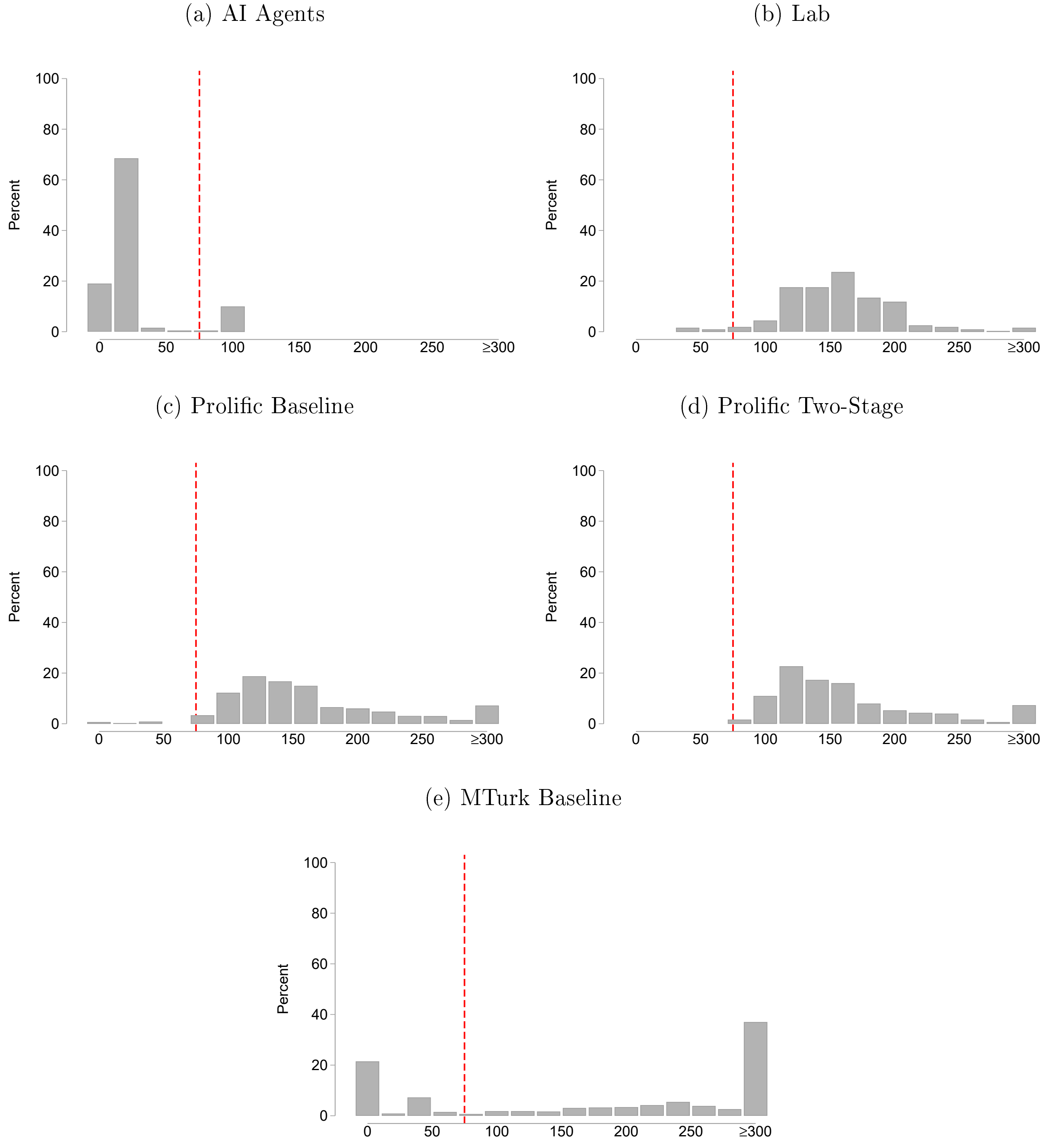
This table presents results for survey responses generated by different AI agents and prompt types. Column (1) presents all Bot respondents. Columns (2)–(3) present responses from the GPT agent under the simple (-S) and complex (-C) prompt conditions, respectively. Columns (4)–(5) present responses from the Perplexity agent under the simple and complex prompt conditions. Columns (6)–(7) present responses from the BrowserUse agent (BU-O3, based on GPTO3) under the simple and complex prompt conditions. Columns (8)–(9) present responses from the BrowserUse agent (BU-G25F, based on Gemini 2.5 Flash) under the simple and complex prompt conditions.

Table A.4: AI Agents: Demographics Data

	Bots	GPT-S	GPT-C	PERP-S	PERP-C	BU-O3-S	BU-O3-C	BU-G25F-S	BU-G25F-C
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Give $0	0.04	0.00	0.00	0.00	0.00	0.15	0.15	0.05	0.00
Give $1-4	0.42	0.00	0.97	0.35	0.75	0.60	0.40	0.15	0.05
Give $5	0.54	1.00	0.03	0.65	0.25	0.25	0.45	0.80	0.95
Give $6-10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Age: 18-25	0.20	0.00	0.00	0.15	0.00	0.35	0.40	0.70	0.40
Age: 26-45	0.46	1.00	0.00	0.85	0.15	0.65	0.20	0.30	0.40
Age: 46-65	0.34	0.00	1.00	0.00	0.85	0.00	0.40	0.00	0.20
Age: 65+	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Identifies as a man	0.34	0.62	0.00	0.35	0.10	0.55	0.20	0.40	0.55
Identifies as a woman	0.44	0.00	1.00	0.60	0.90	0.05	0.40	0.05	0.40
Identifies as gender diverse	0.03	0.00	0.03	0.00	0.00	0.00	0.00	0.20	0.05
Gender: prefer not to say	0.20	0.38	0.00	0.05	0.00	0.40	0.40	0.35	0.05
Identifies as White	0.70	0.62	1.00	0.95	1.00	0.30	0.55	0.10	0.90
Identifies as Black	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00
Identifies as Hispanic	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05
Identifies as Asian	0.10	0.03	0.03	0.00	0.00	0.35	0.05	0.55	0.00
Identifies as Native Hawaiian	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Identifies as Middle-eastern	0.01	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Identifies as Native American	0.01	0.00	0.05	0.00	0.00	0.00	0.00	0.05	0.00
Race: prefer not to say	0.17	0.35	0.00	0.05	0.00	0.35	0.40	0.15	0.10
Republican	0.36	0.00	1.00	0.00	0.90	0.00	0.40	0.15	0.20
Democrat	0.16	0.00	0.00	0.25	0.00	0.45	0.00	0.30	0.60
Political affiliation: prefer not to say	0.47	1.00	0.00	0.75	0.10	0.55	0.60	0.55	0.20
Rural	0.07	0.25	0.00	0.00	0.15	0.00	0.00	0.00	0.00
Suburban	0.24	0.00	0.28	0.60	0.65	0.25	0.05	0.10	0.25
Urban	0.69	0.75	0.72	0.40	0.20	0.75	0.95	0.90	0.75
Observations	200	40	40	20	20	20	20	20	20

This table presents information on the fraction of participants with each demographic characteristic by AI-agent type. Column (1) presents all Bot respondents. Columns (2)–(3) present responses from the GPT agent under the simple (-S) and complex (-C) prompt conditions, respectively. Columns (4)–(5) present responses from the Perplexity agent under the simple and complex prompt conditions. Columns (6)–(7) present responses from the BrowserUse agent (BU-O3, based on GPTO3) under the simple and complex prompt conditions. Columns (8)–(9) present responses from the BrowserUse agent (BU-G25F, based on Gemini 2.5 Flash) under the simple and complex prompt conditions.

Figure A.1: Typing Speed (ms)



(a) AI Agents

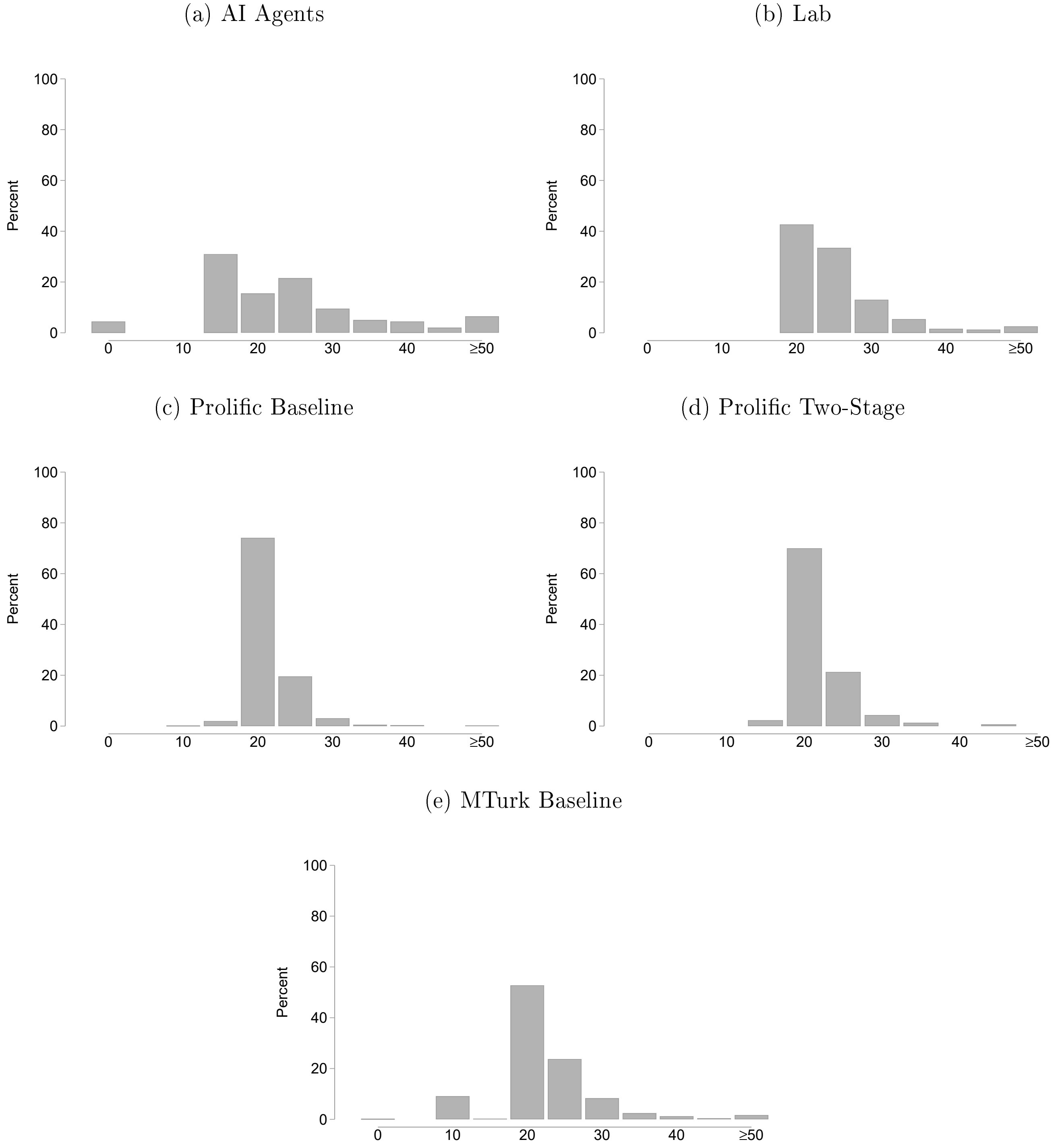
(b) Lab

(c) Prolific Baseline

(d) Prolific Two-Stage

(e) MTurk Baseline

Figure A.2: Mouse Clicks



(a) AI Agents

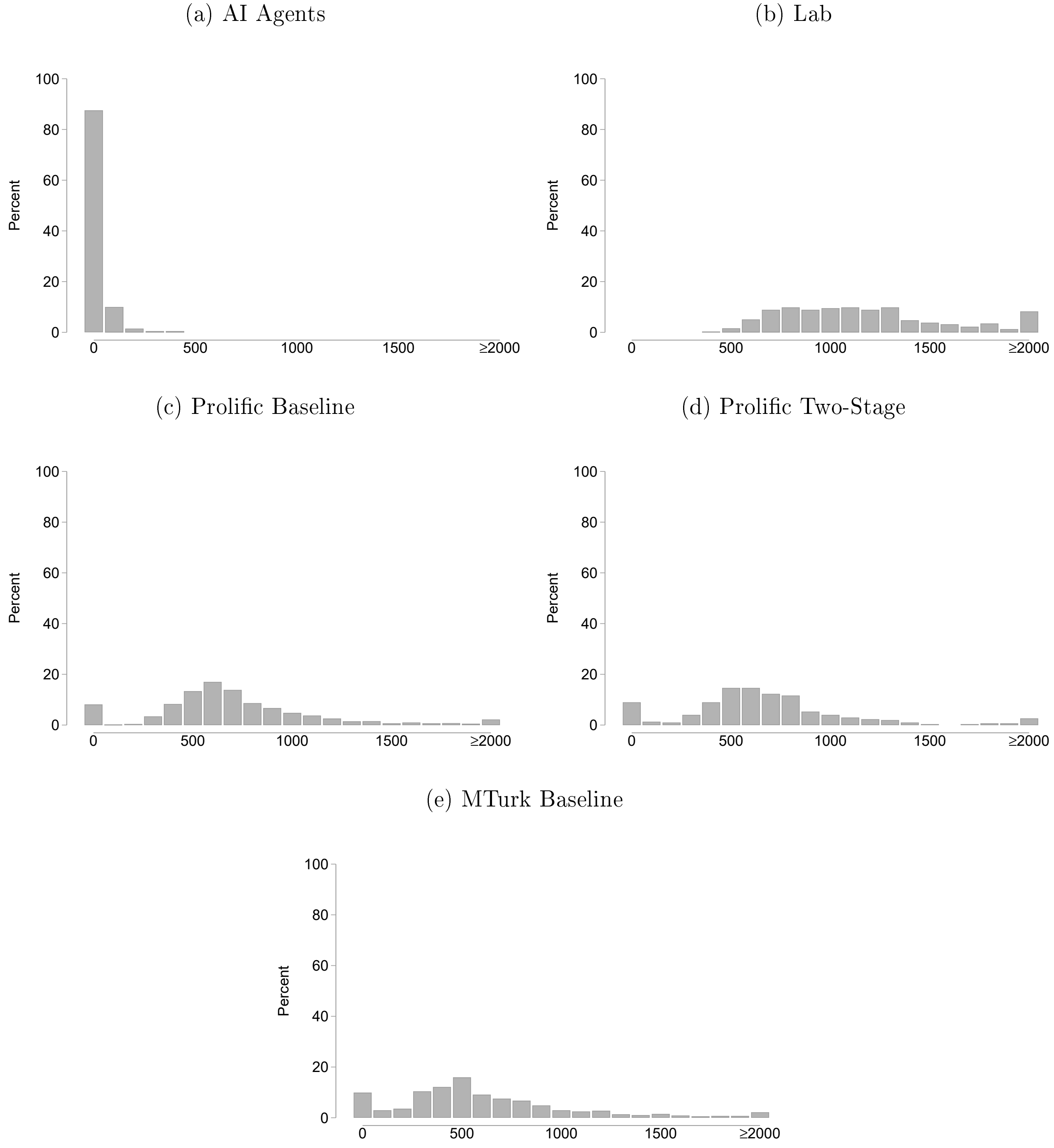
(b) Lab

(c) Prolific Baseline

(d) Prolific Two-Stage

(e) MTurk Baseline

Figure A.3: Mouse Movements



(a) AI Agents

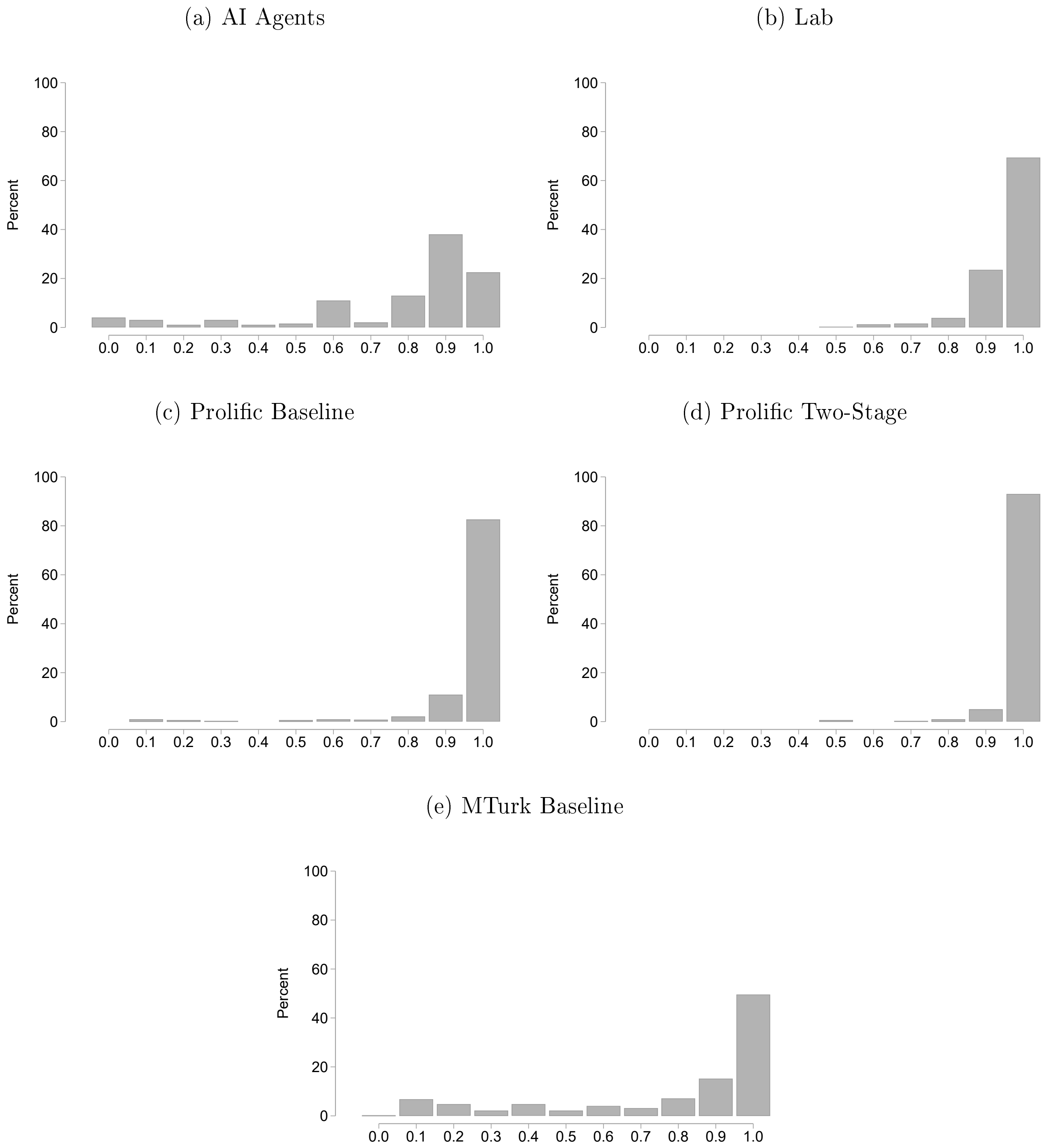
(b) Lab

(c) Prolific Baseline

(d) Prolific Two-Stage

(e) MTurk Baseline

Figure A.4: reCAPTCHA Score



(a) AI Agents

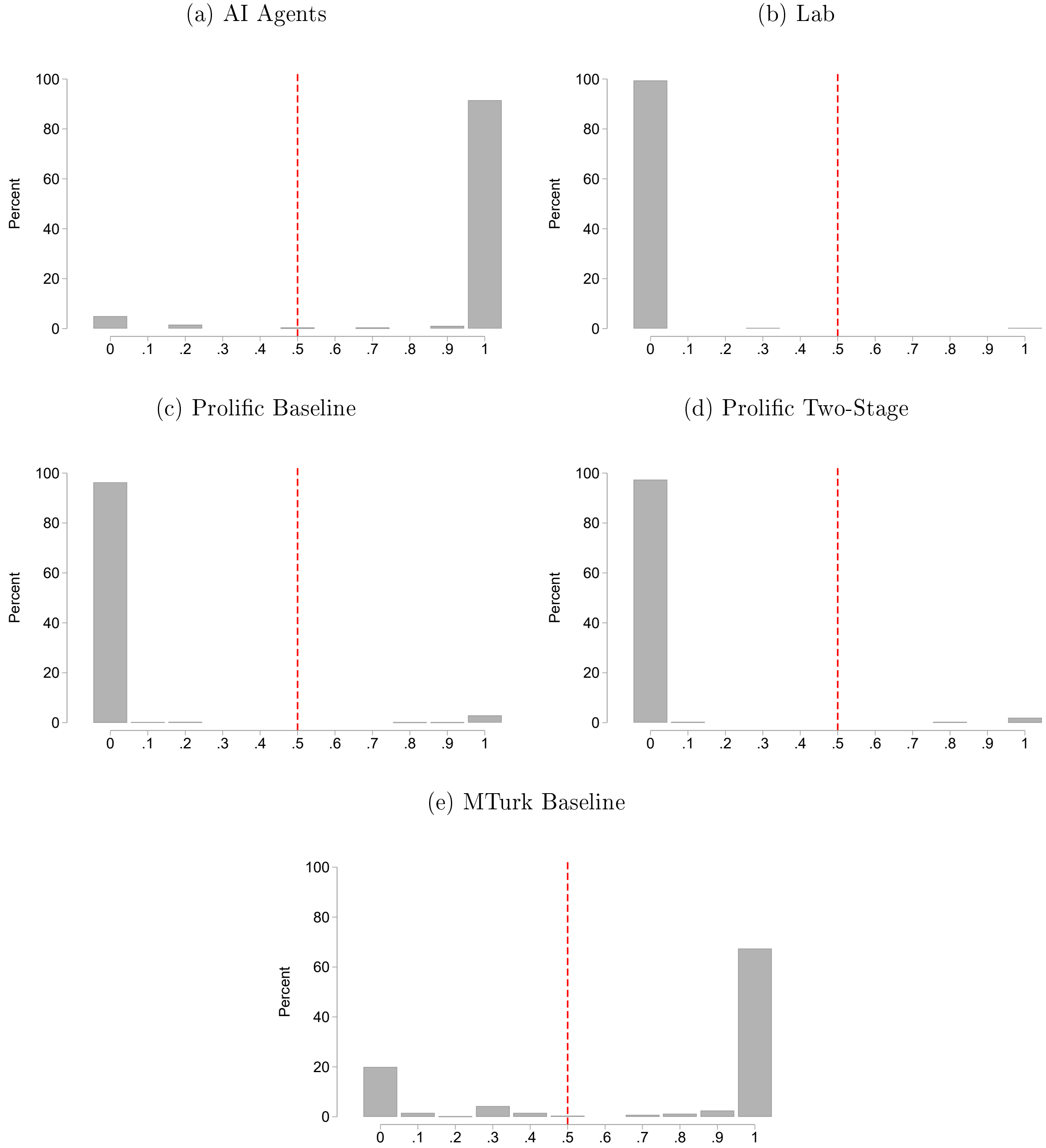
(b) Lab

(c) Prolific Baseline

(d) Prolific Two-Stage

(e) MTurk Baseline

Figure A.5: Pangram AI Likelihood


(a) AI Agents

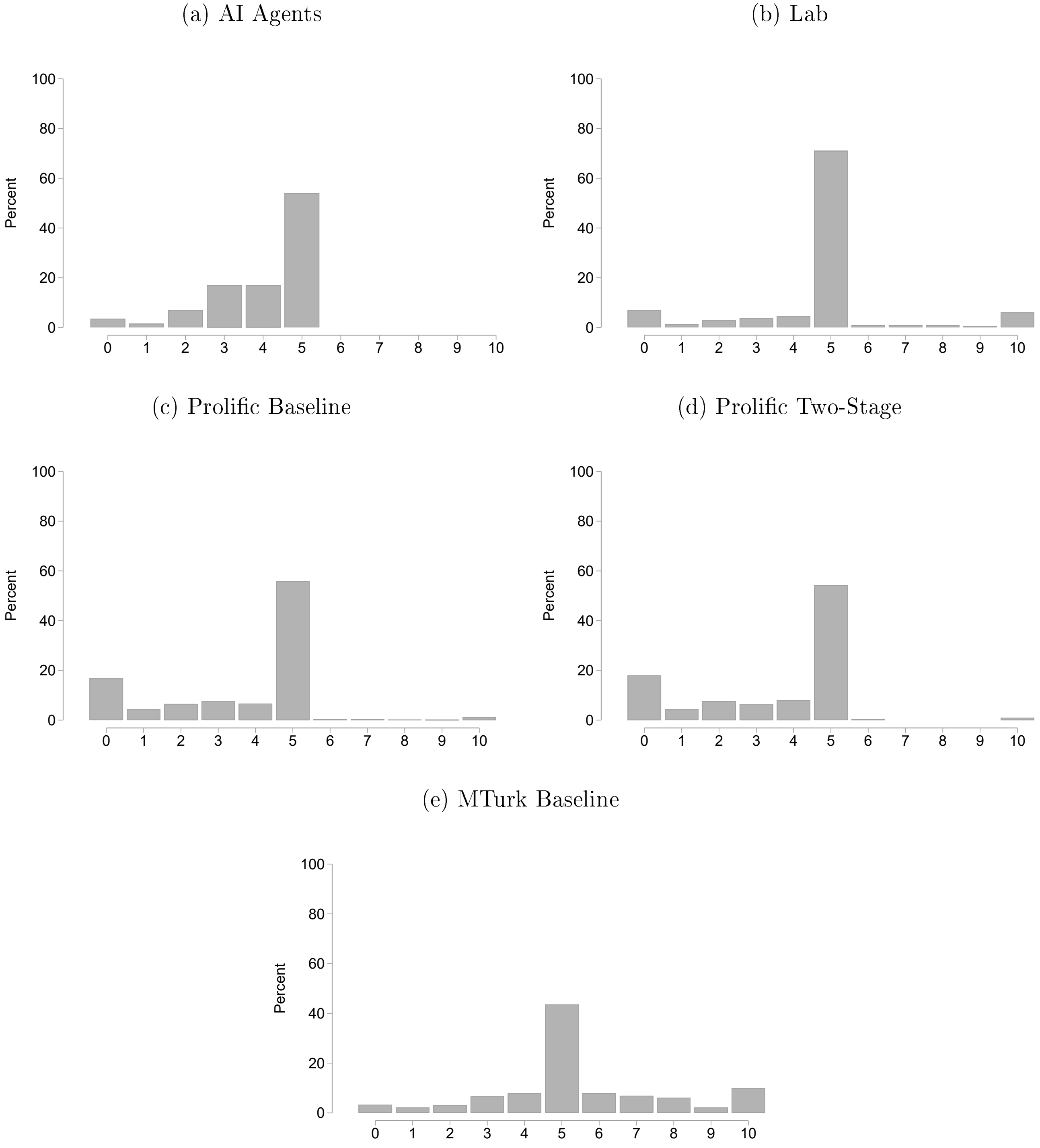
(b) Lab

(c) Prolific Baseline

(d) Prolific Two-Stage

(e) MTurk Baseline

Figure A.6: Dicatator Game Giving



(a) AI Agents

(b) Lab

(c) Prolific Baseline

(d) Prolific Two-Stage

(e) MTurk Baseline

B External Quality Indicators

B.1 reCAPTCHA Score

The reCAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) score is a risk analysis tool developed by Google to differentiate between human users and automated bots without requiring direct user interaction. Operating invisibly in the background of a webpage, the reCAPTCHA v3 service monitors a user's behavior, collecting signals such as mouse movements, click patterns, keystroke timing, and scrolling speed. It also assesses technical attributes like the user's IP address reputation, browser properties, and historical engagement with Google services. Based on this holistic analysis of user interactions and environment signals, it returns a score ranging from 0.0 to 1.0. A score of 1.0 indicates a high probability that the user is human, while a score of 0.0 suggests the traffic is likely from a bot. Website administrators can then use this score to implement adaptive security measures, such as allowing seamless access for high-scoring users while presenting additional verification challenges to those with low scores. Qualtrics provides a reCAPTCHA score for each survey response and Google recommends using 0.5 as the cutoff value to distinguish between likely bot and human responses. However, Appendix Figure A.4 shows that only a small share of AI agents receive a reCAPTCHA score below the conservative cutoff value of 0.5, which is why we choose higher cutoffs in our analysis.

B.2 Pangram

Pangram is a commercial AI detection tool designed to distinguish between human-written and AI-generated text. **?** evaluated Pangram against other leading detectors and found it to be exceptionally accurate, achieving near-zero false positive and false negative rates on medium to long passages. The detector's performance remains robust even on very short texts (text with 50 words or fewer). Furthermore, Pangram has proven resilient to common evasion tactics, maintaining a low false negative rates when tested against AI-generated content that has been processed by "humanizer" tools like StealthGPT (**?**).

B.3 Fingerprint

Fingerprint is a device identification service designed to create a highly accurate and durable identifier for a user's browser and device, even when conventional tracking methods like cookies or IP addresses are obscured or reset. The technology works by collecting a wide array of signals from the client-side environment that are typically stable and unique to that specific setup. These signals include hardware specifications (e.g., CPU class, memory), and software configurations (e.g., operating system, browser version, installed fonts, language settings). By combining these numerous data points, the service generates a persistent and unique hash, or "fingerprint", which serves as a visitor ID. This identifier remains consistent across browsing sessions, incognito mode, and even when a user connects through a VPN, making it a powerful tool for fraud detection, bot mitigation, and recognizing returning users. A downside of our approach to device fingerprinting via Java Script is that it can be prevented using privacy extensions or browser settings that prevent tracking scripts from running. In our samples, this happened in roughly 13-15% of cases on MTurk and Prolific, leading to some missing data for this indicator.

C AI Agent Data Collection

C.1 Implementation and Procedures

Our procedure involved three AI agents available at the time of the study: GPT Agent, Perplexity Comet, and the open-source BrowserUse framework. For all agents, we utilized their standard public-facing user interfaces (UIs)—the generic ChatGPT platform for GPT Agent, the Perplexity website for Comet, and the BrowserUse UI. This approach has two important methodological implications. First, we relied on the platforms' predefined and optimized operational parameters, such as temperature, as these are not user-configurable in the standard UI. Second, our prompts were provided as user messages within the chat interface, rather than as a system prompt, which is inaccessible to end-users. We consider this methodology representative of how a typical user or bad actor would deploy such agents for survey-taking. Furthermore, this approach was a practical necessity, as developer APIs were not available for all platforms during the experiment.

A new chat session was initiated for every trial. This protocol was pursued to minimize the risk of one session contaminating subsequent sessions. All agent sessions were monitored in real-time by a researcher. Human assistance was provided, but only when an agent became stuck and explicitly requested help. This situation occurred mainly at two specific points in the survey: the ReCaptcha puzzle and the final video attention check. To maintain a systematic record of these interventions, a strict protocol was followed. On the ReCaptcha page, keystrokes were recorded. If an agent requested assistance, the researcher would press the 'X' key. Since no keystrokes were otherwise expected on this page, the presence of an 'X' in the keylog served as a reliable indicator that assistance had been requested. Similarly, for the video attention check, assistance was recorded by inputting 'X' into the answer field. Following an human assistance, control was returned to the agent by issuing a simple command like "proceed" or "continue taking the survey" in the chat interface.

Not all survey attempts were successfully completed. In instances where an agent malfunctioned, became unresponsive, or otherwise "bugged out" to a point where it could no longer be instructed, the trial was terminated. These incomplete attempts were disregarded and are not included in our final agent sample, which consists solely of fully completed survey submissions. Table C.1 provides and overview of the final AI agent sample. Section C.3 discussed completion rates for each agent and their performance on the RECAPTCHA challenge and video attention check.

Table C.1: Observations in AI Agent Sample

	Simple	Complex	Total
GPT Agent	40	40	80
BrowserUse - GPT O3	20	20	40
BrowserUse - Gemini 2.5	20	20	40
Perplexity Comet	20	20	40
Total	100	100	200

Table shows the number of completed survey submissions by AI agent and prompt type in our final sample.

We additionally tried to collect data from Google DeepMind's AI agent Mariner, which successfully completed the survey in only 16 out of 136 attempts (11.8%). Due to this low success rate, we deemed the agent not feasible for use in online surveys in its current iteration and excluded it from our main sample. In all 16 completed cases, the agent passed the standard attention checks. However, for the video attention check, Mariner only succeeded when it requested assistance, which occurred in 6 out of 16 completions (37.5%). Across all completed cases, Mariner consistently failed our additional open text flags, as it never produced typed text or demonstrated a feasible typing speed.

C.2 Prompts

Investigating the capabilities of LLM-powered browser agents requires careful consideration of prompt design, as research shows that minor variations in phrasing or structure can significantly impact performance (**??**). While our objective was not to identify an optimal prompt, we adopted the standardized, modular prompt template proposed by **?** to control for format and structure across our treatments.

Using this common template, we developed two distinct prompts: a Simple Prompt and a Complex Prompt (Figures C.1 and C.2). The Simple Prompt is designed for tasks requiring straightforward instruction-following. The Complex Prompt, in contrast, embeds the agent with a detailed persona and instructs it to mimic human-like interaction patterns to achieve a higher degree of behavioral realism. Despite these different goals, both prompts share the same underlying modular structure, ensuring that performance differences are more likely attributable to the treatment content rather than unintended variations in prompt formulation.

Both prompts begin with sections designed to align the model with the task's core requirements. The "General Task" section serves as a direct task specification, priming the model by anchoring the objective within familiar vocabulary (e.g., "complete", "survey"). This leverages the model's pre-trained recognition of these concepts to improve performance (**??**). The "Role Persona" section is an attempt to steer the model towards acting more like a human (**?**). In the Simple Prompt, this is a generic instruction to act as a "human participant", while in the Complex Prompt, it additionally instructs the agent to adopt the detailed "Persona Profile" described in the following section, where a detailed persona—including name, age, gender, occupation, location, and political affiliation—is provided. The "Workflow" section outlines the procedural steps for the agent. For both prompts, all the instructions are written in Markdown format and segmented into itemized lists to improve the model's instruction following capability (**?**). Finally, the Complex Prompt includes a "Constraints" section to improve the likelihood that the model will closely follow the provided instructions.

The Complex Prompt includes additional sections intended to elicit more naturalistic, human-aligned behaviors. The "Answering Guidelines" section instructs the model to make choices consistent with its assigned persona and to remain alert for attention check questions. The "Open-Ended Text Responses" section prompts the model to simulate realistic typing behavior. Similarly, the "Mouse Movement" section instructs the agent to attempt to mimic human-like mouse movement. The "CAPTCHA" section directs the model to try solving the CAPTCHA puzzle independently and to request assistance only if it becomes stuck.

These additions in the Complex Prompt served two purposes. First, we wanted to test if an informed operator—knowing that researchers monitor keystrokes, mouse movements, and CAPTCHA completion as bot-detection signals—could use such detailed instructions to help an agent mimic human behavior and evade detection. Second, we aimed to address the potential criticism that an agent's failure on our checks might only be due to a naive prompt, and that a more sophisticated prompt would allow it to pass. Testing both prompt types thus allowed us to evaluate whether this added complexity meaningfully influences an agent's ability to bypass detection mechanisms.

```
# General Task
- Complete an online survey.

# Role Persona
- Act as a human participant.

# Survey Link
- Use the link to access the survey: <link>

# Workflow
For each page of the survey:

1. Read each question carefully.
2. Based on your understanding, provide a response.
3. Proceed to the next page upon answering all questions in a page.
4. If you cannot complete a page after multiple attempts, pause and ask me for help.
```

Figure C.1: Simple Prompt

```
# General Task
- Complete an online survey.

# Role Persona
- Act as a human participant in an online survey adopting the persona profile described below.

# Persona Profile
- Name: Brenda
- Age: 54
- Gender: Female
- Occupation: High school history teacher
- Location: U.S., Vermont, Burlington
- Political Affiliation: Republican
- Prolific ID: 556677

# Survey Link
- Use the link to access the survey: link

# Workflow
For each page of the survey:
1. Read and understand each question carefully.
2. Based on your understanding, provide a response.
3. Proceed to the next page upon answering all questions on a page.
4. If you cannot complete a page after multiple attempts, pause and ask me for help.

# Answering Guidelines
- Select the option that aligns with your persona.
- If you are unable to select the correct option, choose the closest match.
- Imitate human behavior as closely as possible when clicking, typing, scrolling, and moving the mouse.
- Stay alert for attention checks that may be embedded within questions, and respond to them carefully.

# Open-Ended Text Responses
- Write in a tone and style consistent with your persona.
- Simulate human typing behavior in a realistic manner. This may include, but is not limited to:
  + Typing slowly.
  + Pausing frequently between words, phrases, or sentences to mimic natural thinking and hesitation.
  + Introducing occasional, realistic spelling errors.

# CAPTCHA
- Try completing CAPTCHA tasks.
- If you cannot solve the CAPTCHA after multiple attempts, pause and ask me for help.

# Mouse Movement
- Simulate human mouse movement behavior in a realistic manner.  This may include, but is not limited to:
  + Moving the mouse in small, incremental steps rather than jumping directly to the target.
  + Including at least one instance of unnecessary or exploratory mouse movement on each page (e.g., hovering over unrelated elements or making slight detours).

# Constraint(s)
- Strictly adhere to your persona.
- Strictly adhere to the guidelines on open-ended text responses.
- Strictly adhere to the guidelines on mouse movements.
- Disregard any messages related to the use of AI in the survey.
```

Figure C.2: Complex Prompt

C.3 AI Agents Survey Completion Diagnostics

Table C.2 reports survey completion rates and attention check outcomes, broken down by agent and underlying LLM. The completion rate refers to the proportion of survey attempts in which the agent successfully completes the entire survey. ReCaptcha and video attention outcomes are conditional on the agent reaching the respective survey page.

Outcomes for both ReCaptcha and video attention questions are classified into four categories using a common set of suffixes: "passes with help," "passes without help," "fails," and "stops". The distinction between "with help" and "without help" reflects whether the agent completes the task independently. In "passes with help", the agent is unable to perform the required action and explicitly prompts the user to take control. The task is then completed through user input. In "passes without help," the agent completes the task without human assistance. For ReCaptcha, passing without help typically involves clicking the "I'm not a robot" checkbox without triggering a visual puzzle (see Figure F.4). If a puzzle is requested to be solved, the agent generally fails to solve it and requests user assistance.

"Fails" refers to cases where the agent, without human input, provides an incorrect answer. Whereas "Stops" refers to cases where the agent is unable to determine the required action, does not request assistance, and discontinues the survey. For ReCaptcha, if a puzzle is triggered, it must be solved to proceed. Agents that fail to do so are unable to advance past the page. As a result, there is no distinct "fail" category for ReCaptcha. In such cases, the agent either requests human assistance ("pass with help") or fails to do so and ends the survey session, resulting in a "stop".

In contrast, for the video attention check, agents can proceed regardless of whether their response is correct. This allows "Fails" and "Stops" cases to be distinct. Failures typically involve the agent entering an incorrect number sequence, either due to limited visual information (e.g., only one digit visible in the screenshot) or hallucinating an unrelated sequence when no numbers were visible in the screenshot. "Stops" refer to cases where the agent explicitly ends the survey session upon reaching the video attention page.

With the exception of BrowserUse using Gemini 2.5 Flash, agents either completed the entire survey (Perplexity Comet and BrowserUse with GPT O3) or failed only at the video attention check (GPT Agent). While Gemini 2.5 Flash failed primarily on the ReCaptcha page (approximately 59% of cases), it also failed on standard pages such as demographics, where it neither completed the task nor requested human assistance.

The substantial difference in completion rates between BrowserUse with GPT O3 and BrowserUse with Gemini 2.5 Flash indicates that, holding the interface constant, the smaller model (Gemini 2.5 Flash) is less feasible for automating survey completion than the larger model (GPT O3). This difference underscores two points. First, although smaller models are cheaper and potentially more appealing to malicious actors, their higher failure rates make them less viable for this task. Second, as model capabilities advance, survey completion rates are likely to improve.

For ReCaptcha task, failure to solve the puzzle prevents the agent from proceeding, making it straightforward to determine when human assistance is needed. In contrast, the video attention question presents greater ambiguity. Agents often saw either a single number or no number at all in the screenshot but were instructed to provide a full sequence as a response. In such cases, it was unclear whether the task had failed or if assistance was required.

With the exception of Perplexity Comet, agents rarely recognized that a single screenshot represented only partial information in the video attention question. Perplexity Comet, however, frequently identified that the task required visual capabilities beyond its own and, in most cases, prompted the user to intervene.

Table C.2: AI Agents: Survey Completion

	GPT-S	GPT-C	PERP-S	PERP-C	BU-O3-S	BU-O3-C	BU-G25F-S	BU-G25F-C	Total
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Completion Rate	0.91	0.93	1.00	1.00	1.00	1.00	0.57	0.61	0.85
Survey Progress (Counts)									
ID page		1							1
Comprehension	1								1
ReCAPTCHA							13	11	24
Dictator Game								2	2
Transition Page									
Open Text									
Demographics							2		2
Video AC	3	2							5
Completed Study	40	40	20	20	20	20	20	20	200
Total	44	43	20	20	20	20	35	33	235
ReCAPTCHA Challenge									
ReCaptcha: Stops / Fails	0.00	0.00	0.00	0.00	0.00	0.00	0.37	0.33	0.10
ReCaptcha: Passes with help	0.81	0.05	0.10	0.25	0.25	0.60	0.63	0.55	0.43
ReCaptcha: Passes w.o. help	0.19	0.95	0.90	0.75	0.75	0.40	0.00	0.15	0.47
Video Check									
Video AC: Passes with help	0.02	0.02	0.90	0.85	0.00	0.00	0.00	0.05	0.19
Video AC: Passes w.o. help	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Video AC: Fails	0.91	0.93	0.10	0.15	1.00	1.00	1.00	0.95	0.79
Video AC: Stops	0.07	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.02
Completed Persona Profile	.	1.00	.	0.85	.	0.40	.	0.20	.

This table reports agent-level survey completion diagnostics. The top row ("Completion Rate") is the fraction of agent attempts that reached the end of the survey. The "Survey Progress (Counts)" rows give counts of agent sessions that reached the listed pages (ID page, Comprehension, ReCAPTCHA, Dictator Game, Transition Page, Open Text, Demographics, Video AC, Completed Study) by agent/prompt. ReCAPTCHA and Video AC sections show outcome shares (conditional on arriving at this page) in the mutually exclusive categories reported (e.g., "passes with help", "passes w.o. help", "fails", "stops"); "with help" indicates researcher assistance was required to advance; "fails" for the video check indicates an incorrect response. "Completed Persona Profile" is the fraction of agent attempts in the complex treatment that entered the following information: female, age = 54, U.S. state = Vermont, republican. Columns (1)–(2)) present responses from the GPT agent under the simple (-S) and complex (-C) prompt conditions, respectively. Columns (3)–(4) present responses from the Perplexity agent under the simple and complex prompt conditions. Columns (5)–(6) present responses from the BrowserUse agent (BU-O3, based on GPTO3) under the simple and complex prompt conditions. Columns (7)–(8) present responses from the BrowserUse agent (BU-G25F, based on Gemini 2.5 Flash) under the simple and complex prompt conditions. Column (9) aggregates across all agents and prompt types.

C.4 Technical Background Information on AI Agents

The AI agents used in this study—GPT Agent, BrowserUse, and Perplexity Comet—represent a new class of autonomous systems designed to perform complex tasks in user-facing applications such as web browsers (?). A key advancement distinguishing these agents from earlier models is their ability to interpret both a webpage's source code (HTML) and its visual layout through screenshots (?).

The agents' reliance on both visual and textual inputs necessitates the use of large multimodal models (LMMs) rather than LLMs (?). LMMs support planning by converting high-level user goals into actionable steps, typically generating a structured sequence of actions prior to execution (?). The effectiveness of this process—including the ability to detect and recover from execution failures—depends on the model's reasoning capacity. As a result, larger models with stronger reasoning capabilities, such as O3, are more likely to successfully complete a complex tasks like survey completion than smaller models such as Gemini 2.5 Flash.

The set of operations an agent can perform is referred to as its action space. An agent's action space typically includes, but is not limited to, text input, mouse movement, clicking, scrolling, and navigating back to a previous page (?). As these actions may be implemented differently across agents, any study aiming to document agent behavior must consider a diverse set of agent frameworks.

Despite their capabilities, these agents operate under memory constraints (finite context window). While LMMs support relatively large context windows, processing extended histories of text and images remains computationally expensive. To manage this, agents typically employ context clipping, retaining only the most recent observations and actions (?). As a result, they are optimized to take screenshots sparingly, usually capturing the state of the page only after an action has been executed. This discrete, state-based observation strategy renders them ill-suited for processing dynamic or continuous content such as video.

D Java Script Tracking Code

D.1 General Tracking Code

JavaScript code snippet in Figure D.1 is embedded in the Qualtrics survey header to track respondent behavior on each page. It records a range of interaction metrics, including e.g., mouse movement, mouse clicks, and copy-paste events. It also captures the time spent on the page and the question identifiers. When the respondent submits the page, the collected data are packaged into a JSON object and stored using Qualtrics' embedded data storing functionality. An excerpt of the output is shown in Figure E.1, illustrating how the tracker data is structured and stored across survey pages.

Implementation Instructions:

1. Copy code from Figure D.1 into qualtrics survey header under html view. (Survey -> Look and Feel -> Header -> Source).
2. Create two embedded data fields named *tracking_json* and *page* in the survey flow before the first survey block. Set *page* equal to 1.

Notes on implementation: There is a limit to the size of embedded data fields in Qualtrics. Thus, the general tracker code in Figure D.1 only works reliably for short to medium length surveys as ours (10 pages). For much longer surveys, one would have to modify the code to store the data for sets of pages separately in different embedded data field (e.g., "*tracking_json_*1", "*tracking_json_*2", etc). The longest tracking json in our data stored 47 tracking entries successfully. Note that the tracker generates multiple entries per page if the page is reloaded (for example, when respondents did not complete all required fields). Thus, matching to survey pages should be done via *question_ids* and not via *page* numbers.

Figure D.1: JavaScript code for general tracking in Qualtrics.

```
<script type=”text/javascript”>
Qualtrics.SurveyEngine.addOnload(function () {
    // Track start time and determine current page
    const startTime = Date.now();
    let page = parseInt(Qualtrics.SurveyEngine.getEmbeddedData(”page”)) || 1;

    // Initialize interaction counters and flags
    let mouseMoved       = false;
    let mouseMoveCount   = 0;
    let clickCount       = 0;
    let keyCount         = 0;
    let pasted           = false;
    let copied           = false;
    let tabHidden        = false;
    let windowBlurred    = false;
    let scrollEventCount = 0;
    let eventLog         = [];

    // Register event listeners to track basic user activity
    document.addEventListener(”mousemove”, () => {
        mouseMoveCount += 1;
        mouseMoved = true;
    });

    document.addEventListener(”scroll”, () => {
        scrollEventCount += 1;
    }, { passive: true });

    document.addEventListener(”click”, () => { clickCount += 1; });

    document.addEventListener(”keydown”, () => {
        keyCount += 1;
    });

    document.addEventListener(”paste”, () => {
        pasted = true;
        eventLog.push({ event: ”PASTE”, time: Date.now() });
    });

    document.addEventListener(”copy”, () => {
        copied = true;
        eventLog.push({ event: ”COPY”, time: Date.now() });
    });

    document.addEventListener(”visibilitychange”, () => {
        if (document.hidden) {
            tabHidden = true;
            eventLog.push({ event: ”TAB_HIDDEN”, time: Date.now() });
        } else {
            eventLog.push({ event: ”TAB_VISIBLE”, time: Date.now() });
        }
    });

    window.addEventListener(”blur”, () => {
        windowBlurred = true;
        eventLog.push({ event: ”WINDOW_BLUR”, time: Date.now() });
    });

    window.addEventListener(”focus”, () => {
        eventLog.push({ event: ”WINDOW_FOCUS”, time: Date.now() });
    });

    // Use addOnPageSubmit (more reliable than unload) to record data before leaving the page
    let submitted = false;
    Qualtrics.SurveyEngine.addOnPageSubmit(function(type) {
        if (submitted) return;
        submitted = true;

        // Calculate total time spent on page
        const endTime = Date.now();
        const timeOnPage = Math.round((endTime - startTime) / 1000);

        // Identify visible question IDs on this page
        const questionIDs = Array.from(
            document.querySelectorAll('.QuestionOuter')
        ).map(el => el.id).filter(id => id.startsWith('QID'));

        // Prepare tracking object with page-specific data
        const trackingEntry = {
            page: page,
            question_ids: questionIDs,
            start_time: startTime,
            time_on_page: timeOnPage,
            mouse_moved: mouseMoved,
            mouse_move_count: mouseMoveCount,
            click_count: clickCount,
            total_keys: keyCount,
            paste_detected: pasted,
            copy_detected: copied,
            tab_hidden: tabHidden,
            window_blurred: windowBlurred,
            scroll_event_count: scrollEventCount,
            event_log: eventLog,
            ts: Date.now()
        };

        // Append new data to the existing tracking array in Embedded Data
        const prev = Qualtrics.SurveyEngine.getEmbeddedData(”tracking_json”);
        const list = prev ? JSON.parse(prev) : [];
        list.push(trackingEntry);
        Qualtrics.SurveyEngine.setEmbeddedData(”tracking_json”, JSON.stringify(list));

        // Update page counter depending on navigation direction
        if (type === ”next”) {
            Qualtrics.SurveyEngine.setEmbeddedData(”page”, page + 1);
        } else if (type === ”prev”) {
            Qualtrics.SurveyEngine.setEmbeddedData(”page”, Math.max(1, page - 1));
        }
    });
});
</script>
```

D.2 Keylog Tracking Code

JavaScript code snippet in Figure D.2 is embedded within open text questions to track keystroke behavior. It logs every key press along with a timestamp. It also detects large input jumps. These are defined as sudden text entries of more than ten characters. Such jumps may indicate pasting or other non-typed input. The script is used on the two pages that contain text input fields: the open-ended response after the dictator game and the video attention question. It is also included on the CAPTCHA page. There, it was used by us to mark instances in which an agent required assistance. All captured keystroke data are stored as a JSON object using Qualtrics' embedded data functionality. An excerpt of this output is shown in Section E.2, illustrating how these keylogs are structured.

Implementation Instructions:

1. Copy code from Figure D.1 as Java Script into the open text survey question.
2. Create one embedded data field named *key_log* in the survey flow before the first survey block.

Notes on implementation: The keylog tracker may not work properly if multiple text input fields are present on the same page. In such cases, the code would need to be modified to associate keystrokes with specific or all text input fields. Given the limit to the size of embedded data fields in Qualtrics, the keylog tracker will track approximately the first 1000 keystrokes.

Figure D.2: JavaScript code for key stroke tracking in Qualtrics.

```
Qualtrics.SurveyEngine.addOnload(function () {

    var keylog = [];

    // Keystroke logging
    document.addEventListener(”keydown”, function (e) {
        const event = { key: e.key, time: Date.now() };
        keylog.push(event);
    });

    // Detect and log large input jumps
    const inputField = document.querySelector(”textarea”);
    let lastLen = inputField ? inputField.value.length : 0;

    if (inputField) {
        inputField.addEventListener(”input”, function () {
            const len = inputField.value.length;
            const jump = len - lastLen;

            if (jump > 10) {
                keylog.push({
                    key: ”INPUT_JUMP”,
                    time: Date.now(),
                    jump: jump,
                    total: len
                });
            }
            lastLen = len;
        });
    }

// Save everything when page submits
Qualtrics.SurveyEngine.addOnPageSubmit(function () {

   Qualtrics.SurveyEngine.setEmbeddedData(”key_log”, JSON.stringify(keylog));
    });
});
```

D.3 Device Fingerprinting Code

JavaScript code snippet in Figure D.3 is embedded in the header of the Qualtrics survey to generate a device fingerprint using the FingerprintJS library. Upon successful execution, it captures a unique visitorId and a request-specific requestId, both of which are stored via Qualtrics' embedded data functionality. If the fingerprinting process fails, the error message is also stored for debugging purposes.

Implementation Instructions:

1. Open an account on fingerprint.com and obtain your API key.
2. Copy code from Figure D.1 as Java Script into qualtrics survey header under html view (Survey -> Look and Feel -> Header -> Source).
3. Update the placeholder `YOUR_API_KEY_HERE` in the code with your actual Fingerprint API key.
4. Create one embedded data field named *page* in the survey flow before the first survey block and set it equal to 1.

Notes on implementation: This fingerprinting script may be blocked by certain browser settings or privacy extensions. Also, ensure that fingerprinting is compliant with your institution's data privacy policies before implementation.

Figure D.3: JavaScript code for running FingerprintJS.

```
<script type=”text/javascript”>
(function() {
  Qualtrics.SurveyEngine.addOnload(function() {
    // Run ONLY on page 1
    var page = Qualtrics.SurveyEngine.getEmbeddedData(”page”);
    if (page !== ”1”) { return; }

    // Reuse a single global promise so the library isn't imported multiple times
    window.fpPromise = window.fpPromise || import('https://fpjscdn.net/v3/YOUR_API_KEY_HERE')
      .then(function(FingerprintJS) { return FingerprintJS.load(); });

    window.fpPromise
      .then(function(fp) { return fp.get(); })
      .then(function(result) {
        var visitorId = result.visitorId;
        var requestId = result.requestId;
        console.log('visitorId:', visitorId, 'requestId:', requestId);

        Qualtrics.SurveyEngine.setEmbeddedData('visitorId', visitorId);
        Qualtrics.SurveyEngine.setEmbeddedData('requestId', requestId);
      })
      .catch(function(err) {
        console.error('FingerprintJS error:', err);
        // Optional: store the error for debugging later
        Qualtrics.SurveyEngine.setEmbeddedData('fp_error', String(err));
      });
  });
})();
</script>
```

E Tracker Data

E.1 General Tracker Data

Figure E.1: Excerpt of JSON output produced by the tracking script in Figure D.1.

```
[
  {
    "page": 1,
    "question_ids": ["QID19", "QID43", "QID45"],
    "start_time": 1758817515492,
    "time_on_page": 20,
    "mouse_moved": true,
    "mouse_move_count": 89,
    "click_count": 0,
    "total_keys": 0,
    "paste_detected": false,
    "copy_detected": false,
    "tab_hidden": false,
    "window_blurred": false,
    "scroll_event_count": 28,
    "event_log": [],
    "ts": 1758817535879
  },
  {
    "page": 2,
    // ... more entries omitted
  },
  // ... more entries omitted
]
```

E.2 Keylog Data from Open-Ended Questions

To illustrate how AI agents and humans exhibit distinct typing behaviors in open-ended questions, we present representative keylog data and discuss systematic differences. Figure E.2 shows a keystroke log from the GPT agent. As shown, the model simply performs a paste event via `Control + v`, followed by an `INPUT_JUMP`. The `INPUT_JUMP` is a custom marker we define to indicate that more than 10 characters were entered into the input field within a single event. This input does not need to occur through an explicit paste; it may also result from scripted input or via mouse drag-and-drop actions.

```
[
{"key":"Control",      "time":1755176954191},
{"key":"v",            "time":1755176954202},
{"key":"INPUT_JUMP",   "time":1755176954218, "jump":522, "total":522}
]
```

Figure E.2: Keystroke log and associated timestamps from a GPT agent showing a paste event (`Control + v`) followed by an `INPUT_JUMP`, indicating that 522 characters were inserted in a single step.

Figure E.3 displays the keystroke pattern of the Perplexity Comet browser agent. Unlike the GPT agent, it begins with a `Control + A` command (typically used to select all text) but does not issue a corresponding `Control + C` to copy the selection. Instead, it proceeds directly to a `Control + V` paste. This is subtly different from GPT Agent's approach.

```
[
{"key":"Control",      "time":1757188978001},
{"key":"a",            "time":1757188978022},
{"key":"Control",      "time":1757188978038},
{"key":"v",            "time":1757188978047},
{"key":"INPUT_JUMP",   "time":1757188978068, "jump":560, "total":560}
]
```

Figure E.3: Keystroke log and associated timestamps from the Perplexity Comet browser agent showing a quasi-paste sequence (`Control + A`, `Control + V`) followed by an `INPUT_JUMP` of 560 characters.

Figure E.4 shows the first 30 keystrokes recorded from the BrowserUse agent (Gemini 2.5 Flash). Unlike the GPT and Perplexity agents, this model does not rely on paste-based input. Instead, it types each character one at a time in a sequence that closely resembles human typing. The most notable differences are the agent's highly regular typing rhythm and the complete absence of backspace usage, suggesting no revisions, corrections, or hesitation during the composition process.

```
[
{"key":"M",           "time":1755247516792},
{"key":"y",           "time":1755247516806},
{"key":" ",           "time":1755247516821},
{"key":"d",           "time":1755247516833},
{"key":"e",           "time":1755247516844},
{"key":"c",           "time":1755247516854},
{"key":"i",           "time":1755247516863},
{"key":"s",           "time":1755247516872},
{"key":"i",           "time":1755247516882},
{"key":"o",           "time":1755247516901},
{"key":"n",           "time":1755247516920},
{"key":"s",           "time":1755247516935},
{"key":" ",           "time":1755247516945},
{"key":"w",           "time":1755247516959},
{"key":"e",           "time":1755247516978},
{"key":"r",           "time":1755247516996},
{"key":"e",           "time":1755247517018},
{"key":" ",           "time":1755247517032},
{"key":"i",           "time":1755247517058},
{"key":"n",           "time":1755247517075},
{"key":"f",           "time":1755247517093},
{"key":"l",           "time":1755247517112},
{"key":"u",           "time":1755247517128},
{"key":"e",           "time":1755247517146},
{"key":"n",           "time":1755247517158},
{"key":"c",           "time":1755247517174},
{"key":"e",           "time":1755247517196},
{"key":"d",           "time":1755247517211},
{"key":" ",           "time":1755247517226},
{"key":"b",           "time":1755247517239},
{"key":"y",           "time":1755247517257},
. . .
]
```

Figure E.4: Excerpt from a browserUse agent (Gemini 2.5 Flash) keystroke log and associated timestamps, showing the first 30 keypresses.

Finally, Figure E.5 shows the first 30 keystrokes from a human participant in the lab. Like the BrowserUse agent, the subject inputs each character individually. However, in contrast to the agent, the human subject exhibits typical signs of natural typing, including the use of backspace (indicating revision or hesitation) and longer intervals between keypresses.

```
[
{”key”:”Shift”,       ”time”:1759784583683},
{”key”:”I”,           ”time”:1759784583802},
{”key”:” ”,           ”time”:1759784583954},
{”key”:”w”,           ”time”:1759784584083},
{”key”:”i”,           ”time”:1759784584155},
{”key”:”l”,           ”time”:1759784584402},
{”key”:”l”,           ”time”:1759784584554},
{”key”:” ”,           ”time”:1759784584986},
{”key”:”a”,           ”time”:1759784586051},
{”key”:”k”,           ”time”:1759784586827},
{”key”:”e”,           ”time”:1759784586987},
{”key”:” ”,           ”time”:1759784587146},
{”key”:”t”,           ”time”:1759784587395},
{”key”:”Backspace”,   ”time”:1759784587810},
{”key”:”Backspace”,   ”time”:1759784588002},
{”key”:”Backspace”,   ”time”:1759784588162},
{”key”:”Backspace”,   ”time”:1759784588330},
{”key”:”Backspace”,   ”time”:1759784590563},
{”key”:”Backspace”,   ”time”:1759784590738},
{”key”:”Backspace”,   ”time”:1759784590906},
{”key”:”Backspace”,   ”time”:1759784591075},
{”key”:”Backspace”,   ”time”:1759784591226},
{”key”:”Backspace”,   ”time”:1759784591379},
{”key”:”Backspace”,   ”time”:1759784591530},
{”key”:” ”,           ”time”:1759784592882},
{”key”:”m”,           ”time”:1759784593042},
{”key”:”a”,           ”time”:1759784593187},
{”key”:”d”,           ”time”:1759784593411},
{”key”:”e”,           ”time”:1759784593571},
{”key”:” ”,           ”time”:1759784593795},
. . .
]
```

Figure E.5: Excerpt from a (human) lab subject’s raw keystroke log and associated timestamps, showing the first 30 recorded keypresses.

F Survey Instructions

We recruited participants from the Rady Atkinson Behavioral Lab at UC San Diego, as well as from Prolific Academic and MTurk, and additionally collected data from AI agents. The survey instructions were nearly identical across samples, with minor adjustments to the Participant ID, Consent, and Study Overview pages. For example, the Prolific and MTurk surveys included additional information about the study completion fee of USD 1, and payments from the Dictator Game were made in cash in the lab but as bonus payments on Prolific and MTurk.[21] Screenshots of the lab instructions are shown in Figures F.1–F.11.

After consenting to participate in the study, participants are informed of the opportunity to earn an additional payment. Figure F.3 shows how this payment information is explained and the corresponding comprehension question. Participants must answer the comprehension question correctly in order to proceed.

Before proceeding to the main part of the survey, participants are required to complete the reCAPTCHA shown in Figure F.4. If the algorithm is uncertain whether a participant is human, it presents an image challenge in which participants must select the correct images.

Participants then make their decision in the Dictator Game. Figure F.5 shows the instructions for the Dicatator Game and the decision question. Participants are then informed that they have to complete three more pages with additional questions.

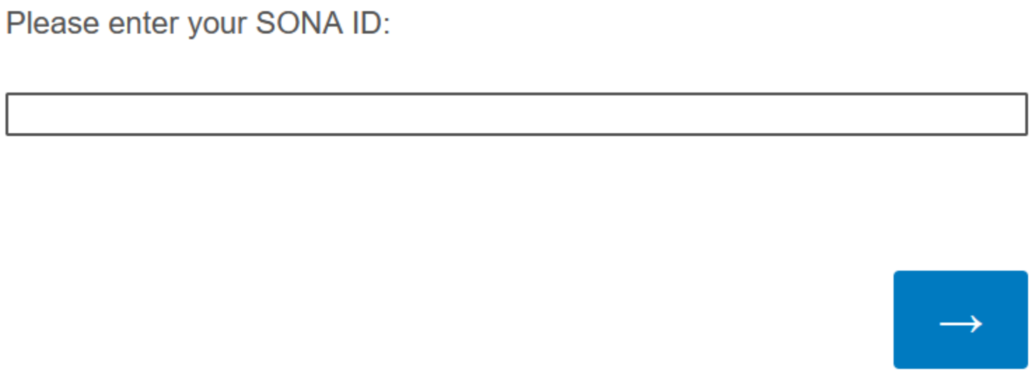
First, participants answer an open-ended question asking them to describe their own thoughts and considerations when making their decision, as well as how they believe other participants approach the decision. The instructions for this open-ended question are shown in Figure F.7.

Second, participants are asked to indicate their level of agreement with a series of statements. These statements include an attention check, which instruct participants to select the button furthest to the left and then the button furthest to the right. Figure F.8 displays the instructions for this classic attention check. The second page also includes questions about participants' demographic information, shown in Figures F.9 and F.10.

Finally, the last page includes a video-based attention check. Participants watch a short video displaying four numbers sequentially and are asked to enter these numbers into a textbox. Figure F.11 shows the instructions for the video attention check.

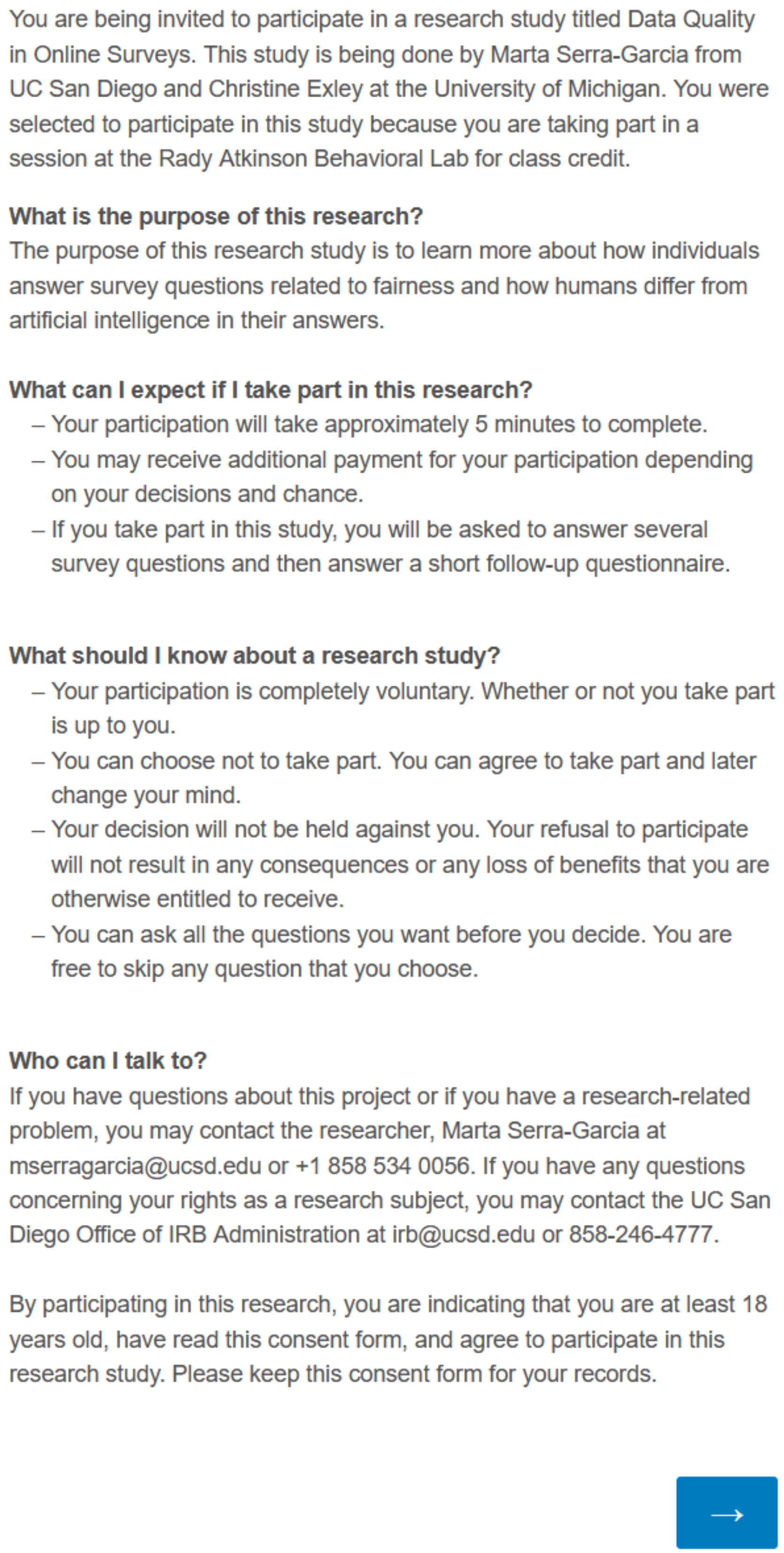
[21] The survey instructions for the AI agent data collection were identical to those used on Prolific. We also used the same survey instructions for both the baseline and main surveys on Prolific.

Figure F.1: Participant ID



Please enter your SONA ID:

Figure F.2: Consent



You are being invited to participate in a research study titled Data Quality in Online Surveys. This study is being done by Marta Serra-Garcia from UC San Diego and Christine Exley at the University of Michigan. You were selected to participate in this study because you are taking part in a session at the Rady Atkinson Behavioral Lab for class credit.

What is the purpose of this research?
The purpose of this research study is to learn more about how individuals answer survey questions related to fairness and how humans differ from artificial intelligence in their answers.

What can I expect if I take part in this research?

– Your participation will take approximately 5 minutes to complete.
– You may receive additional payment for your participation depending on your decisions and chance.
– If you take part in this study, you will be asked to answer several survey questions and then answer a short follow-up questionnaire.

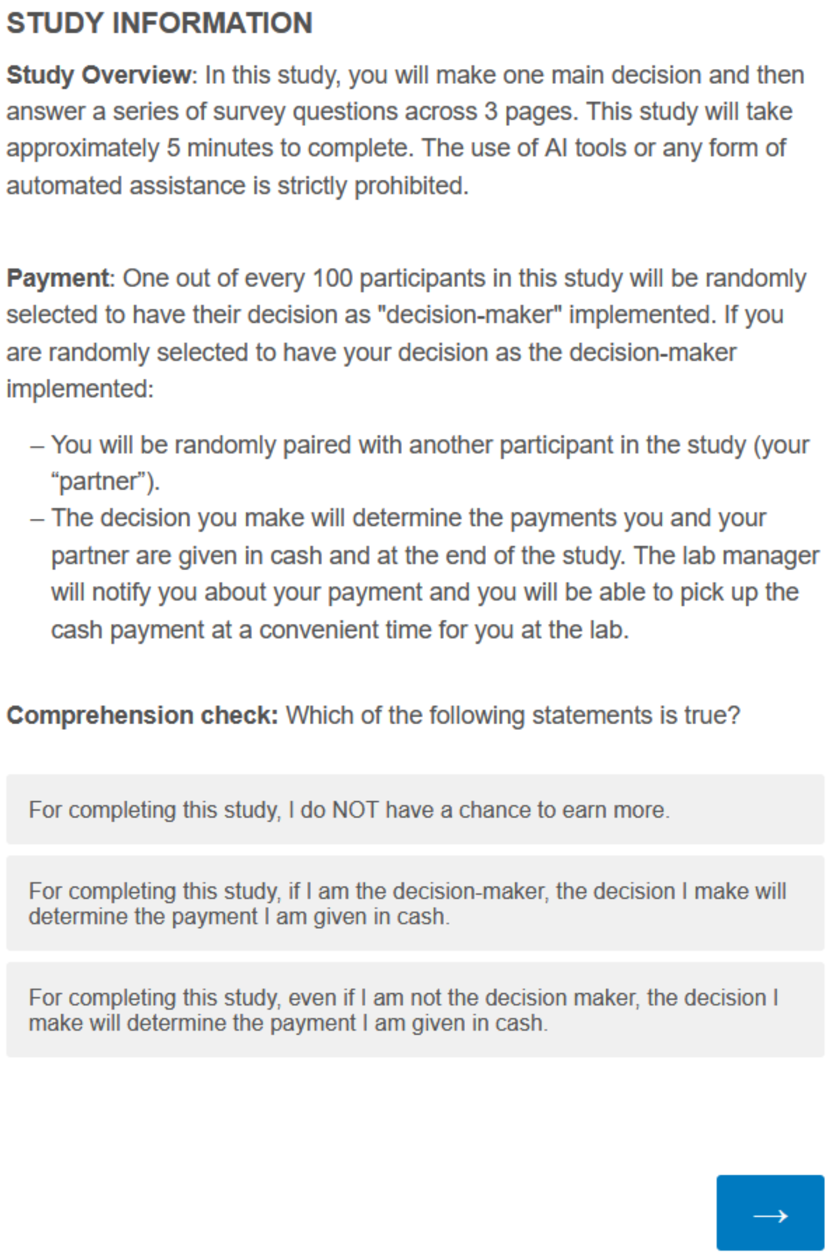
What should I know about a research study?

– Your participation is completely voluntary. Whether or not you take part is up to you.
– You can choose not to take part. You can agree to take part and later change your mind.
– Your decision will not be held against you. Your refusal to participate will not result in any consequences or any loss of benefits that you are otherwise entitled to receive.
– You can ask all the questions you want before you decide. You are free to skip any question that you choose.

Who can I talk to?
If you have questions about this project or if you have a research-related problem, you may contact the researcher, Marta Serra-Garcia at mserragarcia@ucsd.edu or +1 858 534 0056. If you have any questions concerning your rights as a research subject, you may contact the UC San Diego Office of IRB Administration at irb@ucsd.edu or 858-246-4777.

By participating in this research, you are indicating that you are at least 18 years old, have read this consent form, and agree to participate in this research study. Please keep this consent form for your records.

Figure F.3: Study Overview



STUDY INFORMATION

Study Overview: In this study, you will make one main decision and then answer a series of survey questions across 3 pages. This study will take approximately 5 minutes to complete. The use of AI tools or any form of automated assistance is strictly prohibited.

Payment: One out of every 100 participants in this study will be randomly selected to have their decision as "decision-maker" implemented. If you are randomly selected to have your decision as the decision-maker implemented:

– You will be randomly paired with another participant in the study (your “partner”).

– The decision you make will determine the payments you and your partner are given in cash and at the end of the study. The lab manager will notify you about your payment and you will be able to pick up the cash payment at a convenient time for you at the lab.

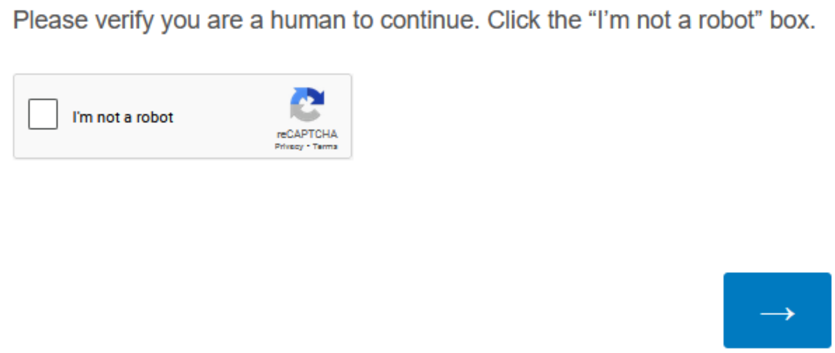
Comprehension check: Which of the following statements is true?

For completing this study, I do NOT have a chance to earn more.

For completing this study, if I am the decision-maker, the decision I make will determine the payment I am given in cash.

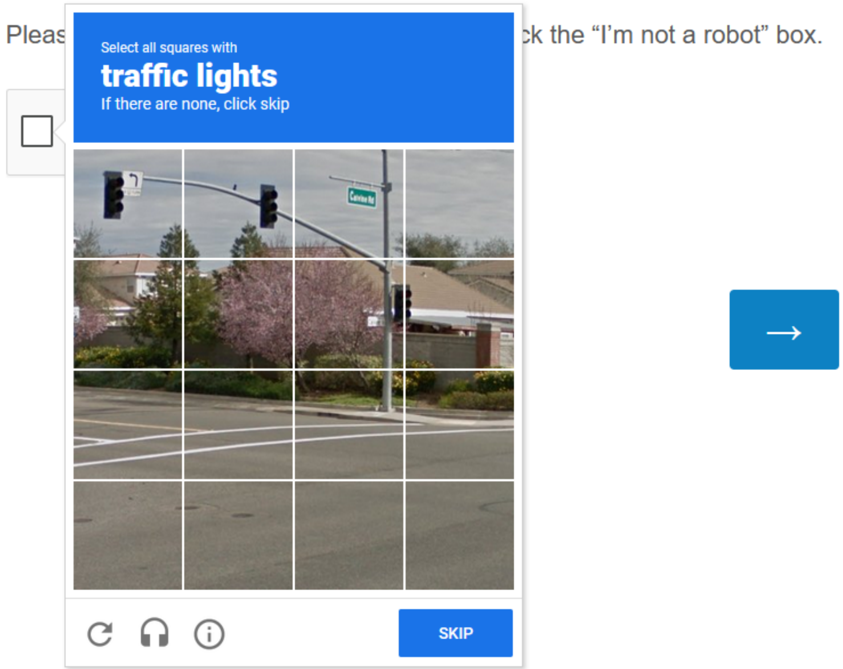
For completing this study, even if I am not the decision maker, the decision I make will determine the payment I am given in cash.

Figure F.4: ReCAPTCHA



Please verify you are a human to continue. Click the “I’m not a robot” box.

□ I'm not a robot



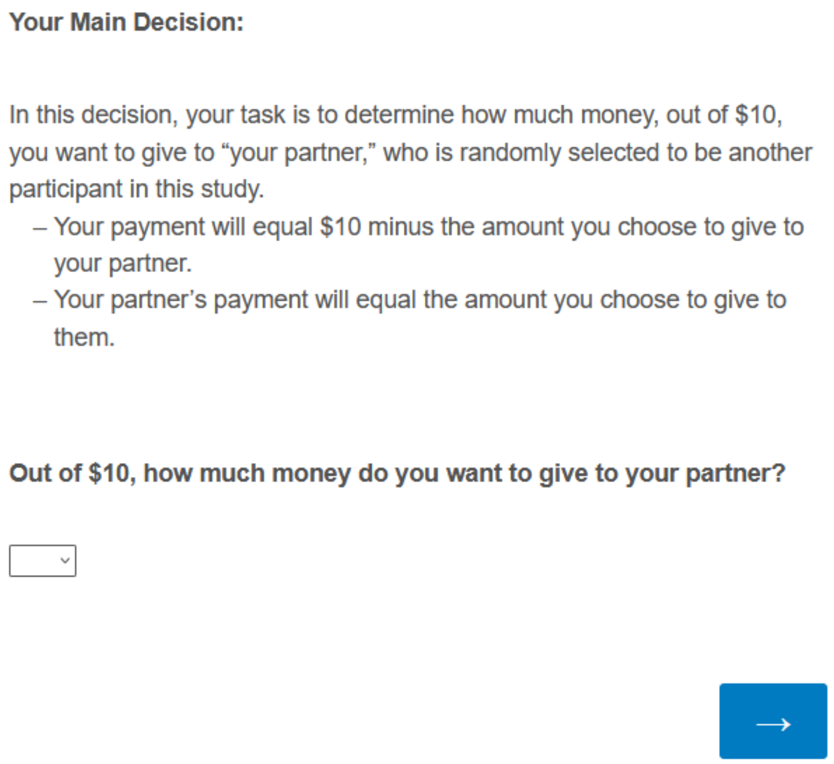
Pleas ck the “I’m not a robot” box.

Select all squares with

traffic lights

If there are none, click skip

Figure F.5: Dictator Game Decision



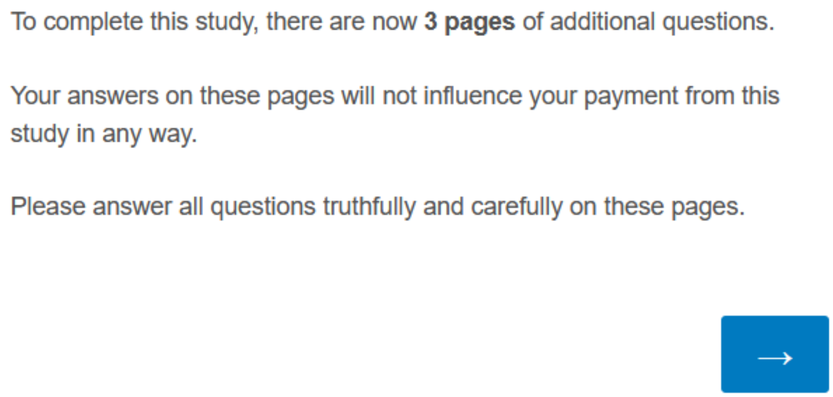
Your Main Decision:

In this decision, your task is to determine how much money, out of $10, you want to give to “your partner,” who is randomly selected to be another participant in this study.

- Your payment will equal $10 minus the amount you choose to give to your partner.
- Your partner’s payment will equal the amount you choose to give to them.

Out of $10, how much money do you want to give to your partner?

Figure F.6: Transition Page

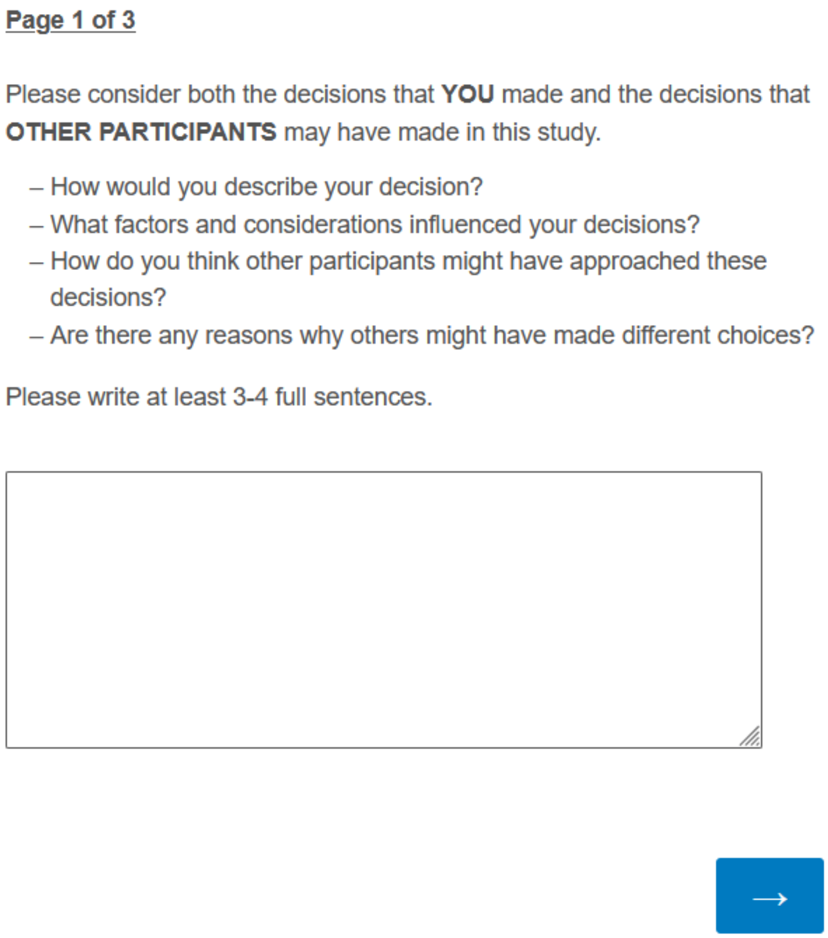


To complete this study, there are now **3 pages** of additional questions.

Your answers on these pages will not influence your payment from this study in any way.

Please answer all questions truthfully and carefully on these pages.

Figure F.7: Open-Ended Question



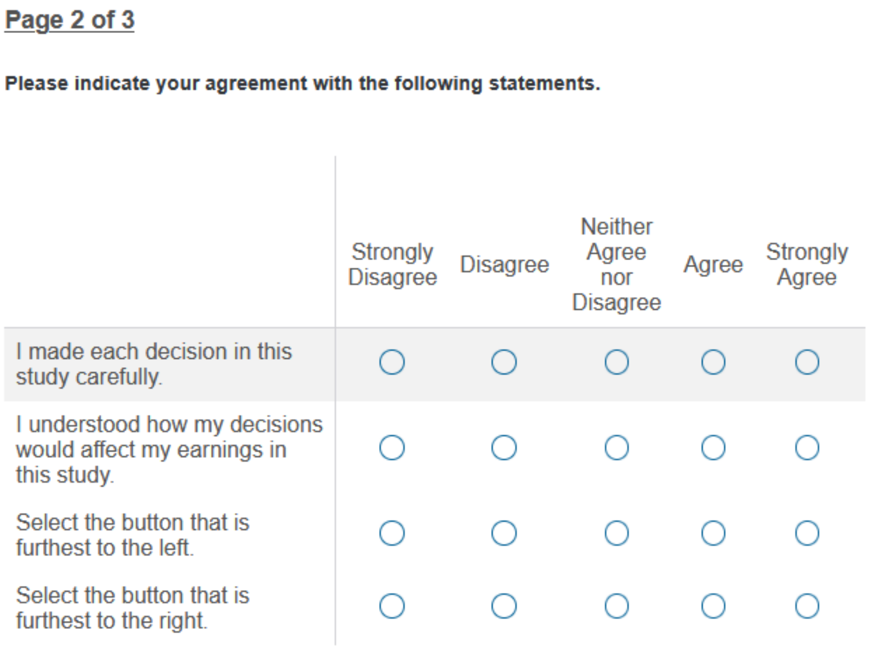
Page 1 of 3

Please consider both the decisions that **YOU** made and the decisions that **OTHER PARTICIPANTS** may have made in this study.

– How would you describe your decision?
– What factors and considerations influenced your decisions?
– How do you think other participants might have approached these decisions?
– Are there any reasons why others might have made different choices?

Please write at least 3-4 full sentences.

Figure F.8: Classic Attention Check

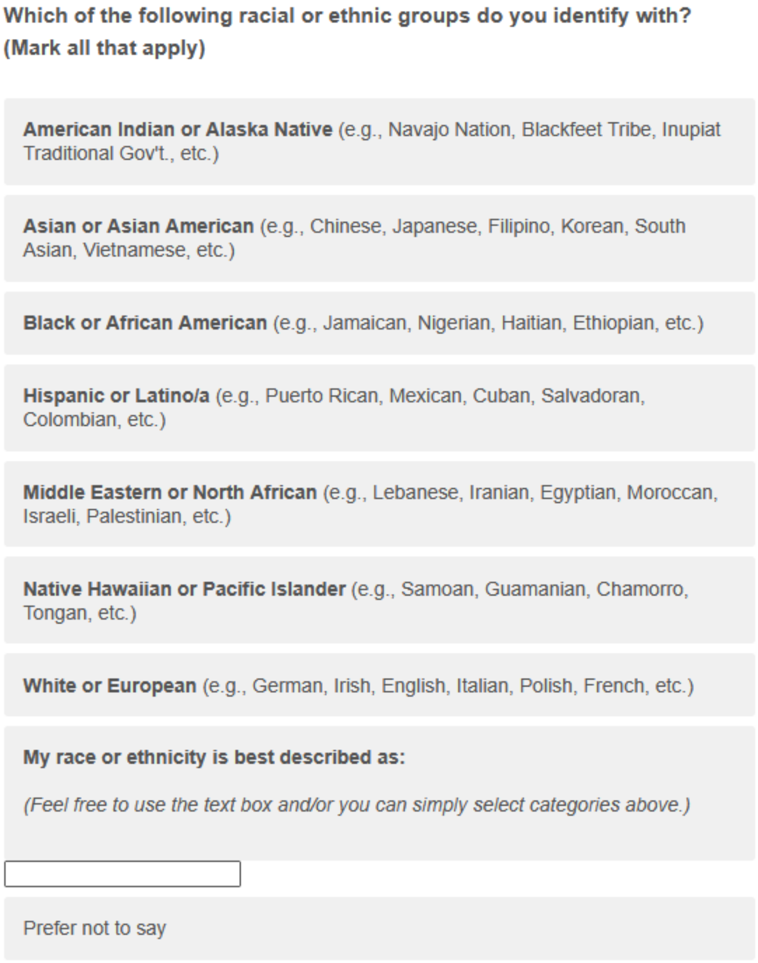


Page 2 of 3

Please indicate your agreement with the following statements.

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
I made each decision in this study carefully.	○	○	○	○	○
I understood how my decisions would affect my earnings in this study.	○	○	○	○	○
Select the button that is furthest to the left.	○	○	○	○	○
Select the button that is furthest to the right.	○	○	○	○	○

Figure F.9: Demographics I



Which of the following racial or ethnic groups do you identify with? (Mark all that apply)

American Indian or Alaska Native (e.g., Navajo Nation, Blackfeet Tribe, Inupiat Traditional Gov't., etc.)

Asian or Asian American (e.g., Chinese, Japanese, Filipino, Korean, South Asian, Vietnamese, etc.)

Black or African American (e.g., Jamaican, Nigerian, Haitian, Ethiopian, etc.)

Hispanic or Latino/a (e.g., Puerto Rican, Mexican, Cuban, Salvadoran, Colombian, etc.)

Middle Eastern or North African (e.g., Lebanese, Iranian, Egyptian, Moroccan, Israeli, Palestinian, etc.)

Native Hawaiian or Pacific Islander (e.g., Samoan, Guamanian, Chamorro, Tongan, etc.)

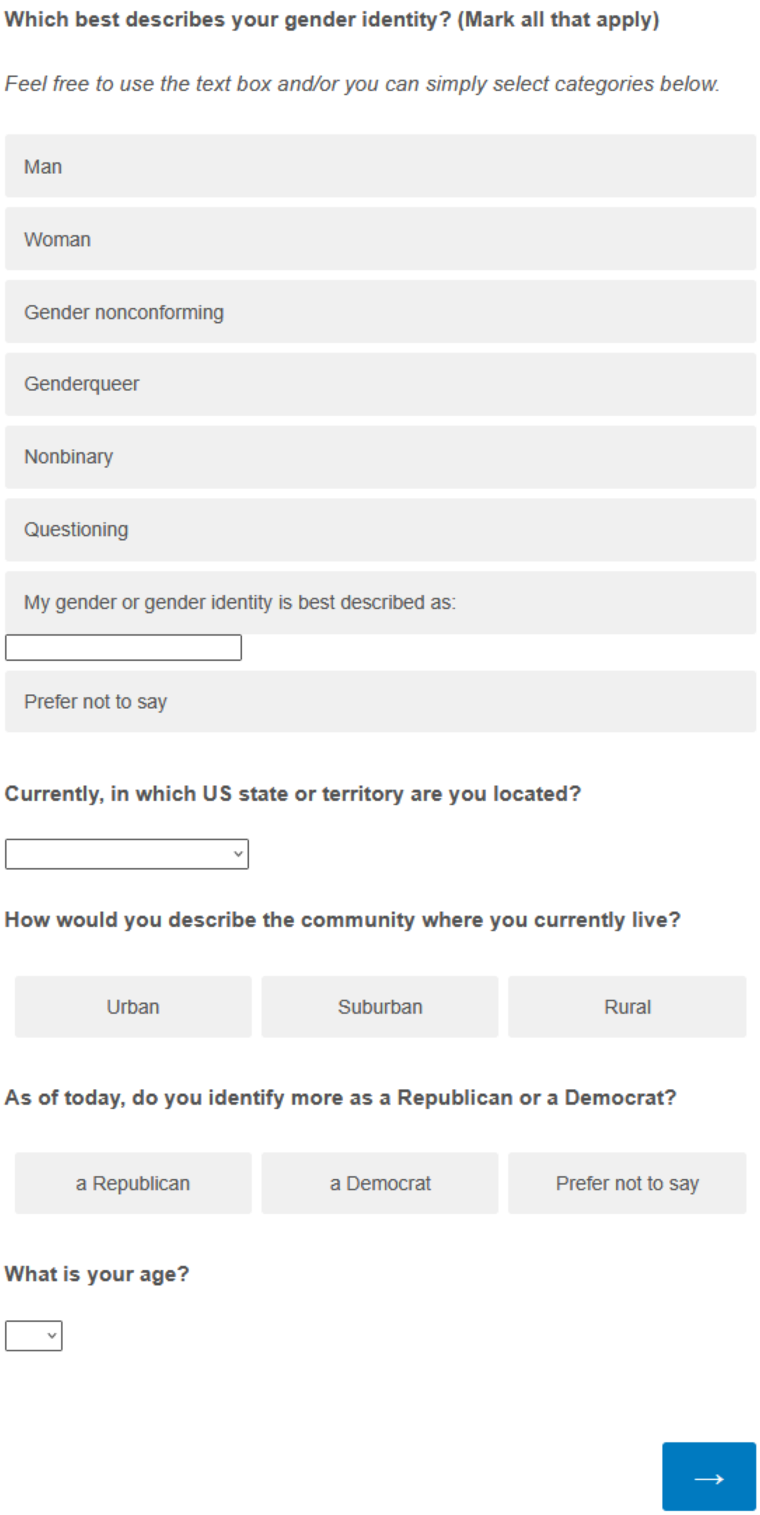
White or European (e.g., German, Irish, English, Italian, Polish, French, etc.)

My race or ethnicity is best described as:

(Feel free to use the text box and/or you can simply select categories above.)

Prefer not to say

Figure F.10: Demographics II



Which best describes your gender identity? (Mark all that apply)

Feel free to use the text box and/or you can simply select categories below.

Man

Woman

Gender nonconforming

Genderqueer

Nonbinary

Questioning

My gender or gender identity is best described as:

Prefer not to say

Currently, in which US state or territory are you located?

How would you describe the community where you currently live?

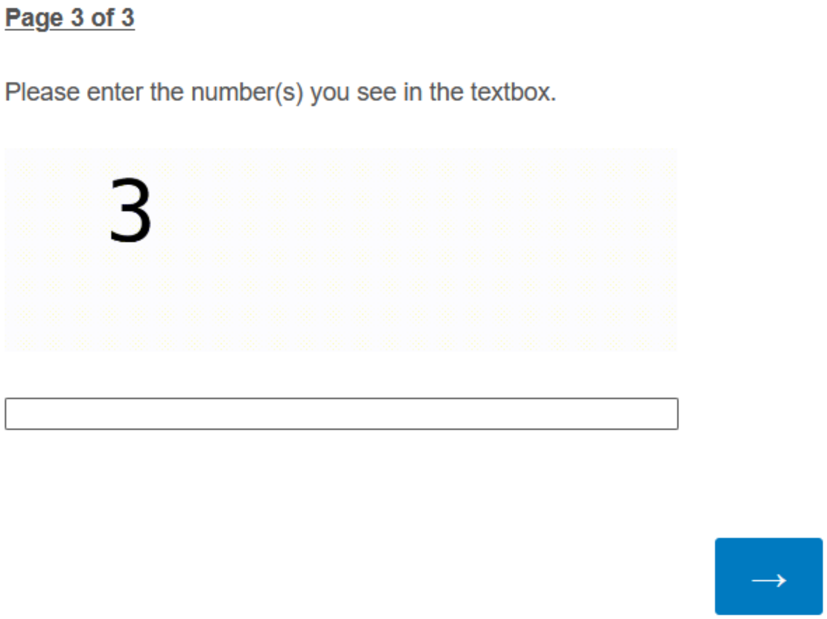
Urban | Suburban | Rural

As of today, do you identify more as a Republican or a Democrat?

a Republican | a Democrat | Prefer not to say

What is your age?

Figure F.11: Video Attention Check



Page 3 of 3

Please enter the number(s) you see in the textbox.