Mission Possible: The Collection of High-Quality Data

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

December 3, 2025

Abstract

Absent high-quality online data, research questions would be constrained conceptually and in study population. To inform the debate about online data quality, this paper provides *empirical evidence* that compares data quality of responses from online participants, AI agents, and human subjects in the lab. Corresponding results support Prolific—but not MTurk—data quality. This paper also highlights a viable path for high-quality online data in an evolving landscape: use a *two-stage recruitment method* to broadly recruit online subjects in a baseline study and then limit recruitment for the main study to the resulting subset of “high quality" subjects.

JEL Classification: C81, C83, C90, 033.

Keywords: Experiments, data quality, AI agents, 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. We provide code and materials via https://github.com/survey-data-quality-lab/mission-possible.

1 Introduction

The collection of high-quality data is essential to social science research. Absent the ability to collect high-quality data online, the scope of research questions that can be answered is severely limited, both conceptually and in study population. With respect to study population, data from online surveys and experiments often allow for larger samples at lower costs; more diverse samples in terms of geography and demographics; and the ability to reach otherwise unreachable populations since online surveys and experiments can be completed at more flexible times and in more accessible environments. Online data can also facilitate important insights into human decision-making by enabling, for example, a degree of anonymity that may not be otherwise possible.[1] Unsurprisingly, online surveys and experiments have become a common source of data on human behavior over the last decade.[2]

However, there are significant concerns about data quality from online surveys and experiments. A fundamental threat to inference and interpretation from online data is being certain that the response was provided by a single, attentive human. Concerns around inattentive subjects in online settings are commonly raised. In addition, there is a growing skepticism about whether the insights delivered by online samples are reflective of humans in the desired study population. Bots—particularly AI agents[3]—may emulate human participants (Westwood, 2025). Participants may use large language models (LLMs) to aid in their choices and provide responses (Zhang, Xu and Alvero, 2025). Participants may even engage in account fraud, for instance by using multiple accounts to participate more than once.

Threats to online data quality are thus essential to seriously consider. At the same time, before forgoing insights from online surveys and experiments—due to such threats—it is important to gather careful *empirical evidence* to make carefuly assessments of online data quality. Further, even when researchers observe low-quality data online, it is important to determine whether there are procedures that can be followed to acquire high-quality data online. The lack of careful procedures—properly targeted for the study population of interest—can certainly lead to low-quality online data just as it can lead to low-quality lab or field data.[4]

[1] For example, online experiments allow for types of anonymity—in which no research staff ever meets study participants—that may be infeasible with laboratory studies.

[2] This pattern is, for example, evident in the share of papers in AEA Journals and NBER Working Papers that mention online surveys and experiments, which has risen sharply from 2010 to 2025. Data were obtained from the *Economics Literature Search Tool* by Paul Goldsmith-Pinkham: https://paulgp.com/econlit-pipeline/. See Appendix Figure F.1 for the complete time trend.

[3] We specifically refer to AI agents that are powered by large language models (LMMs) such as the ChatGPT Agent and Perplexity Comet.

[4] For instance, it is always possible to write sufficiently confusing experimental instructions or survey questions to generate low quality data. Ensuring the instructions and questions are written in a manner that can be understood by the study population is essential, and, depending on the research question of interest, often constrains which study populations are suitable.

This paper provides an evidence-based assessment of the measurement concerns for online surveys and experiments by comparing data quality checks and behavioral responses of 314 human subjects in a controlled laboratory setting, 200 responses from AI agents (ChatGPT Agent, Perplexity Comet, and Browser Use), and approximately 2,400 online participants recruited on two widely-used platforms, MTurk and Prolific Academic. Relying on a range of data quality measures—pertaining to classic attention checks, video attention checks, typing speed and patterns in an open-ended question as well as other interaction and external validation metrics—the results in this paper support high-quality data on one platform (Prolific) but not the other (MTurk).

Beyond providing empirical evidence of online data quality, this paper proposes a new method, the *two-stage recruitment method*, to identify high-quality human responses in an evolving landscape. The first stage of data collection under this method consists of a short baseline survey, which includes five main data quality checks that are simple to implement. These checks are then used to screen high quality respondents and only invite these respondents to complete the second stage of data collection, which implements the main study of interest for the researcher.

In the baseline survey, the rates at which respondents pass the main data quality checks substantially differs when responses are from Prolific, MTurk, the lab, or AI agents. Approximately 90% of Prolific respondents satisfy all main data quality checks while less than 10% of MTurk respondents satisfy them all. This results in Prolific, but not MTurk, data quality comparing favorably to lab data quality, where we find that 80% of human subjects in the lab pass all main data quality checks. Importantly, 0% of AI agents pass all main data quality checks, highlighting their effectiveness at screening out AI agents.

In the baseline survey, the pass rates for specific types of quality checks further differs when responses are from Prolific, MTurk, the lab, or AI agents. This is evident from the rates at which respondents satisfy our main data quality checks. Our first main check is a novel video attention check, which requires respondents to correctly type in the four numbers after viewing a video that displays these numbers. This video atteniton check screens out the vast majority of AI agents but essentially no one else. The pass rates for the novel video attention check are 19% for AI agents, 96-98% for MTurk respondents, 98-99% for Prolific respondents, and 97% for lab participants. AI agents pass this video check only when asking for human assistance.

Our second main check involves a more classic multiple-choice-based attention checks. This classic attention check is less effective at screening out AI agents but appears more effective at catching human inattention.The pass rates for the classic attention checks are 98% for AI agents, 52-66% for MTurk respondents, 97-98% for Prolific respondents, and 84% for lab participants. The classic attention checks could be more effective at catching human inattention than the video attention check because they are less salient; they require respondents to correctly select the option furthest to the “left” and “right” when asked to do so in 5-point likert questions. These results make clear the importance of even “simple” attention checks that AI agents can complete.

Our third and fourth main checks leverage typing patterns and speed by asking respondents to provide an open-ended text response. This type of question can help to capture AI agents and LLM use. While 93-98% of Prolific respondents and 97% of lab participants raise no flags when we consider measures of typing patterns and speed, only 10% of AI agents and 11-22% of MTurk participants pass the main quality checks related to typing patterns and speed.

Our fifth and final main check reveals, by identifying duplicate IP addresses, considerable levels of account fraud on MTurk (>20%) but close to none on Prolific (<1%). We validate these five main data quality checks by showing how they compare to other data quality checks across all samples.

In addition to beinh informative themseleves, can these main data quality checks allow researchers to implement the two-stage recruitment method that we propose? That is, recall that the two-sage method is intended to allow researchers to (i) broadly recruit online participants for a baseline survey and then (ii) recruit from the identified pool of high-quality participants, i.e., participants who pass the data quality checks in the baseline survey, 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 that one uses for study recruitment purposes. One can also consider this approach as akin to the approach that is commonly used for field studies: have a surveyor administer a baseline survey and then only invite a subset of eligible people, given their answers in the baseline survey, to participate in the main study.

The two-stage recruitment method proves effective on Prolific. For our main study, we only recruit Prolific participants from the set who, given their responses in the baseline survey, passed all main data quality checks. This method results in high data quality for the main study: 93% of Prolific participants who complete our main study pass all data quality checks. Thus, although Prolific data quality was already very high, the two-stage recruitment method directionally increases data quality on Prolific. In contrast to Prolific, we did not complete the two-stage recruitment method on MTurk due to a lack of high-quality participants available for recruitment: only 9% of MTurk respondents pass all data quality checks in the baseline survey. Therefore, the two-stage recruitment method can make clear data quality concerns before a researcher expends more resources on their main study, as evident via our MTurk data collection, and it can also alleviate data quality concerns that may otherwise make a researcher inclined to forgo online data collection and thereby insights from online data, as evident via our Prolific data collection.

There are several features of the two-stage recruitment method that make it broadly useful and accessible to social science researchers. First, it is an intently simple method, easy to implement across a variety of platforms. Online platforms often 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.

Second, the method relies on data-quality checks that are designed to be "broad" and easy to incorporate into a range of study topics. The attention checks we use are not specific to the research question of interest (e.g., they do not ask about study-specific instructions).[5] Even the open-ended text question—which allows one to capture typing speed and patterns—is written in a way to be broad and easy to add to the end of most studies with very minimal changes.

Third, AI agents could effectively emulate human decisions, making them difficult to detect from a reliance on answers to common economic decisions or questions. In our baseline survey and in our main study, we ask respondents to answer an incentivized dictator game questions. The modal answer provided by AI agents in a dictator game is the same as that of humans in the lab.

Fourth, the two-stage method has some important strengths when compared to ex-post data quality checks. The ex-ante screening—via the first stage baseline survey—ensures that researchers only collect main study data from the set of respondents that they ex-ante deem to be high-quality. The ex-ante classification limits the scope for p-hacking and helps to prevent wasted resources that follow from low-quality data collection. In addition, ex-post screens can always be used to complement ex-ante screens.[6]

Our work contributes to a body of literature examining data quality across a variety of fields (e.g., Paolacci, Chandler and Ipeirotis (2010); Horton, Rand and Zeckhauser (2011); Goodman, Cryder and Cheema (2013); Berinsky, Margolis and Sances (2014); Berinsky et al. (2024); Thomas and Clifford (2017); Danz et al. (2021); Snowberg and Yariv (2021); Fréchette, Sarnoff and Yariv (2022); Peer et al. (2022); Ward and Meade (2023); Aksoy and Nevo (2025)). Much of this work importantly compares demographic characteristics across subject populations or examines whether similar empirical findings—in terms of answers provided to cognitive assessments or classic economic decisions—emerge across subject populations. These comparisons are essential for understanding the representativeness, cognitive abilities, and other important features of study populations. Motivated to instead focus on whether each study response reflects a single attentive human response—an important consideration for all research questions—the main contributions of our paper are: (i) to propose and empirically validate data quality checks given the new concerns related to AI agents and LLM use, and (ii) to propose and empirically validate the two-stage recruitment method, which can be adjusted to new checks over time, as AI agents and tools to detect them necessarily evolve.

In considering the tools that may prove necessary for data quality over time, we expect two parallel developments: just as AI agents may become better at mimicking humans, AI-detection software

[5] While not our focus here, we of course do believe study-specific instructions are often essential to answering a desired research question.

[6] Researchers can, for example, also pre-register how they will ex-post exclude participants from their main study. Even with pre-registration, however, researchers may waste resources on recruiting low-quality data fro their main study.

may improve its detection of these agents (e.g., Höhne et al. (2025), Imas and Jabarian (2025)). That is, while specific attention checks may not be valid for longer periods of time, improved checks may appear over 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.[7] Many survey tools allow one to easily add in questions to capture reCAPTCHA scores. In addition, shortly before the launch of our study, Prolific appeared to roll out new procedures for detecting AI agents and low-quality data.[8]

A wide range of tools are needed to answer questions in social science, including online data. Some research questions may be best answered via laboratory data, e.g., if the researcher needs more control over the decision environment or if the research question is best suited to be answered by more "normatively rational" undergraduate student populations (see, e.g., the discussion in Snowberg and Yariv (2021)). Other research questions may be best suited by field data, e.g., if choices are uninformative when there is an awareness of study participation (see, e.g., the discussion in De Quidt, Haushofer and Roth (2018)). Other questions, however, may be best answered by online surveys and experiments that allow for the researcher to control the decision environment in more ways than many field settings while still recruiting a large, diverse study population in a rather anonymous setting. More generally, a diversity of methods and data collection allow for a diversity of research ideas and innovation.

Thus, rather than viewing high-quality online data as a new "mission impossible," we hope this paper provides a viable pathway forward for the continued collection of high-quality online data. At the same time, we emphasize that close attention to data quality is warranted—when collecting data from online settings, laboratory settings, field settings, and representative samples.[9]

2 Experimental Design

In this section, we first describe the survey instrument used for all samples.[10] 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

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

[8] Indeed, our project was originally motivated by a much-discussed drop in data quality on Prolific during the earlier part of 2025. While it is reassuring this data quality issue did not appear when we turned to design and run this study in the Summer and Fall of 2025, we note that this could be reflective of what appear to be, at least to the public, new procedures about how they screen for AI-generated responses and data quality checks. See https://www.prolific.com/protocol-data-quality.

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

[10] See Appendix G for screenshots of the baseline survey instructions.

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 platforms, 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 for the baseline survey and the main survey. Since individuals or AI agents may complete the survey, we often refer to each submission as being provided by a "respondent."

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, and they complete a reCAPTCHA verification challenge.

The baseline and main surveys then proceed 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 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 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.[11]

2.2 Data Quality Measures

Next, we describe all the data quality measures we collect with the baseline survey. To apply the two-stage recruitment method—that is, to create a pool of high-quality respondents—we use a subset of these measures, which we describe in Section 2.3.3. The JavaScript code used within the survey to collect these measures, and a link to a permanent repository is provided in Appendix D.

Two data quality checks that are based on the answers respondents provide to the questions on the survey instrument are:

- *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 5-point likert questions in the follow-up questionnaire.
- *Passed video check* is an indicator variable equal to 1 if respondents correctly type four numbers that are shown sequentially in a short video.[12]

With the open-ended survey question, we use JavaScript to record keystrokes and input events with timestamps, tracking how respondents enter text into the text box:[13]

- *Typed text* is an indicator for whether the respondent entered the open-text response through manual typing. It takes the value 1 if three conditions are jointly satisfied: (i) no paste event is recorded; (ii) there is no large discrete increase in text length without corresponding

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

[12] You can watch the video here.

[13] We present representative tracker data in Appendix E.2 to illustrate how AI agents and humans exhibit distinct text entry behavior. We also show how typing speed is calculated and interaction events are recorded, providing several examples of these measures.

keystrokes (i.e., no input jump event of more than 50 characters); and (iii) at least one keystroke is recorded. Together, these conditions are designed to detect the most common forms of non-typed text input, including copy–paste (i and ii), drag-and-drop (ii), and fully automated insertion (i,ii and iii).[14]

- *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.[15]

On the questionnaire page, we track mouse movements and clicks with two indicator variables:

- *Mouse clicks >0* is an indicator variable that takes value 1 if the respondent clicked on the screen at least once.[16]
- *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.[17]

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, ≥0.9 or ≥0.5* is an indicator variable equal to 1 if Google's reCAPTCHA score is equal to 1, or higher than 0.9 or 0.5, 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.[18]
- *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.[19]

[14] See Appendix E.2 to E.4 for details about keystroke data and input events. *Typed text* produces highly consistent classifications with an indicator variable taking the value of 1 if the number of keystrokes is larger than the number of characters of the submitted text (>97.5% consistency), see Appendix E.6. *Typed text* also produces highly consistent classifications with an indicator variable equal to 1 if the submitted open text response is sufficiently similar to a reconstructed response based on recorded keystrokes (>98% consistency), see Appendix E.6.

[15] 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 AI agents frequently type faster than 75 milliseconds.

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

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

[18] See Appendix B.1 for a more detailed discussion of reCAPTCHA scores. ReCAPTCHA challenges, in contrast, use visual or audio puzzles that users must solve to prove they are human.

[19] See Appendix B.3 for a more detailed discussion of Pangram scores. Appendix Figure A.5 shows the distribution of Pangram scores for the different samples.

To identify potential fraudulent accounts, we use IP addresses, geolocations, and device fingerprints for the online samples (Prolific and MTurk). For our AI agent and lab samples, these measures are not applicable and therefore recoded as missing, as shared IP addresses and device fingerprints are not indicative of account fraud in these samples.[20]

- *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 located inside the United States (provided by Maxmind.com).[21]
- *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.
- *No duplicate 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 using cookies.
- *No duplicate device fingerprint* is an indicator variable that takes value 1 if the submission has no duplicate device fingerprint (provided by Fingerprint.com). Device fingerprinting assigns a unique and persistent identifier to a device (visitor ID). This device fingerprint allows to track devices even when IPs change (e.g. through VPNs) or cookies are deleted.[22]

2.3 Samples

We conducted the experiment in the lab, with undergraduate students, on two widely used platforms, Prolific and MTurk, and with AI agents. 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 24 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

[20] AI agents may in general not share the same IP addresses or device fingerprints if they are run on different devices and networks. Thus, we do not use these measures to identify AI agents in our data.

[21] See Appendix B.2 for more details.

[22] Device fingerprints are missing for 14.2% of the MTurk sample and 14.6% of the Prolific sample. These observations have no duplicate device fingerprint. See Appendix B.4 for details.

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. The students' dictator game decision was still incentivized as described.

2.3.2 AI Agent 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 (August to October 2025): ChatGPT Agent, Perplexity Comet, and BrowserUse. ChatGPT Agent and Perplexity Comet are leading commercial products, while BrowserUse is a very popular open-source web automation framework. For BrowserUse, we implemented agents using OpenAI's O3 model and Google's Gemini 2.5 Flash.

We instructed the AI agents to complete the survey under two prompt conditions. The *Simple* prompt provided the survey link and directed the agent to act as a human participant. The *Complex* prompt added instructions on persona profile and answer behavior to imitate human responses. This design allows us to test the sensitivity of our checks to prompt complexity.

AI agents complete the survey in 85% of cases, though some agents requested assistance from the researcher to solve the reCAPTCHA challenge or the video attention check. We recorded every instance in which an agent requested such human assistance. The AI agent sample consists of those 200 AI agents that completed the entire survey. Comparing this sample with the lab sample provides an empirical benchmark to evaluate which measures effectively distinguish human from AI-generated responses.[23]

2.3.3 Prolific Sample

We recruited 1200 participants on Prolific Academic. Specifically, we recruited 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 higher data quality. In the Prolific (*All*) treatment, participants could be any Prolific participant who is based in the US, with any approval rating and any prior experience. In the Prolific *T95*

[23] See Appendix C for more details on the implementation of the AI agent data collection, a detailled discussion of the prompts, details on completion rates, and technical background information about AI agents.

treatment, we maintain the requirement to be based in the US and 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 the *T99* treatment, we maintain the requirement to be based in the US and introduce more stringent requirements, requiring that participants have at least a 99% approval rating and more than 1000 prior submissions.

For the Prolific two-stage treatment (2S), we only recruit from a pool of high-quality participants who are identified as high-quality given their responses to the baseline survey in the T(All), T(95), or T(99) treatment. We specifically label a participant as high-quality if they pass all main data quality checks. We select main data quality checks that: (i) allow us to effectively screen out all AI agents and (ii) allow us to capture human inattention as well. For simplicity, we do not include data quality checks that are effectively redundant with each other, but we do show the robustness of our results to all data quality checks.[24] This results in five main data quality checks, which require that:

- Respondents correctly answer the two attention checks, satisfying the conditions for *Passed classic checks* and *Passed video check*,
- Respondents have typed text at typical speeds, measured by *Typed text* and *Typed with typical speed*, and
- Respondents 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 around 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.[25]

Data quality was extremely low on MTurk: only 9% of respondents passed the main data quality checks that would make them eligible for the two-stage treatment. This resulted in only 79 of 881

[24] There are only two measures that allow us to capture human inattention: the classic and video attention checks. Table 2 shows the non-main data quality checks. We did not include any of these additional non-main data quality checks as main data quality checks because, as shown in Column 8, this would classify little-to-no additional data as low quality in the two-stage recruitment method. The most notable exception occurs with *ReCAPTCHA score = 1*, but we emphasize that industry practice is to set a threshold of 0.50 not 1.

[25] From our MTurk sample, we exclude 16 of 897 observations that are repeated submissions from the same MTurk IDs, leading to our final MTurk sample size of 881 observations.

respondents being part of the eligible recruitment pool for 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 describe additional data quality checks. Finally, we describe the economic decisions and sociodemographic characteristics of each sample.

3.1 Main Data Quality Checks

Table 1 shows the fraction of respondents that passed the main data quality checks simultaneously and individually, by sample. Figure 1 provides a graphical representation of these results when pooling all Prolific treatments and pooling all MTurk treatments, since there are little within-platform differences across treatments.

Column (1) of Table 1 confirms that our main data quality checks succeed at screening out AI agents: 0% of AI agents pass all checks.[26] While Columns (2)–(4) reveal broadly low data quality on MTurk with only 7-11% of MTurk participants passing all main data quality checks, Columns (5)–(8) reveal consistently high data quality on Prolific with 89-93% of Prolific participants passing all main data quality checks. Column (9) then shows that Prolific data quality—in addition to comparing very favorably to MTurk data quality—compares quite favorably to lab data. Only 80% of lab participants pass all main data quality checks.

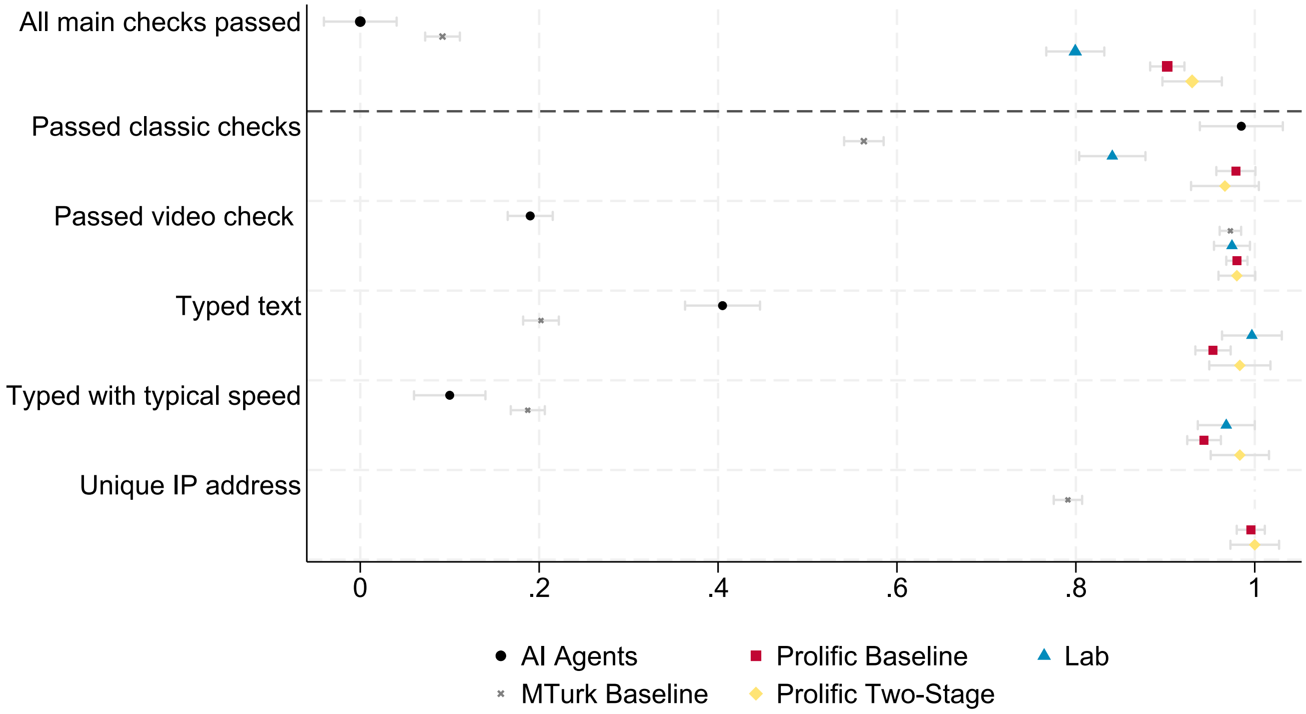
How do the data quality measures contribute to these differences across samples? Several patterns become immediately evident.

First, while the classic attention checks involving multiple choice questions do little to screen out AI agents (98% of AI agents pass these check), it still seems potentially quite important at capturing human inattention. Only 52-66% of MTurk participants pass the classic attention checks, and while much higher, only 84% of lab participants pass the classic attention checks. By contrast, 97-98% of Prolific participants pass these attention checks.

Second, there are some attention checks that—as of now—are rather effective at screening out AI agents. Only 19% of AI agents pass the novel video attention check (and only those AI agents pass that ask for human assistance). By contrast, 96-98% of MTurk, Prolific, and lab participants pass the video attention check. That the video check is passed at similar rates in the online surveys as in the lab may indicate very low usage rates of AI agents on MTurk and Prolific - at least of AI

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

Figure 1: Data Quality - Main Checks



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

Table 1: Main Data Quality Checks

	AI	**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 AI agent responses. 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 passing all data quality checks. The remaining variables show the fraction passing the noted check (see Section 2.2 for specific definitions.)

agents that would typically fail the video attention check (e.g. ChatGPT Agent and BrowserUse).[27] That the video attention checks are passed at a higher rate than the classic attention checks among MTurk and lab participants is notable. It may reflect the video attention checks being more salient and hence still point to the benefit of less salient, simple attention checks for capturing human inattention—while using other checks to capture AI agents.

Third, typing patterns and speed—e.g., because text appears to be pasted into the open-text response question or otherwise typed too quickly—appear to be the most binding data quality measures. Only 10% of AI agents pass these checks, and only 11-23% of MTurk participants pass these checks. One concern with these measures, however, could be that they are “too restrictive." For instance, a participants in an online study could type a free response question in a word document and then paste their answer into the survey. This would result in their answer being flagged by these measures. Thus, while we think caution with interpreting these measures is informative, we choose to keep them in our main data quality checks because both Prolific participants and lab participants in our sample largely succeed at passing these checks.

Fourth, while examining participant's IP addresses is reassuring for Prolific participants (with 99-100% of participants having unique IP addresses), it raises some account fraud concerns on MTurk since only 74-83% have unique IP addresses on MTurk. Note that we do not present unique IP address results for the AI agent sample or for the lab sample, since these responses have the same IP addresses by design.

3.2 Robustness Measures of Data Quality

Table 2 shows additional data quality measures.[28] We obtain three main insights. First, using these additional measures does not improve the detection of AI agents, 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 Prolific participants recruited via our two-stage recruitment method pass additional data quality checks relating to mouse clicks and mouse movements. Results from a reCAPTCHA challenge also reveal that 100% of these Prolific participants pass the challenge and have a reCAPTCHA score than exceeds the recommended threshold of 0.50. Even 98% or 93% of these Prolific participants have reCAPTCHA

[27] As the ChatGPT Agent is one of the most widely known and commercially available AI agents, an absence of ChatGPT Agents may be indicative of very low usage rates of AI agents in general. When analyzing additional information about the IP addresses of our AI agents, we observe that all ChatGPT Agents originated from IPs associated with the company Cloudflare — suggesting that ChatGPT Agents are hosted on Cloudflare's server infrastructure. No response on Prolific — and only one response on MTurk — has an IP address associated with Cloudflare, suggesting very low usage rates of the ChatGPT Agent (see Appendix B.2 for details).

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

scores that exceed much higher thresholds of 0.9 or 1, respectively.[29] In addition, 98% of these Prolific participants have a Pangram AI likelihood that suggests LLM was unlikely to generate the response to the open-ended question.[30]

To provide a plausible upper-bound estimate of the share of respondents using either (i) one of our three AI agents to complete the survey, or (ii) an LLM to generate a response for the open-text question, we can use the combination of *Passed video check* and *Typed with typical speed.* The combination of these two checks detects 100% of AI agents in our sample, and detects when text produced by an LLM was copy-pasted or drag-and-dropped into the text box. This estimate is 82.3% for MTurk, 7.7% for Prolific in the baseline survey, and 3.7% for Prolific in the two-stage procedure.[31]

Appendix Table A.1 provides further insight in terms of how respondents, across all samples, made decisions in the dictator game, their demographic information, and response times.[32] Decisions in the dictator game made by AI agents are mainly aligned with those on Prolific and in the lab, with AI agents choosing to give $5 in 54% of the cases, compared to Prolific participants choosing this amount between 54% and 56% of the time and lab participants choosing it 71% of the time. The median response time of AI agents (6.3 minutes) is slightly higher than in the lab (5.6 minutes), on Prolific (4.5 minutes), and MTurk (3.6 minutes), but is within a plausible range for a 5-minute survey.[33] In terms of age, gender identity, race, political affiliation, and rural compared to (sub)urban differences, there are no systematic differences between AI agents and the online and lab samples. The age distribution reported by AI agents is similar to that of Prolific, while—as expected—lab participants are much younger (97% between 18 and 25 years old). Moreover, when AI agents are assigned a persona profile in the complex prompt condition (age, gender, state, political affiliation), they enter it correctly in 69% of cases. Overall, the behavioral decisions, demographic characteristics, and total response times of AI agents are not clearly distinct from those of other samples, indicating that AI agent responses would appear plausibly human based

[29] Note that the threshold of 1 is only met by 69% of human participants in the lab and hence inaccurately flags some (clearly) human behavior.

[30] The shares of respondents that pass all main checks and all additional checks shown in Table 2 (using the standard 0.5 cutoff for ReCAPTCHA and Pangram) are 0% for AI agents, 4.4% for MTurk, 87.4% for Prolific in the baseline surveys, 91.4% for Prolific in the two-stage procedure, and 79.9% for the lab.

[31] On Prolific, we also implemented Prolific's new authenticity check (see https://prolific-researcher.dixa.help/en/article/6bb6d8). In the baseline survey, only 21 responses (2.3%) are flagged by Prolific as having submitted inauthentic responses in our open-text question. This is a lower share than our AI usage estimate (7.7%), the share of respondents who have not *Typed with typical speed* (5.7%), and those who have no *Pangram AI likelihood < 0.5* (3.1%).

[32] In the Appendix, we show the distribution of the dictator game decisions in Figure A.6 and response times in Figure A.7 for the different samples. Appendix Table A.4 shows the demographic information separately for each AI agent model and prompt type.

[33] The observation that AI agents are, on average, slower than humans may run against one's prior expectation. In fact, AI agents are much faster than humans in the lab on some questions (study overview page, open text response), but much slower on others (reCAPTCHA challenge, demographics, video check) (see Appendix Figures A.7 and A.8 for details). Given that some AI agents are slow in completing the reCAPTCHA challenge, and some fail it, the challenge can be useful to detect and deter AI agents and not only traditional bots.

Table 2: Additional Data Quality Checks

	AI	**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 ≥ 0.5	0.88	0.75	0.76	0.94	0.97	0.98	0.99	1.00	1.00
ReCAPTCHA score ≥ 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	—	0.95	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	—
No duplicate submission	—	0.91	0.93	0.94	1.00	1.00	1.00	1.00	—
No duplicate device fingerprint	—	0.50	0.51	0.35	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 AI agent responses. 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 passing 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).

only on these measures.

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 AI agents that cannot (yet) assess video content. The typing patterns serve to detect text stemming from LLMs or AI agents. Finally, IP address checks are helpful in detecting potentially fraudulent submissions on online platforms.

When considering these main 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 AI agents 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 in a baseline survey to complete the main study. 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 online surveys and experiments. Even as AI evolves in the ways it can assist (or impersonate) humans, we hope the two-stage recruitment method allows the continued collection of high online data quality. In addition, we hope this paper encourages researchers to carefully and empirical assess their data quality for data collected from any number of sources, including field data, data from representative samples, and data from well-established sources. Different data sources are needed to answer important questions in social science research. Close attention to the data quality of these sources—which may confirm or oppose a researcher's prior and which may change over time—is essential.

References

Aksoy, Billur, and Saggi Nevo. 2025. "Unlocking Insights into Prolific: Research Implications, Participant Behavior and Motivations." *Participant Behavior and Motivations (March 21, 2025).*

Berinsky, Adam J, Alejandro Frydman, Michele F Margolis, Michael W Sances, and Diana Camilla Valerio. 2024. "Measuring attentiveness in self-administered surveys." *Public Opinion Quarterly*, 88(1): 214–241.

Berinsky, Adam J, Michele F Margolis, and Michael W Sances. 2014. "Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys." *American Journal of Political Science*, 58(3): 739–753.

Çelebi, Can, and Stefan P. Penczynski. 2024. "Using Large Language Models for Communication Classification." Centre for Behavioural and Experimental Social Science, University of East Anglia. Working Paper Series 24-01.

Danz, David, Neeraja Gupta, Marissa Lepper, Lise Vesterlund, and K Pun Winichakul. 2021. "Going virtual: A step-by-step guide to taking the in-person experimental lab online." *Available at SSRN 3931028.*

De Quidt, Jonathan, Johannes Haushofer, and Christopher Roth. 2018. "Measuring and bounding experimenter demand." *American Economic Review*, 108(11): 3266–3302.

Fréchette, Guillaume R, Kim Sarnoff, and Leeat Yariv. 2022. "Experimental economics: Past and future." *Annual Review of Economics*, 14(1): 777–794.

Goodman, Joseph K, Cynthia E Cryder, and Amar Cheema. 2013. "Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples." *Journal of Behavioral Decision Making*, 26(3): 213–224.

He, Hongliang, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. 2024. "Webvoyager: Building an end-to-end web agent with large multimodal models." *arXiv preprint arXiv:2401.13919.*

Höhne, Jan Karem, Joshua Claassen, Saijal Shahania, and David Broneske. 2025. "Bots in web survey interviews: A showcase." *International Journal of Market Research*, 67(1): 3–12.

Horton, John J, David G Rand, and Richard J Zeckhauser. 2011. "The online laboratory: Conducting experiments in a real labor market." *Experimental Economics*, 14(3): 399–425.

Imas, Alex, and B. Jabarian. 2025. "Artificial Writing and Automated Detection." *National Bureau of Economic Research Working Paper No. w34223.*

Lampinen, Andrew K, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. "Can language models learn from explanations in context?" *arXiv preprint arXiv:2204.02329.*

Mishra, Swaroop, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021*a*. "Cross-task generalization via natural language crowdsourcing instructions." *arXiv preprint arXiv:2104.08773*.

Mishra, Swaroop, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2021*b*. "Reframing Instructional Prompts to GPTk's Language." *arXiv preprint arXiv:2109.07830*.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. "Running experiments on amazon mechanical turk." *Judgment and Decision Making*, 5(5): 411–419.

Peer, Eyal, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2022. "Data quality of platforms and panels for online behavioral research." *Behavior Research Methods*, 54(4): 1643–1662.

Sclar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting." *arXiv preprint arXiv:2310.11324*.

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

Thomas, Kyle A, and Scott Clifford. 2017. "Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments." *Computers in Human Behavior*, 77: 184–197.

Ward, Mary K, and Adam W Meade. 2023. "Dealing with careless responding in survey data: Prevention, identification, and recommended best practices." *Annual Review of Psychology*, 74(1): 577–596.

Westwood, Sean J. 2025. "The potential existential threat of large language models to online survey research." *Proceedings of the National Academy of Sciences*, 122(47): e2518075122.

White, Jules, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. "A prompt pattern catalog to enhance prompt engineering with chatgpt." *arXiv preprint arXiv:2302.11382*.

Zhang, Simone, Janet Xu, and AJ Alvero. 2025. "Generative ai meets open-ended survey responses: Research participant use of ai and homogenization." *Sociological Methods & Research*, 00491241251327130.

Mission Possible: Data Quality in Online Survey

December 3, 2025

(For Online Publication Only)

Online Appendix

A Additional Results **2**

B External Data Quality Measures **16**

B.1 Google reCAPTCHA Score . . . 16

B.2 MaxMind IP Information . . . 16

B.3 Pangram AI Detection . . . 17

B.4 Fingerprint Device Identification . . . 17

C AI Agent Data Collection **19**

C.1 Implementation and Procedures . . . 19

C.2 Prompts . . . 22

C.3 Survey Completion Diagnostics . . . 25

C.4 Technical Background Information on AI Agents . . . 27

D JavaScript Tracking Code **28**

D.1 General Tracking Code . . . 28

D.2 Keylog Tracking Code . . . 32

D.3 Device Fingerprinting Code . . . 34

E Tracker Data **36**

E.1 General Tracker Data . . . 36

E.2 Keylog Data from Open-Ended Questions . . . 37

E.3 Empty Keylogs . . . 39

E.4 Paste and Input Jump Events . . . 40

E.5 Typing Speed . . . 41

E.6 Validation of Text Based Indicators . . . 42

F The Rise of Online Surveys and Experiments **48**

G Survey Instructions **49**

A Additional Results

Table A.1: Demographics Data

	AI	MTurk			Prolific				Lab
		All	T95	T99	All	T95	T99	2S	
	(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 AI agents. 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 (2S), respectively. Column (9) presents lab respondents.

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

	AI	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)
All main checks passed	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Passed classic checks	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 AI agent responses. Columns (2)–(3) present responses from the ChatGPT 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. Unique IP address is not applicable and coded as missing for AI agents because they have shared IP addresses by design of our data collection. See Section 2.2 for specific variable definitions.

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

	AI	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)
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 ≥ 0.5	0.88	0.95	1.00	0.95	0.90	0.95	1.00	0.10	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
Text similarity score > 0.2	0.40	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00
US IP address	—	—	—	—	—	—	—	—	—
Not in a geolocation cluster	—	—	—	—	—	—	—	—	—
No duplicate submission	—	—	—	—	—	—	—	—	—
No duplicate device fingerprint	—	—	—	—	—	—	—	—	—

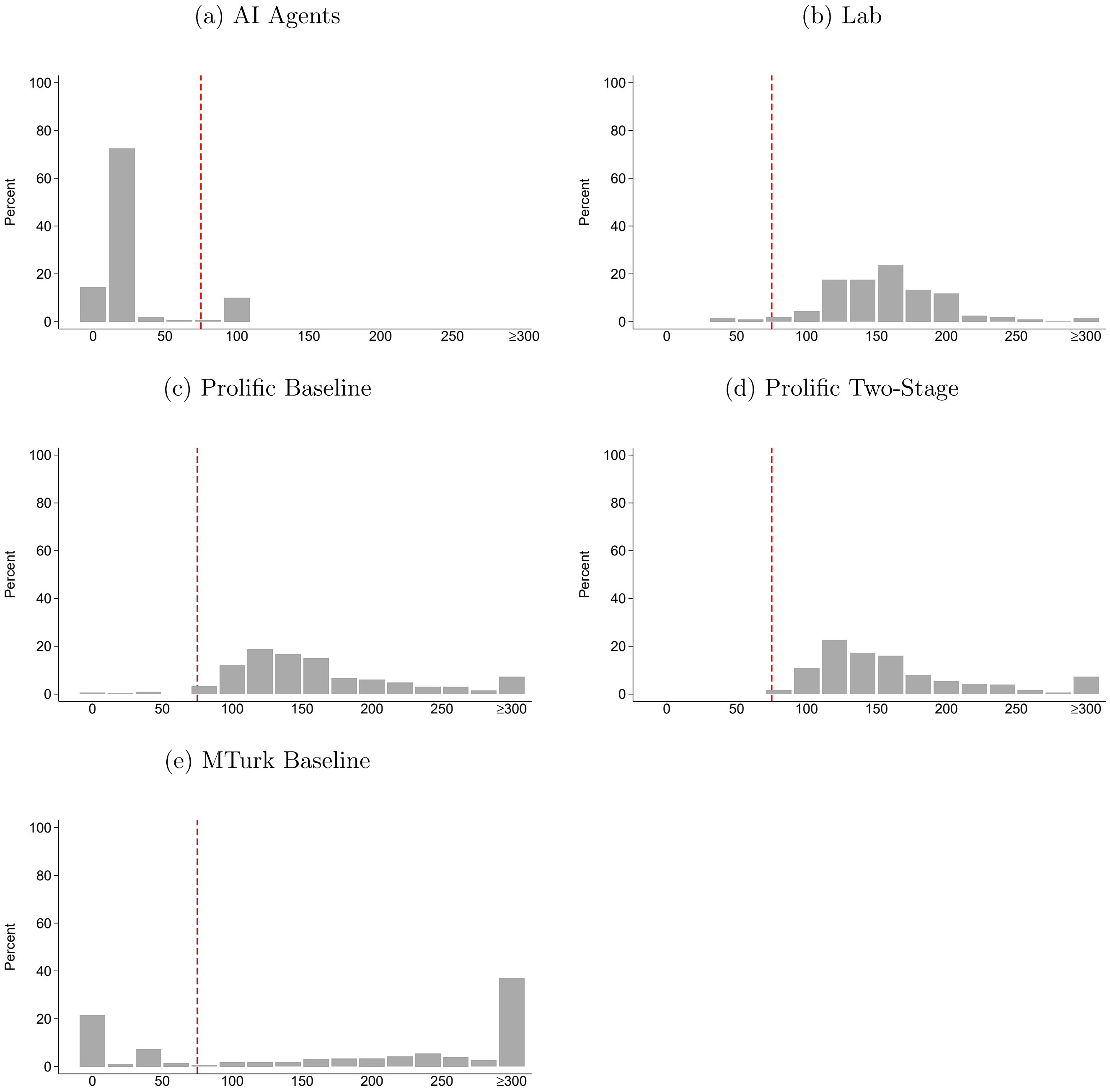
This table presents results for survey responses generated by different AI agents and prompt types. Column (1) presents all AI agent responses. Columns (2)–(3) present responses from the ChatGPT agent under the simple (-S) and complex (-C) prompt conditions, respectively. Columns (4)–(5) present responses 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. Variables marked with (—) are not applicable and coded as missing for AI agents because they have these variables by design of our data collection. See Section 2.2 for specific variable definitions.

Table A.4: AI Agents: Demographics Data

	AI	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 AI agent responses. Columns (2)–(3) present responses from the ChatGPT 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 (in milliseconds)



(a) AI Agents

(b) Lab

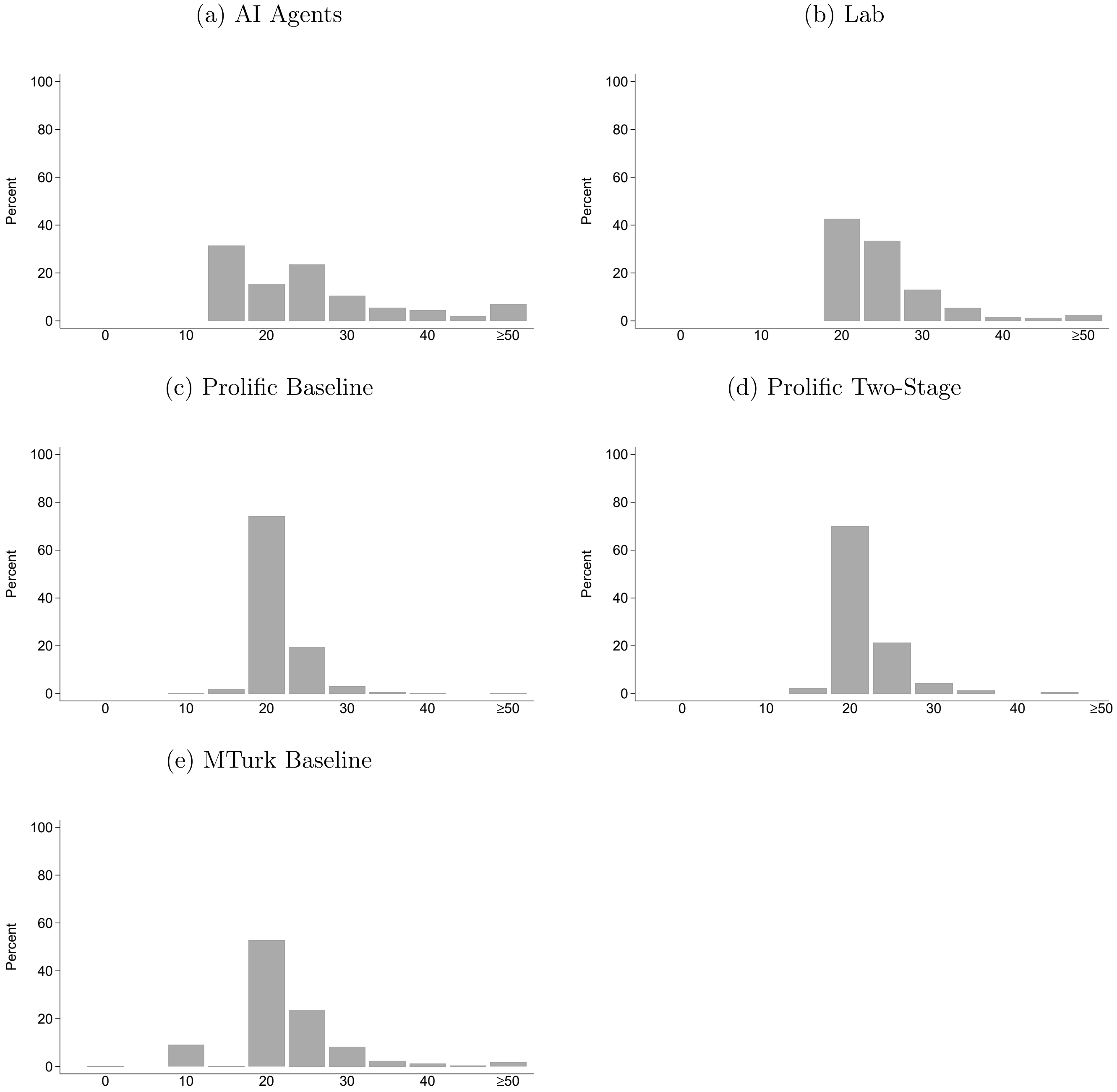
(c) Prolific Baseline

(d) Prolific Two-Stage

(e) MTurk Baseline

Notes: This figure presents histograms of typing speed (in milliseconds) by sample. Each subfigure corresponds to a different sample. Observations with less than two keystrokes have a typingspeed of 0. The vertical dashed line at 75 milliseconds indicates the cutoff for typical typing speeds. See Appendix section E.5 for more details on typingspeed.

Figure A.2: Mouse Clicks


(a) AI Agents

(b) Lab

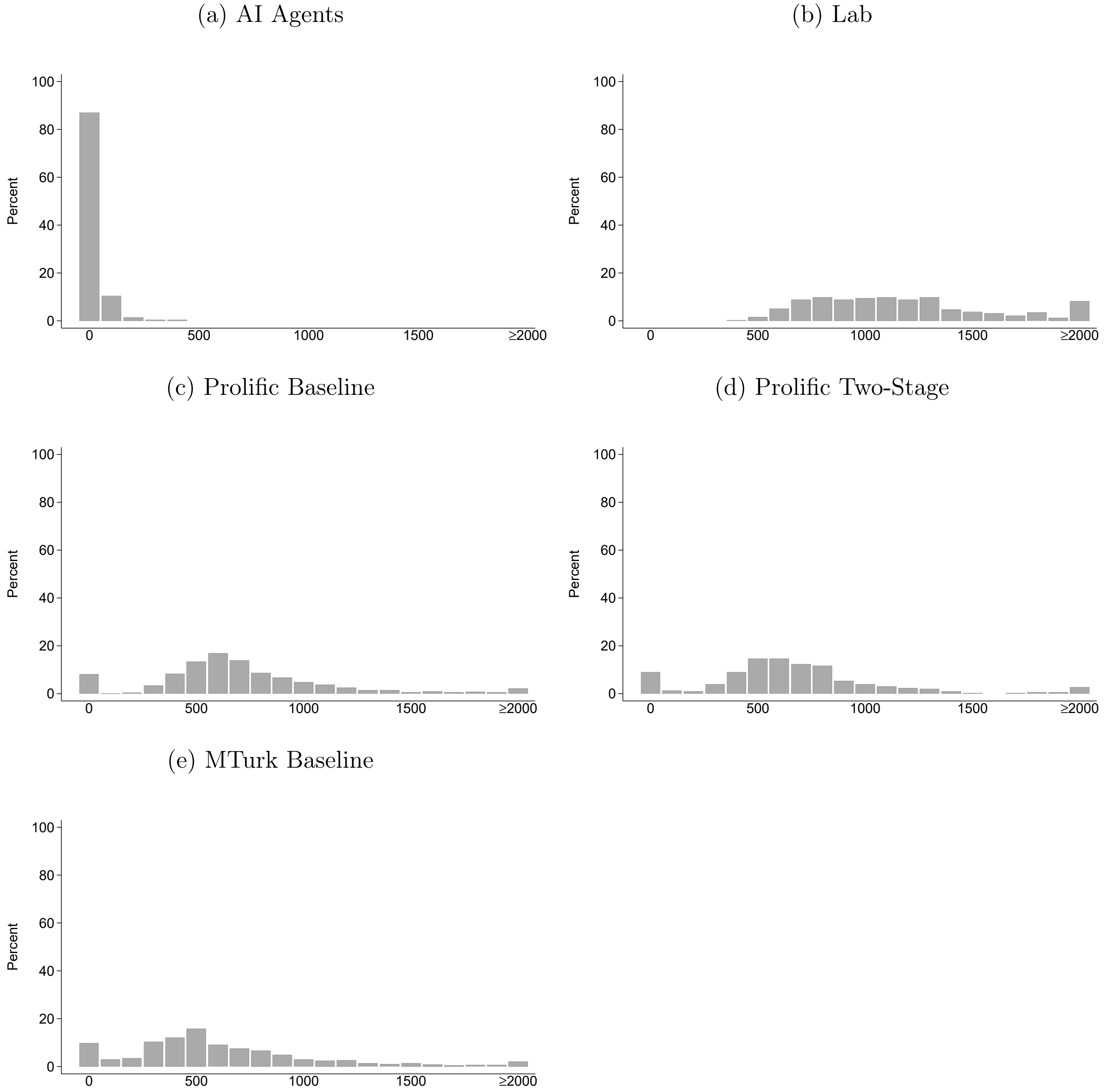
(c) Prolific Baseline

(d) Prolific Two-Stage

(e) MTurk Baseline

Notes: This figure presents histograms of mouse clicks on the demographics page (questionnaire) by sample. Each subfigure corresponds to a different sample.

Figure A.3: Mouse Movements



(a) AI Agents

(b) Lab

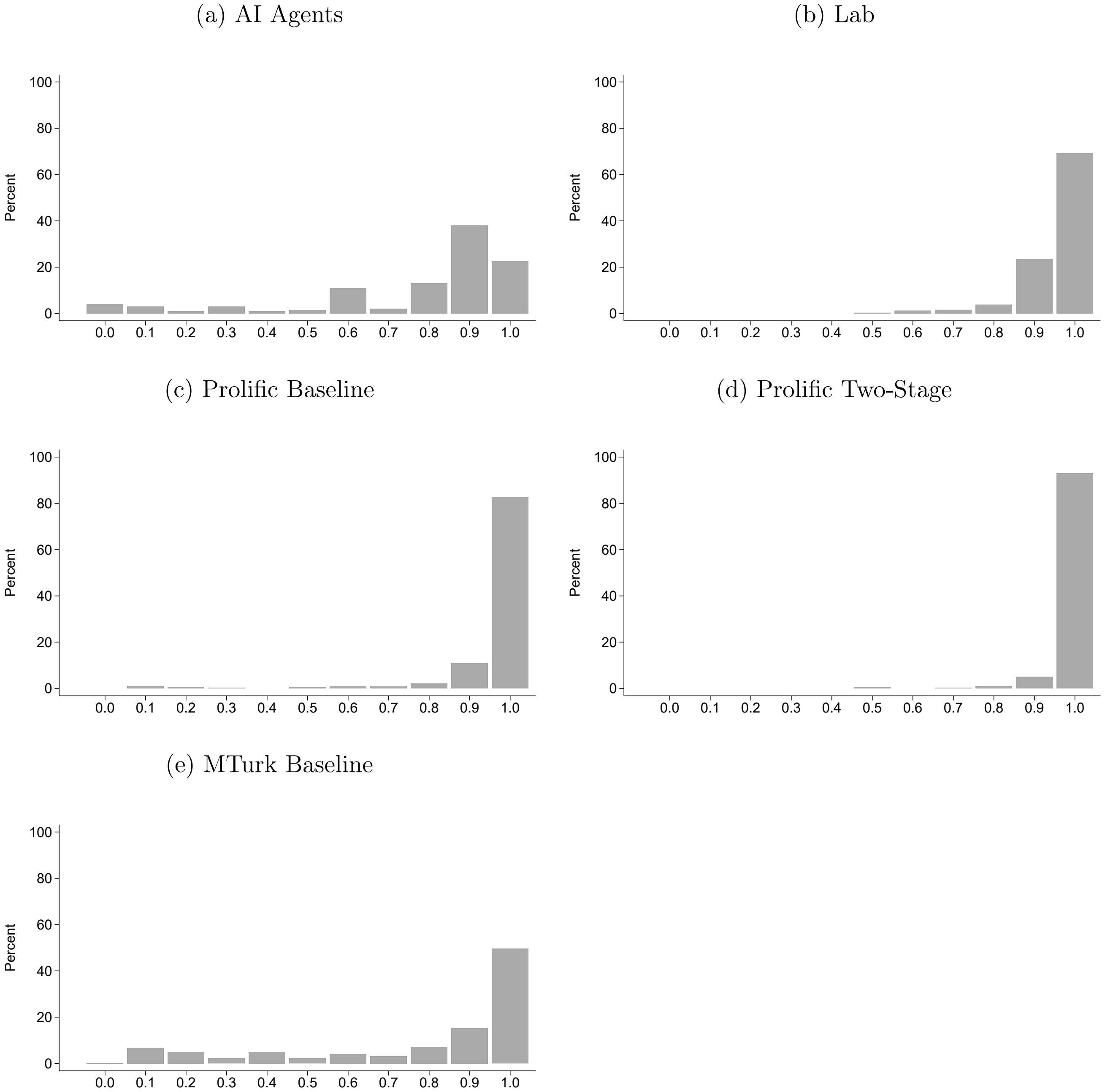
(c) Prolific Baseline

(d) Prolific Two-Stage

(e) MTurk Baseline

Notes: This figure presents histograms of mouse movements on the demographics page (questionnaire) by sample. Each subfigure corresponds to a different sample.

Figure A.4: reCAPTCHA Score



(a) AI Agents

(b) Lab

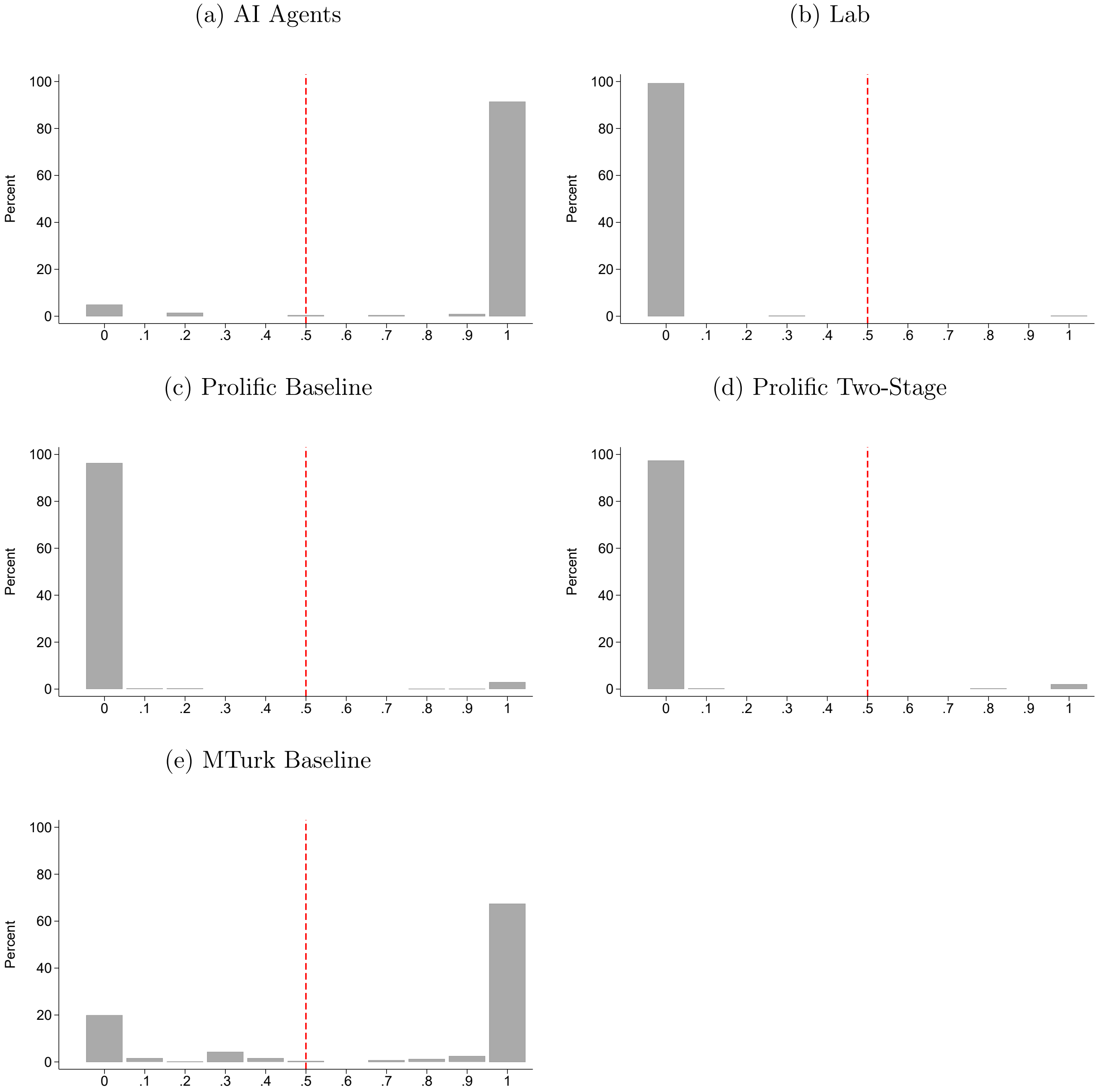
(c) Prolific Baseline

(d) Prolific Two-Stage

(e) MTurk Baseline

Notes: This figure presents histograms of reCAPTCHA scores by sample. Each subfigure corresponds to a different sample. See Appendix section B for more details on ReCAPTCHA scores.

Figure A.5: Pangram AI Likelihood


(a) AI Agents

(b) Lab

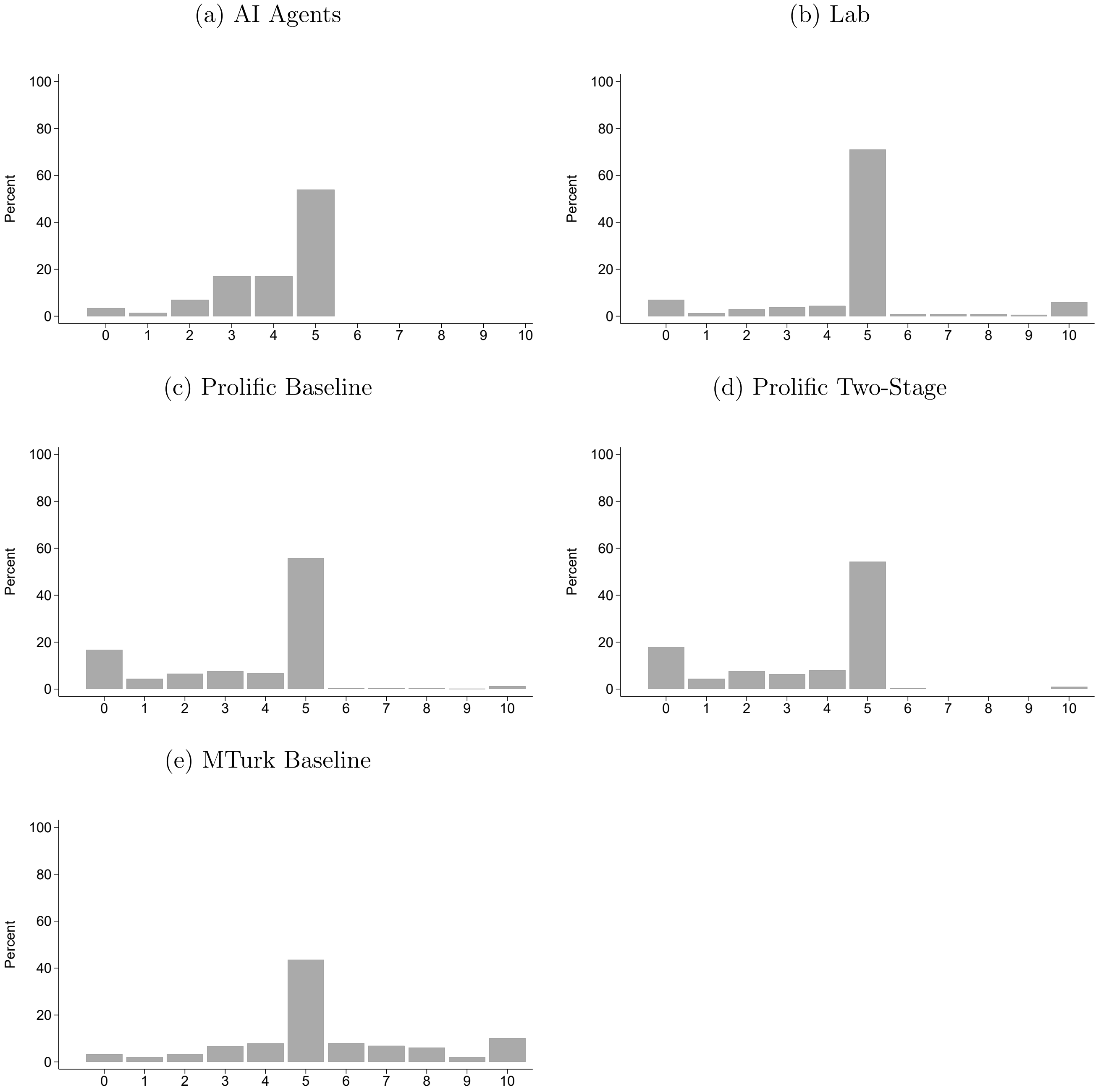
(c) Prolific Baseline

(d) Prolific Two-Stage

(e) MTurk Baseline

Notes: This figure presents histograms of Pangram AI likelihood by sample. Each subfigure corresponds to a different sample. The vertical dashed line at 0.5 indicates the cutoff for AI likelihood. See Appendix section B for more details on Pangram AI likelihood.

Figure A.6: Dictator Game Giving



(a) AI Agents

(b) Lab

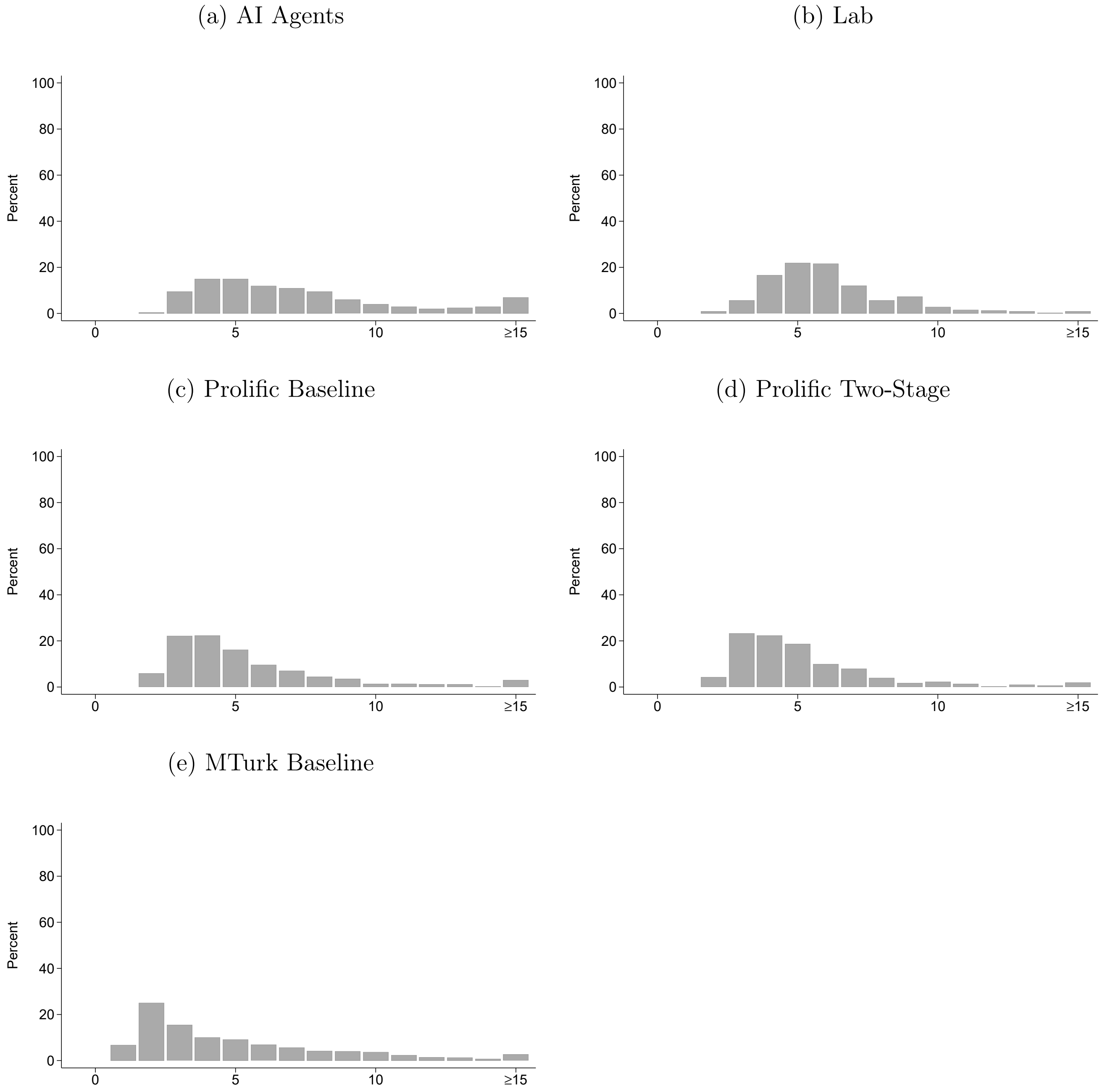
(c) Prolific Baseline

(d) Prolific Two-Stage

(e) MTurk Baseline

Notes: This figure presents histograms of Dictator Game giving by sample. Each subfigure corresponds to a different sample.

Figure A.7: Total Survey Duration (in minutes)



(a) AI Agents

(b) Lab

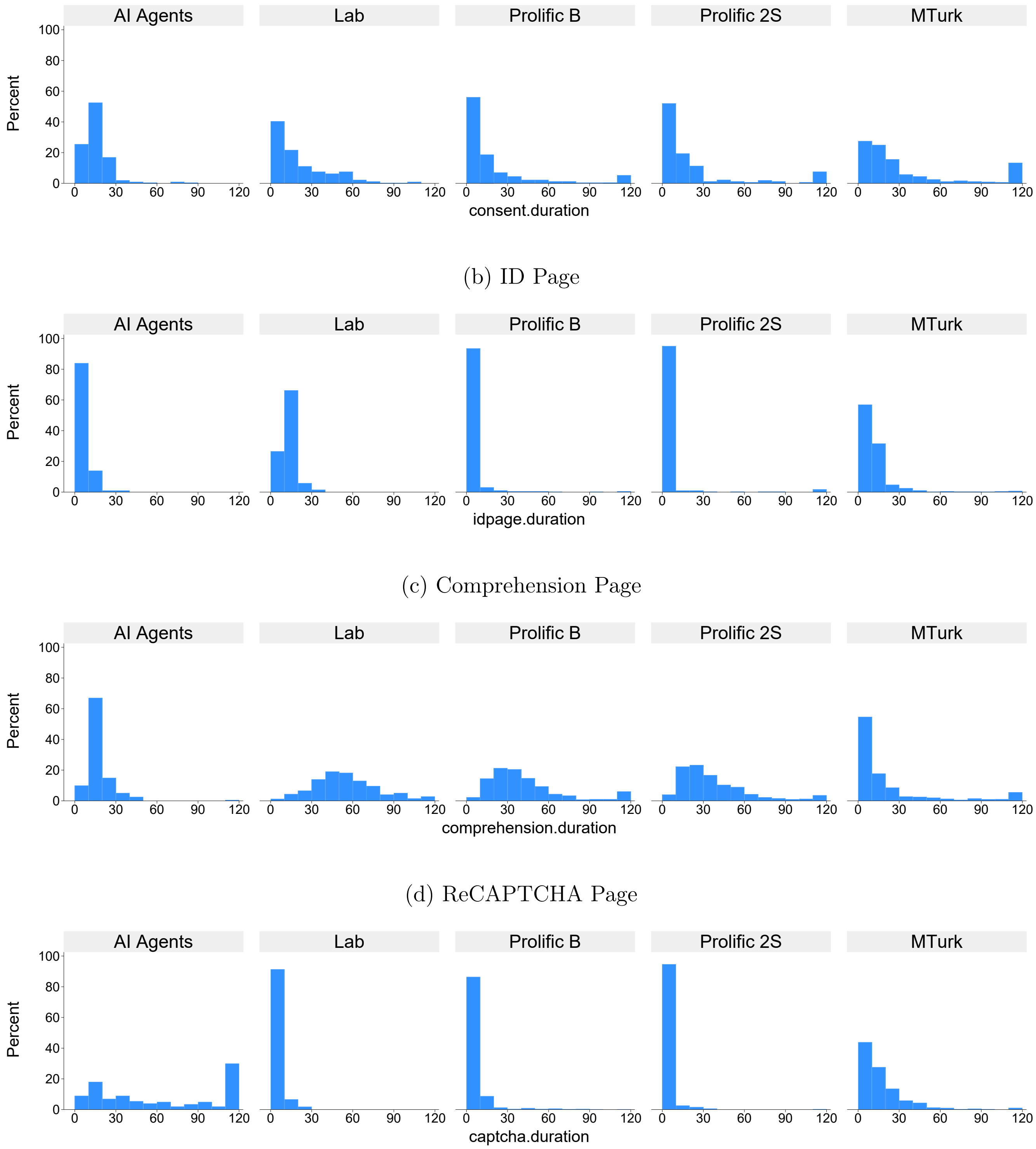
(c) Prolific Baseline

(d) Prolific Two-Stage

(e) MTurk Baseline

Notes: This figure presents histograms of total survey duration by sample. Each subfigure corresponds to a different sample.

Figure A.8: Response Time by Page and Sample (in seconds)

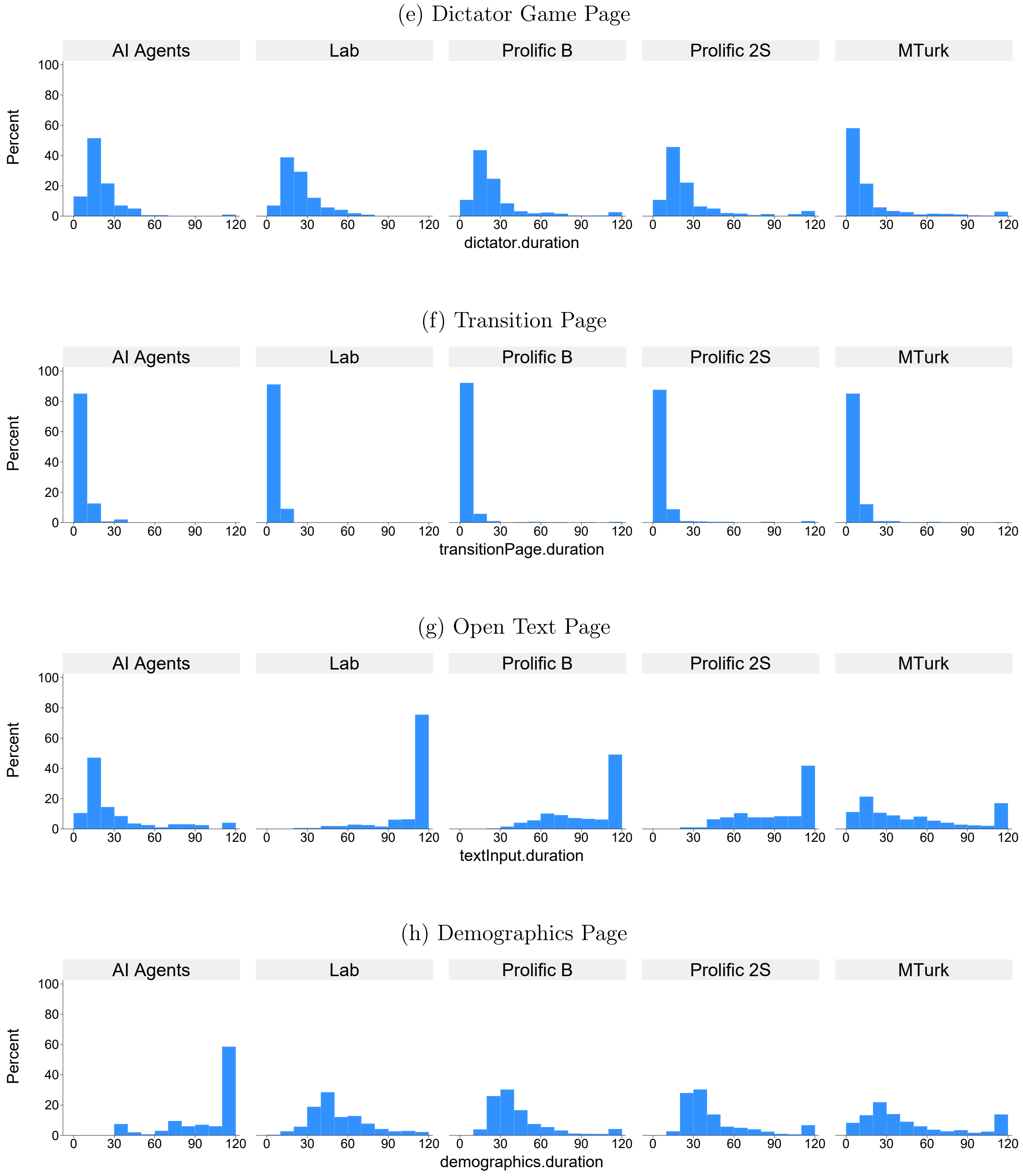


(a) Consent Page

(b) ID Page

(c) Comprehension Page

(d) ReCAPTCHA Page



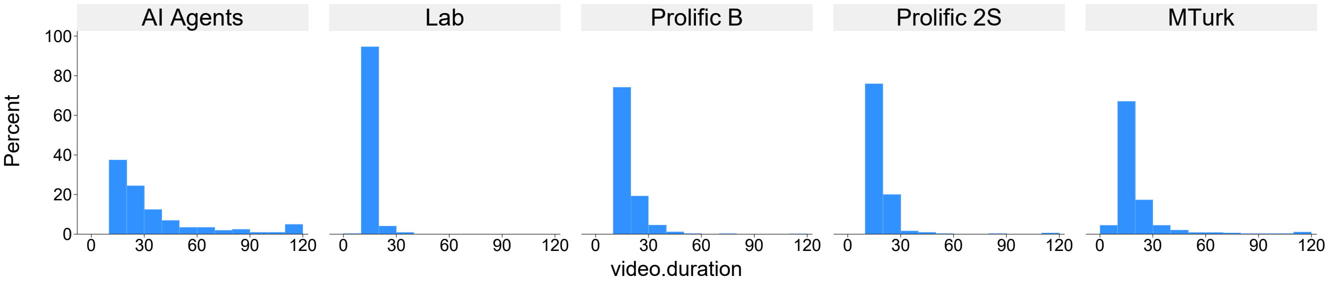
(e) Dictator Game Page

(f) Transition Page

(g) Open Text Page

(h) Demographics Page

(i) Video Page



Notes: This figure presents histograms of response times by page and sample. Each subfigure corresponds to a different survey page, showing the distribution of time (in seconds) that participants from each sample spent on that page. The samples include AI Agents, Lab participants, Prolific Baseline, Prolific Two-Stage, and MTurk Baseline.

B External Data Quality Measures

B.1 Google 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. Qualtrics provides a reCAPTCHA score for each survey response as a built-in feature. Google recommends using 0.5 as the cutoff value to distinguish between likely bot and human responses. In our data, only a small share of AI agents receive a reCAPTCHA score below the standard cutoff value of 0.5 (see Appendix Figure A.4), which is why we also show higher cutoffs in our analysis.

For more information on the ReCAPTCHA v3 score — and the ReCAPTCHA v2 challenge — by Google, see:

https://developers.google.com/recaptcha/docs/versions

For information about the implementation in Qualtrics, see:

https://www.qualtrics.com/support/survey-platform/survey-module/editing-questions/question-types-guide/advanced/captcha-verification/.

https://www.qualtrics.com/support/survey-platform/survey-module/survey-checker/fraud-detection/

B.2 MaxMind IP Information

MaxMind is a commercial IP intelligence service that provides detailed information about the geographical and network characteristics of an IP address. Using the MaxMind GeoIP2 web API, we queried the IP address associated with each survey response. The API returns a structured set of attributes, including country-level location data, the Internet Service Provider (ISP), and the Autonomous System Number (ASN) associated with the network through which the respondent connected. We classify responses based on whether the IP address originates from within the United States or from a non-U.S. location, allowing us to flag geographically inconsistent submissions. We also explored ISP and ASN data to identify consistent patterns among our AI agents. We observed that all ChatGPT Agent responses originated from the ASN "CLOUD-

FLARENET" belonging to ISP "Cloudflare Inc", one of the largest cloud services providers worldwide (https://www.cloudflare.com/). This observation is consistent with the idea that many AI agents are hosted on cloud infrastructure. In contrast, human respondents typically connect through residential or mobile ISPs, resulting in a diverse set of ISPs and ASNs.

For more information about MaxMind and the GeoIP2 web API, see:

https://www.maxmind.com/en/home

https://www.maxmind.com/en/geoip-api-web-services.

B.3 Pangram AI Detection

Pangram is a commercial AI detection tool designed to distinguish between human-written and AI-generated text. Imas and Jabarian (2025) evaluated Pangram against other leading AI detectors and found it to be exceptionally accurate, achieving near-zero false positive and low false negative rates on medium to long passages. The detector's performance remains robust even on 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 (Imas and Jabarian, 2025).

For more information about Pangram and the Pangram API, see:

https://www.pangram.ai/.

https://www.pangram.com/solutions/api

B.4 Fingerprint Device Identification

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 (visitor ID), which serves as a device fingerprint. This identifier is supposed to remain 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 fingerprint may not be generated in case users have browser settings or privacy extensions that prevent tracking scripts from running. This leads to some missing data points in our sample for this indicator. In our samples, this was the case for 14.2% of the MTurk sample, 14.6% of the Prolific sample, 20.7% of the lab sample, and 24.5% of the AI agent sample. We code such cases as having no duplicate fingerprint.

For more information about Fingerprint and Fingerprint integration, see:

https://fingerprint.com/.

https://dev.fingerprint.com/docs/quick-start-guide

C AI Agent Data Collection

C.1 Implementation and Procedures

Our AI agent data collection involved three AI agents available at the time of the study (August to October 2025): ChatGPT 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 ChatGPT Agent, the Perplexity Comet browser, and the BrowserUse Cloud UI. We relied on the platforms' predefined and optimized operational parameters, such as temperature and system prompt, as these are not user-configurable in the standard UI. Our prompts were provided as user messages within the chat interface. We consider this methodology representative of how a typical internet user would deploy such agents for survey-taking.

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. When an agent became stuck and explicitly requested help, human assitance was provided by the researcher. This situation occurred mainly at two specific points in the survey: the ReCAPTCHA challenge and the video attention check. To maintain a systematic record of these interventions, a strict protocol was followed. On the ReCAPTCHA challenge page, keystrokes were recorded. If an agent requested assistance, the researcher would press the 'X' key twice. Since no keystrokes were otherwise expected on this page, the presence of 'XX' in the keylog served as a reliable indicator that assistance had been requested. Similarly, for the video attention check, assistance was recorded by inputting 'XX' into the answer field. Following an instance of human assistance, control was returned to the AI agent by issuing a simple command like "proceed" or "continue taking the survey" in the chat interface.

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

Table C.1: Observations in AI Agent Sample

	Simple	Complex	*Total*
ChatGPT 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.

Costs of AI Agents At the time of the study, the ChatGPT Agent was available for subscribers to ChatGPT Plus at a cost of 20$ per month (with a limit of 40 ChatGPT Agent runs per month) or ChatGPT Pro at a cost of 200$ per month (with an unlimited number of runs).

For more information about the ChatGPT Agent, see: https://chatgpt.com/features/agent/.

At the time of the study, the Perplexity Comet browser was available via a waiting list procedure. We were granted access after applying through the Perplexity AI website. By now, the Comet browser is publicly available for free download and usage.

For more information about Perplexity Comet, see: https://www.perplexity.ai/comet.

The BrowserUse framework is an open-source browser automation library. It can be used with a subscription to BrowserUse Cloud, or with API keys and pay-as-you-go pricing. It is also also available for download via GitHub. The costs per agent run depend on the LLM model within the BrowserUse framework. For the GPT O3 and Gemini 2.5 Flash models, the costs were in a range of about 0.25$ to 0.75$ per completed survey run.

For more information about BrowserUse, see: https://browseruse.com/ and https://github.com/browser-use/browser-use

Google Mariner Agents In addition to the AI agents presented in Table C.1, we tried to collect data with Google DeepMind's AI agent Mariner. This AI agent successfully completed the survey in only 16 out of 136 attempts (11.8%). The low completion rate was primarily due to Mariner's inability to navigate past the reCAPTCHA challenge. 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 Mariner agent failed one of our main data quality checks. Therefore, including Mariner in our main sample would not have meaningfully changed our findings. The Mariner agents passed the standard attention checks. 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 open text flags, as it never produced typed text with typical typing speed (0%).

At the time of the study, the Google Mariner Agent was available for Google AI Ultra subscribers in the US at a cost of 250$ per month (with an unlimited number of runs).

For more information about Google DeepMind Mariner, see: https://deepmind.google/models/project-mariner/.

C.2 Prompts

Investigating the capabilities of LLMs and AI agents should involve careful consideration of prompt design, as variations in phrasing or structure may impact performance (Sclar et al., 2023; Mishra et al., 2021*a*). We therefore adopted the standardized, modular prompt template proposed by Çelebi and Penczynski (2024) to control for format and structure across our treatments. For both prompts, all the instructions are written in Markdown format and segmented into itemized lists to improve the model's instruction following capability (Mishra et al., 2021*b*). 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* provides the agent with basic instructions. The *Complex Prompt*, in contrast, provides the agent with a detailed persona and instructions to mimic human-like behavior.

The additions in the *Complex Prompt* served two purposes. First, we wanted to test if a sophisticated user—knowing that researchers monitor keystrokes, mouse movements, and CAPTCHA completion as bot-detection signals—could use detailed instructions in the prompt to evade detection. Second, we wanted to preempt 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.

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 (Mishra et al., 2021*a*; White et al., 2023). The "Role Persona" section is an attempt to steer the model towards acting more like a human (Lampinen et al., 2022). 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.

The *Complex Prompt* includes additional subsections intended to elicit more naturalistic, human-like 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. Finally, the Complex Prompt includes a "Constraints" section to improve the likelihood that the model will closely follow the provided instructions.

Figure C.1: Simple Prompt

```
# 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.2: Complex 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.
```

C.3 Survey Completion Diagnostics

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

With the exception of BrowserUse using Gemini 2.5 Flash, all agents either completed the survey in all attempts (Perplexity and BrowserUse with GPT O3) or failed in a small number of cases, primarily at the video attention check (ChatGPT agent). BrowserUse with Gemini 2.5 Flash showed lower completion rates, with most failures occurring at the ReCAPTCHA challenge. Completion rates do not differ systematically across prompt types. The 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) has a lower performance than the larger model (GPT O3).

For the ReCAPTCHA challenge, failure to solve the puzzle directly prevents agents from continuing, making it straightforward to infer that human assistance may be required. In contrast, the video attention check presents ambiguity: the task involves a video displaying a sequence of numbers, but agents perceive it only through intermittent screenshots. As a result, they often capture a single number or a blank frame between numbers. In such cases, agents must decide whether to respond with partial information, generate a full sequence without visual evidence, or request assistance. Perplexity Comet frequently identified that the task exceeded its perceptual capabilities and often prompted user intervention. In contrast, other agents rarely recognized that their limited visual sampling prevented them from completing the task as instructed and seldom requested assistance. This behavior showed minimal variation across prompt types and did not differ systematically between smaller (Gemini 2.5 Flash) and larger (GPT O3) models within the same framework (BrowserUse).

The "Completed Persona Profile" row applies only to the complex prompt condition, where agents received a demographic persona to follow (see Figure C.2). This measure captures whether agents provided demographic responses that fully matched the assigned persona (age, gender, state, political affiliation). The ChatGPT agent consistently matched the persona profile, while Perplexity Comet failed in 15% of cases. BrowserUse agents failed in the majority of cases, with failure rates higher when using the smaller LLM (Gemini 2.5 Flash) compared to the larger one (GPT O3).

Table C.2: AI Agents: Survey Completion

	GPT-S (1)	GPT-C (2)	PERP-S (3)	PERP-C (4)	BU-O3-S (5)	BU-O3-C (6)	BU-G25F-S (7)	BU-G25F-C (8)	Total (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 Check	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: Fails / Stops	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 Check: Fails	0.91	0.93	0.10	0.15	1.00	1.00	1.00	0.95	0.79
Video Check: Stops	0.07	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.02
Video Check: Passes with help	0.02	0.02	0.90	0.85	0.00	0.00	0.00	0.05	0.19
Video Check: Passes w.o. help	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
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 Check, Completed Study) by agent/prompt. ReCAPTCHA and Video Check sections show outcome shares (conditional on arriving at this page) in the mutually exclusive categories reported (e.g., "fails", "stops", "passes with help", "passes w.o. help"); "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 ChatGPT 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 GPT O3) 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, Perplexity Comet, and BrowserUse—represent a new class of agentic AI systems designed to perform complex tasks in user-facing applications such as web browsers (?). A key advancement distinguishing these AI agents from earlier scripted bots is their ability to interpret both a webpage's source code (HTML) and its visual layout through screenshots (He et al., 2024).

The agents' reliance on both visual and textual inputs necessitates the use of large multimodal models (LMMs) rather than LLMs (He et al., 2024). 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 may be more likely to successfully complete a complex tasks like survey completion than smaller models.

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 (?). These actions may be implemented differently across agents.

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 and thus economically costly. To operate under these constraints, agents typically employ context clipping, retaining only the most recent observations and actions (He et al., 2024). As a result, they are optimized to take screenshots sparingly, usually capturing the state of the page only after an action has been executed.

The discrete, state-based observation and action strategy of AI agents may render them ill-suited for processing dynamic or continuous content such as videos, and may lead to discrete action sequences that are easily distinguishable from human browsing behavior.

D JavaScript Tracking Code

We provide this code via https://github.com/survey-data-quality-lab/mission-possible.

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., copy-paste events, mouse movements, and mouse clicks. 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 JavaScript code 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. 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. The tracker generates multiple entries per page if the page is reloaded (for example, when respondents did not complete all required fields). This should be taken into account when considering the expected number of tracking entries for a survey. Also note that 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() });
    });

    // Save data when the page is submitted
    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 JSON
        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 discrte input jumps by comparing the amount of text in the input field before and after each input event. The keylog tracker records each input jump of more than ten characters. Such input jumps may indicate pasting or other non-typed input like drag-and-drop or auto-completion.

The script is used on the two pages that contain text input fields: the open-text 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 JavaScript 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. When using the keylog tracker on multiple pages, make sure to create separate embedded data fields for each page (e.g., “*key_log*_1", “*key_log*_2", etc) and update the code accordingly on each page:

```
Qualtrics.SurveyEngine.setEmbeddedData("key_log_1", JSON.stringify(keylog)
   );
```

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 JavaScript into qualtrics survey header under html view (Survey -> Look and Feel -> Header -> Source).
3. Update `YOUR_API_KEY_HERE` with your Fingerprint API key.
4. Create embedded data fields named *page*, *visitorId*, *requestId*, and *fp_error* in the survey flow before the first survey block. Set *page* 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; }

    // Import library
    window.fpPromise = window.fpPromise || import(’https://fpjscdn.net/v3/YOUR_API_KEY_HERE’)
      .then(function(FingerprintJS) { return FingerprintJS.load(); });

    // Generate fingerprint
    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);

    // Store in Qualtrics embedded data
    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.

```json
[
  {
    "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 **ChatGPT agent**. As shown, the model simply performs a paste event via `Control + v`, followed by an `INPUT_JUMP`. The `INPUT_JUMP` indicates that 522 characters were entered into the input field within a single input event leading to a total of 522 characters in the text input field.

```
[
{"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 ChatGPT 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 ChatGPT 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 ChatGPT 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 ChatGPT 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 writing 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 time 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.

E.3 Empty Keylogs

In our dataset, 3.1% of respondents have empty keylogs (N=86, all on MTurk), indicating no recorded keystrokes and no input jump events during their interaction with the open-ended question. At the same time, these respondents also did not trigger any paste event, but submitted text with an average length of 698 characters. This pattern is consistent with some form of automatic or scripted text input into the open text box.

E.4 Paste and Input Jump Events

Based on our JavaScript tracker data, 70.9% of open-text responses involve no paste event, and 70.3% involve no large input jump event (>50 characters). These two indicators show high consistency (>98% consistency). Table E.1 presents a cross-tabulation of the occurrence of paste events and large input jump events across all respondents. Around 0.5% of responses involve a paste event with no corresponding large input jump event, and 1.1% involve a large input jump event with no paste event (e.g., respondents using drag-and-drop).

Table E.1: Cross-Tabulation of No Paste Event and No Input Jump Event

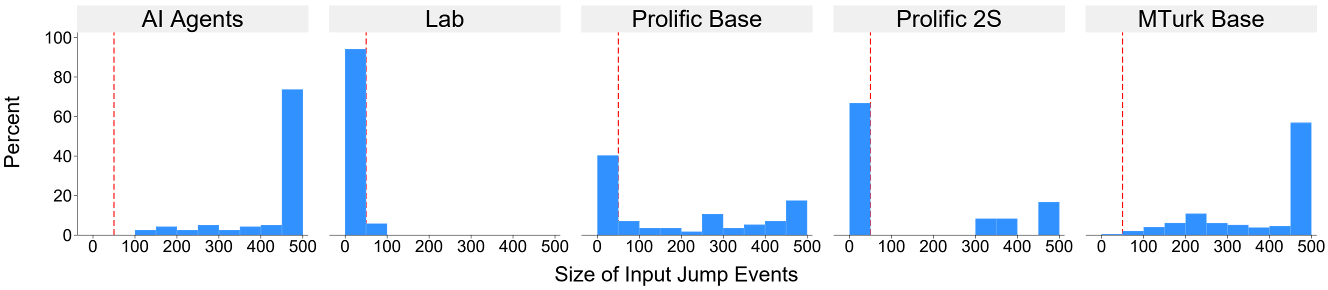
	No input jump event > 50		
No paste event	= 0	= 1	Total
= 0	744 (28.67)	12 (0.46)	756 (29.13)
= 1	28 (1.08)	1,811 (69.79)	1,839 (70.87)
Total	772 (29.75)	1,823 (70.25)	2,595 (100.00)

Notes: Entries report frequencies, with cell percentages in parentheses.

Figure E.6 illustrates the distribution of input jump sizes across all samples. 68.3% of responses involve no input jump event (> 10 characters), while 31.7% involve at least one such event, with an average input jump size of 482 characters.

In the lab, input jump events are rare (5.4%) and most input jumps are small (<50 characters). This shows that while some lab participants may make minor corrections or additions to their text, larger input jumps (e.g., pasting or drag-and-dropping of entire sentences) are virtually absent in this controlled environment. In contrast, in our AI agent sample, input jump events are frequent (59%) and are all large (>100 characters), with most exceeding 500 characters.

Figure E.6: Size of Input Jump Events



Notes: This figure shows the distribution of input jump sizes across our lab, online, and AI agent samples. The vertical dashed line at 50 characters indicates the cutoff for large input jumps.

E.5 Typing Speed

Typing speed is calculated using the keylog data. We compute the time intervals between consecutive keystrokes and then calculate the median of these intervals: "*Median Typing Speed*". A respondent is classified as typing with typical speed if their median typing speed is slower than 75 milliseconds per keystroke. Observations with less than two recorded keystrokes have a typing speed of 0 and are thus classified as not typing with typical speed.

Using the keylog data from the BrowserUse agent in Figure E.4 as an example, we compute the time intervals between each pair of consecutive keystrokes. The first few intervals are as follows:

```
[14, 15, 12, 11, 10, 9, 9, 10, 19, 19, 15, 10, 14, 19, 18, 22, 14, 26, 17,
18, 19, 16, 18, 12, 16, 22, 15, 15, 13, 18]
```

Sorting these 30 differences in ascending order yields:

```
[9, 9, 10, 10, 11, 12, 12, 13, 14, 14, 14, 15, 15, 15, 16, 16, 17, 18, 18,
18, 18, 19, 19, 19, 19, 22, 22, 26]
```

The median inter-keystroke interval for this agent is 16 ms. Since this value is below the 75 ms threshold, this agent would **not be classified as typing with typical human speed**.

Using the keylog data from a human participant in the lab from Figure E.5, we compute the time intervals between each pair of consecutive keystrokes. The first 29 intervals are as follows:

```
[119, 152, 129, 72, 247, 152, 432, 1065, 776, 160, 159, 249, 415, 192, 160,
168, 2233, 175, 168, 169, 151, 153, 151, 1352, 160, 145, 224, 160, 224]
```

Sorting these 29 differences in ascending order yields:

```
[72, 119, 145, 151, 151, 152, 152, 153, 159, 160, 160, 160, 160, 168, 168,
169, 175, 175, 192, 224, 224, 247, 249, 415, 432, 776, 1065, 1352, 2233]
```

The median inter-keystroke interval for this human participant is 168 ms. Since this value is above the 75 ms threshold, this participant would be **classified as typing with typical human speed**.

E.6 Validation of Text Based Indicators

To validate our main text-based indicators *Typed text* and *Typed with typical speed*, we present comparisons with two additional measures derived from the keylog data, *Text similarity score* and *Key count*, as well as with *Pangram AI likelihood.*

E.6.1 Text Similarity Score

To capture how closely the recorded keystroke log reproduces the final submitted text, we reconstruct the typed content from the raw keylog using deterministic parsing rules. Specifically, we ignore modifier and navigation keys (e.g., Shift, Control, Arrow keys, Tab, Enter); skip common Control-based shortcuts (c/v/a); treat Backspace as deletion of the most recent character; and append all other keys to build a reconstructed string R. We then compare R to the actual submitted response O (the text field value recorded by Qualtrics) using the Levenshtein edit distance $d_L(R, O)$, which counts the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one string into the other.

To obtain a scale-free metric, we normalize this distance by the length of the longer string and report the similarity score

$$S = 1 - \frac{d_L(R, O)}{\max\{|R|, |O|\}},$$

where $S \in [0, 1]$. Higher values of S indicate closer agreement between the reconstructed and submitted text, with $S = 1$ if and only if $R = O$. By convention, we set $S = 0$ if one of R or O is empty. This measure provides a direct assessment of the extent to which the keylog reproduces the submitted answer.

Lower scores can be the result of:

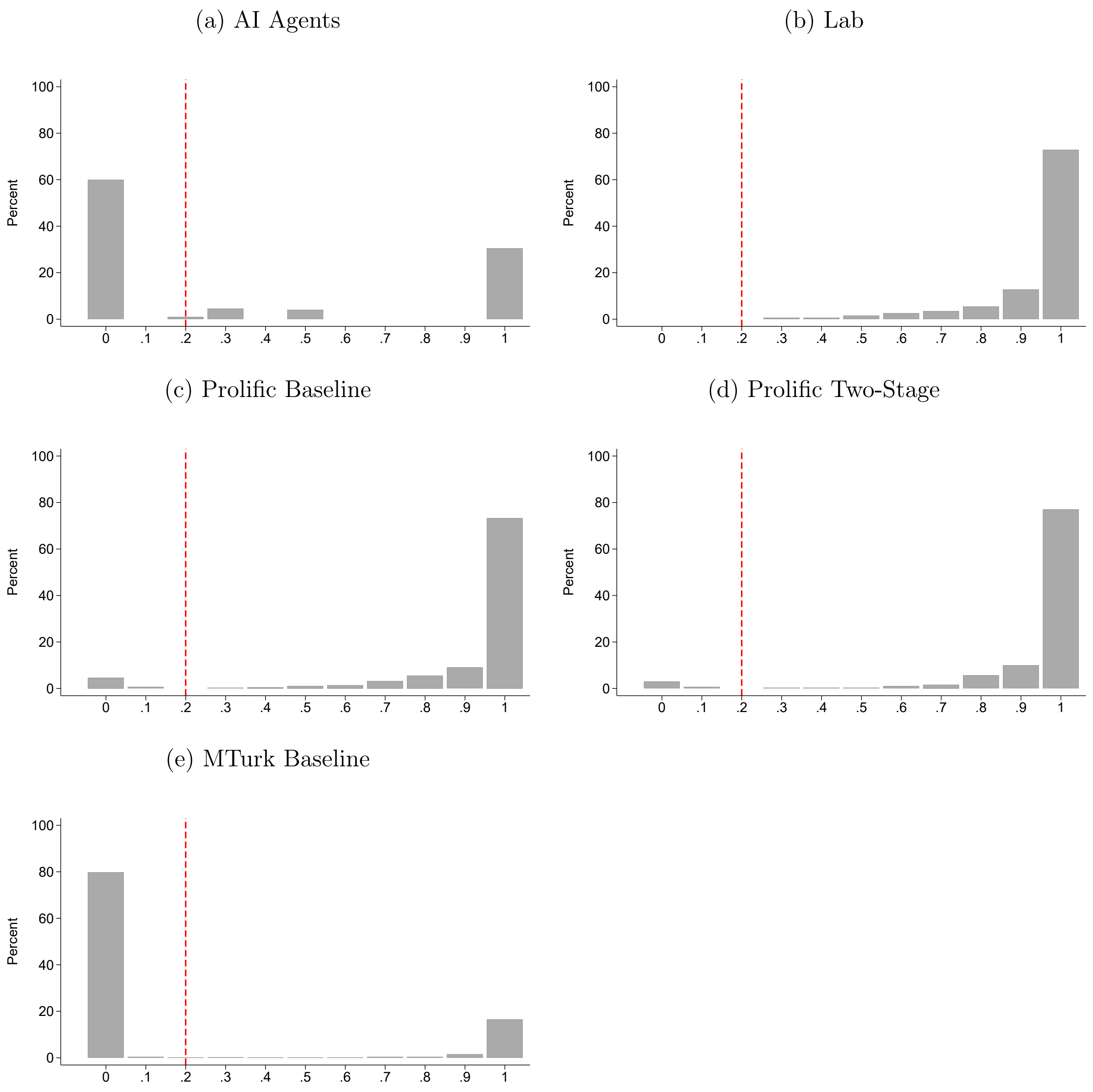
- Pasting content from external sources
- Using drag-and-drop to insert text from external sources
- Using other non-typing input methods

But also:

- Deleting larger chunks of text at once (instead of character-by-character backspacing)
- Moving parts of the text around (e.g., changing the order of sentences)
- Using mobile devices which may not log keys accurately (the keylog may contain "Unidentified" entries instead of keys)

Figure E.7 shows the distribution of text similarity scores across our lab, online, and AI agent samples. Lab subjects exhibit scores of 0.3 and larger. Hence, we choose a cutoff of > 0.2 to classify respondents as having low text similarity.

Figure E.7: Text Similarity Score



(a) AI Agents

(b) Lab

(c) Prolific Baseline

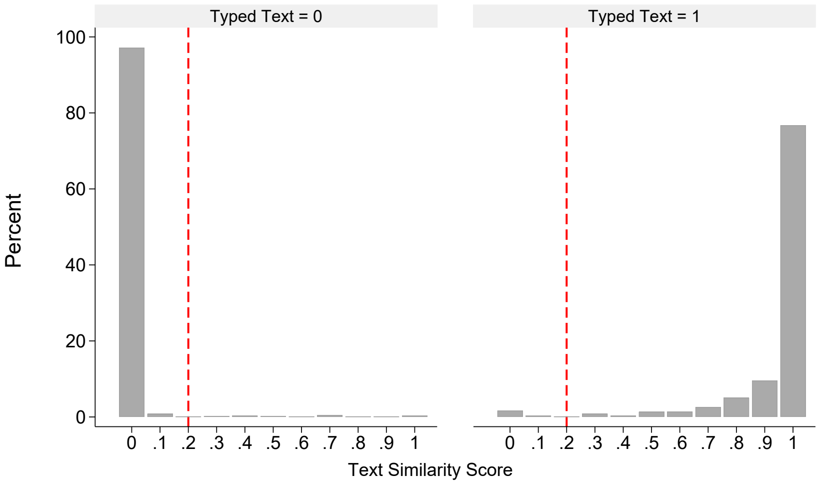
(d) Prolific Two-Stage

(e) MTurk Baseline

Notes: This figure shows the distribution of text similarity scores (based on Levenshtein distance between reconstructed and submitted text) across our lab, online, and AI agent samples. The vertical dashed line at 0.2 indicates the cutoff for low text similarity.

Figure E.8 and Table E.2 show that our *Typed text* indicator (based on paste detection, large input jumps, and empty key logs) produces highly consistent classifications with the *Text similarity score* (consistency >98%).

Figure E.8: Typed Text and Text Similarity Score



Notes: This figure shows the relationship between our main *Typed text* check (based on paste detection, large input jumps, and empty key logs) and the *Text similarity score* (based on Levenshtein distance between reconstructed and submitted text). The vertical dashed line at 0.2 indicates the cutoff for low text similarity. The left panel shows those who fail the *Typed text* check and the right panel shows those who pass the *Typed text* check.

Table E.2: Cross-Tabulation of Typed Text and Text Similarity Score

	Typed text		
Text similarity score	= 0	= 1	Total
≤ 0.2	853 (32.87)	34 (1.31)	887 (34.18)
> 0.2	17 (0.66)	1,691 (65.16)	1,708 (65.82)
Total	870 (33.53)	1,725 (66.47)	2,595 (100.00)

Notes: Entries report frequencies, with cell percentages in parentheses.

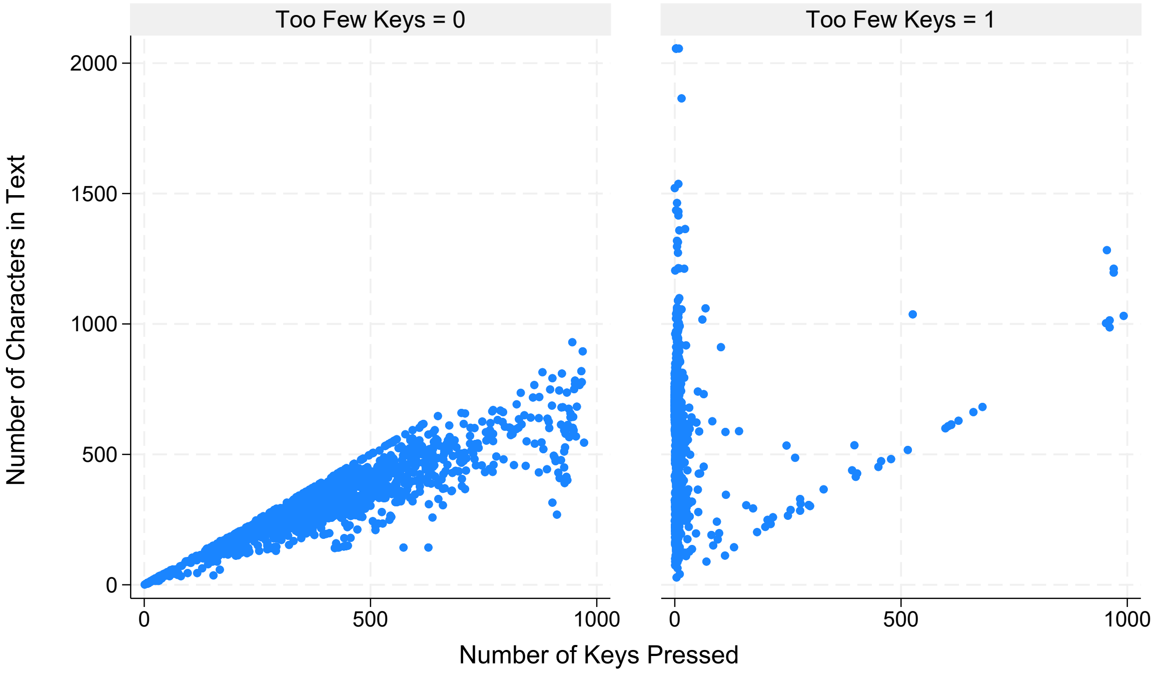
E.6.2 Key Count

We calculate the total number of keystrokes recorded in the keylog for each respondent. This count includes all key presses, encompassing character keys, modifier keys (e.g., Shift, Control), navigation keys (e.g., Arrow keys, Tab, Enter), and special keys (e.g., Backspace).

We compare this total key count to the length of the final submitted text. A respondent is classified as having *Too few keys* if the total number of keystrokes is less than the length of the submitted text.

Note that this measure may flag respondents as having too few keys in cases where the keylog is reaching it's capacity limit (approximately 1000 keystrokes). In such cases, the keylog may be truncated, leading to an undercount of total keystrokes. Therefore, respondents with keylogs at or near this limit should be interpreted with caution. In rare instances the keylog tracker may also miss a keystroke due to technical issues, which could also lead to an undercount of keystrokes.

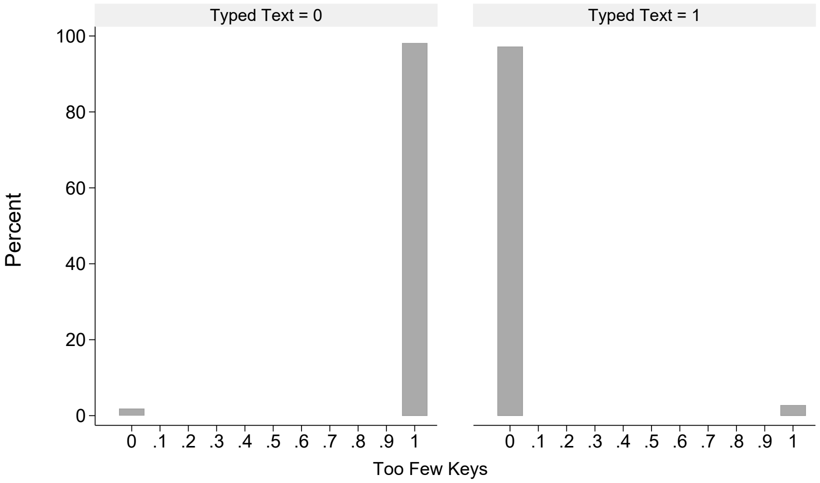
Figure E.9: Too Few Keys



Notes: This figure shows the relationship between the total number of keystrokes recorded in the keylog (x-axis) and the length of the submitted text response (y-axis). Respondents below the 45-degree line are classified as having enough keys (left panel) because their total keystrokes are larger than the length of their submitted text. Respondents above the 45-degree line are classified as having *Too few keys* (right panel) because their total keystrokes are less than the length of their submitted text.

Figure E.10 and Table E.3 show that our *Typed text* indicator (based on paste detection, large input jumps, and empty key logs) produces highly consistent classifications with the *Too few keys* indicator (consistency >97.5%).

Figure E.10: Typed Text and Too Few Keys



Notes: This figure shows the relationship between our main *Typed text* check (based on paste detection, large input jumps, and empty key logs) and the *Too few keys* indicator (based on total keystrokes being less than the length of the submitted text). The left panel shows those who fail the *Typed text* check and the right panel shows those who pass the *Typed text* check.

Table E.3: Cross-Tabulation of Typed Text and Too Few Keys

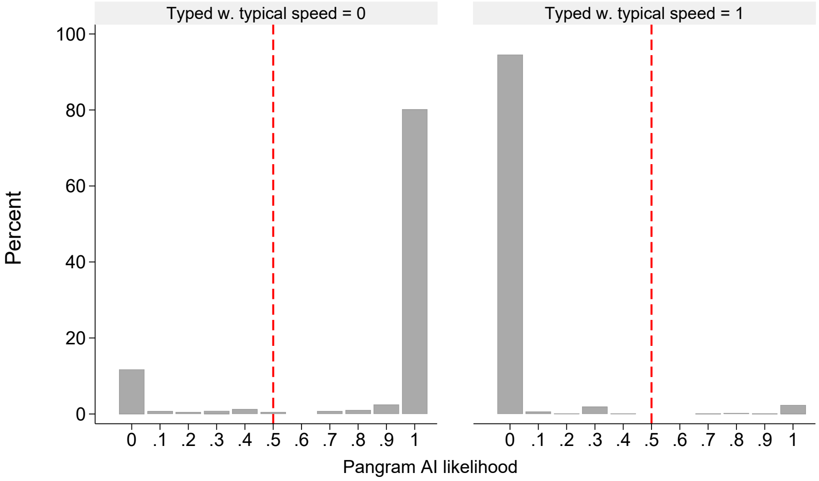
	Typed text		
Too few keys	= 0	= 1	Total
= 1	854 (32.91)	48 (1.85)	902 (34.76)
= 0	16 (0.62)	1,677 (64.62)	1,693 (65.24)
Total	870 (33.53)	1,725 (66.47)	2,595 (100.00)

Notes: Entries report frequencies, with cell percentages in parentheses.

E.6.3 Pangram AI Likelihood

Figure E.11 and Table E.4 show that the *Typed with typical speed* indicator produces very consistent classifications with the *Pangram AI likelihood* $<$ *0.5* indicator (>92.5% consistency). See Appendix Section B.3 for details on the Pangram AI detection method.

Figure E.11: Typed with Typical Speed and Pangram AI Likelihood



Notes: This figure shows the relationship between our *Typed with typical speed* check (based on keystroke timing and input patterns) and the Pangram AI likelihood. The left panel shows those classified as not having *Typed with typical speed* and the right panel shows those classified as having *Typed with typical speed.* The vertical dashed line at 0.5 indicates the standard cutoff for Pangram AI likelihood.

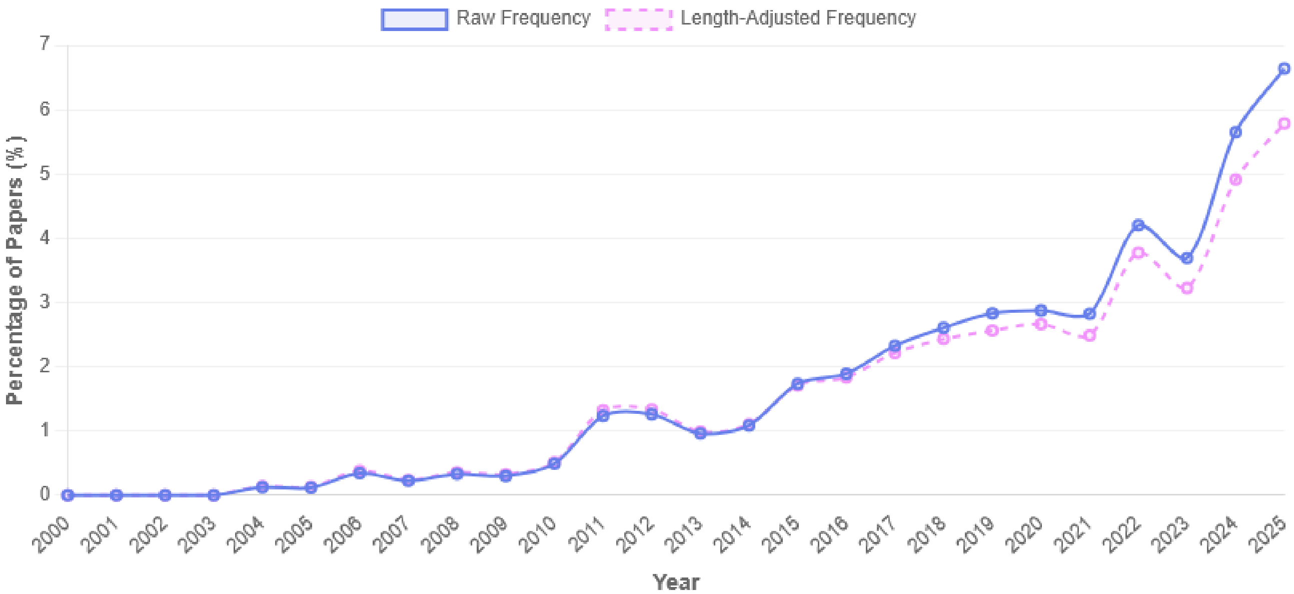
Table E.4: Cross-Tabulation of Typed with Typical Speed and Pangram AI Likelihood

	Typed with typical speed		
Pangram AI likelihood	= 0	= 1	Total
≥ 0.5	812 (31.29)	44 (1.70)	856 (32.99)
< 0.5	150 (5.78)	1,589 (61.23)	1,739 (67.01)
Total	962 (37.07)	1,633 (62.93)	2,595 (100.00)

Notes: Entries report frequencies, with cell percentages in parentheses.

F The Rise of Online Surveys and Experiments

Figure F.1: Use of Online Surveys and Experiments Over Time



Notes: This figure shows the share of papers in AEA journals and NBER Working Papers that contain the terms "online survey" or "online experiment" from 2000 to 2025. Source: https://paulgp.com/econlit-pipeline/

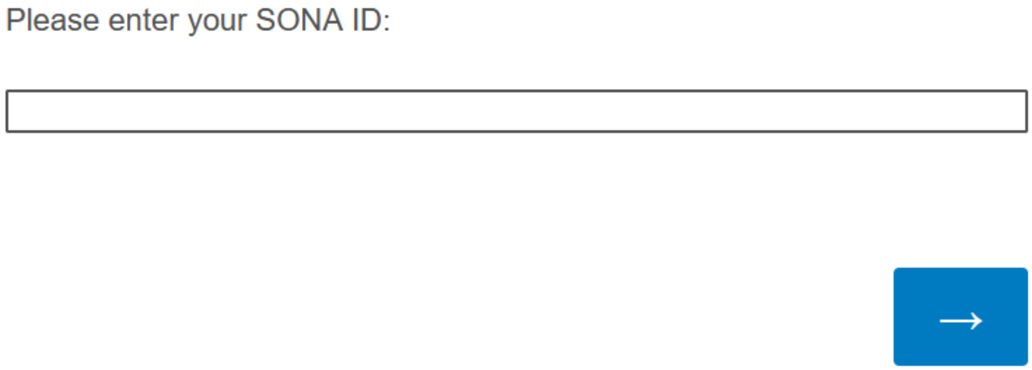
G Survey Instructions

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 baseline 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. The baseline 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.Screenshots of the lab instructions are shown in Figures G.1–G.11. After consenting to participate in the study, participants are informed of the opportunity to earn an additional payment. Figure G.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 G.4. If the algorithm is uncertain whether a participant is human, it presents an image challenge in which participants must select the correct images.

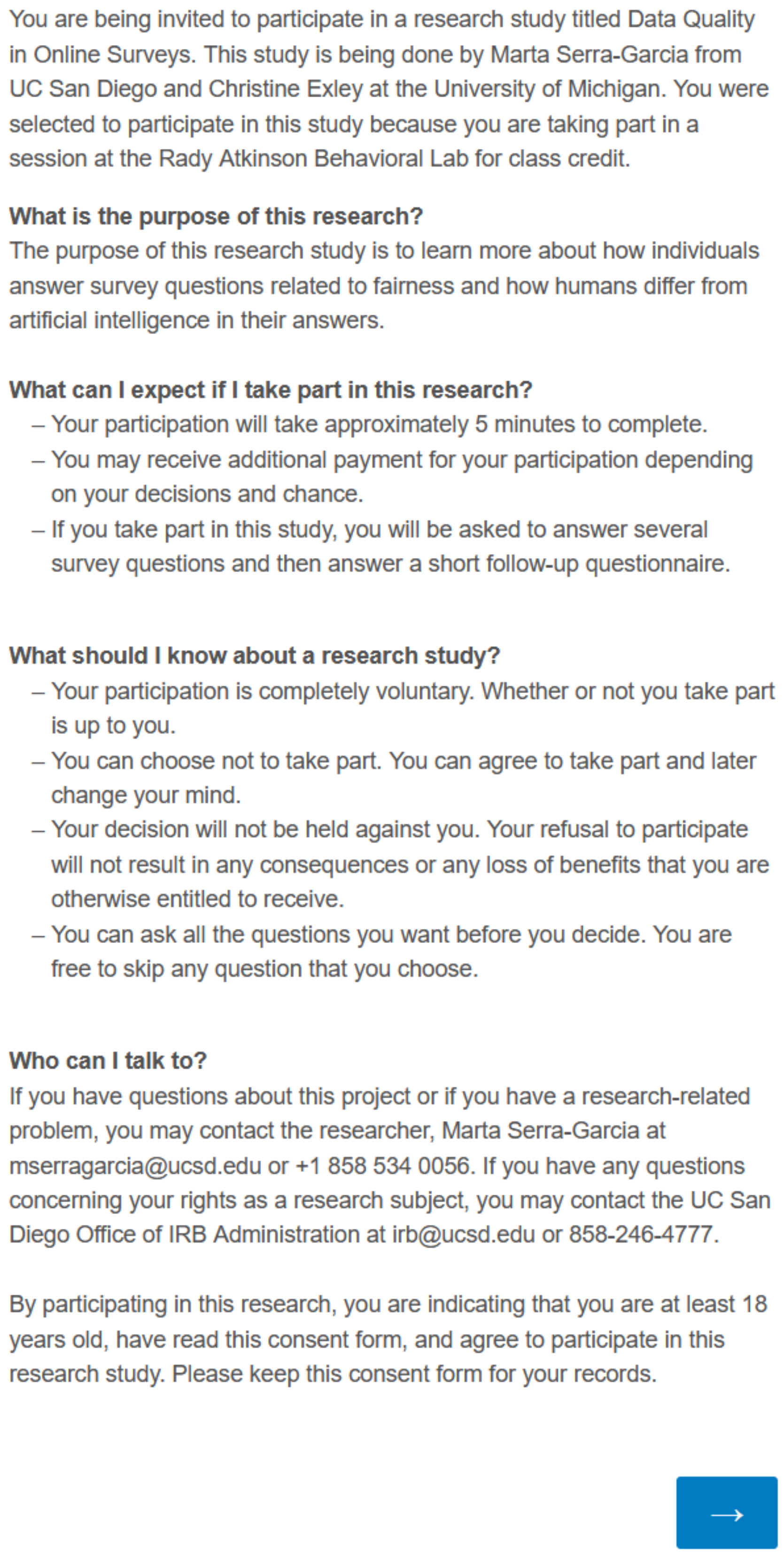
Participants then make their decision in the Dictator Game. Figure G.5 shows the instructions for the Dicatator Game and the decision question. Participants are then informed that they have to complete three more pages with additional questions. First, participants answer an open-ended question asking 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 G.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 G.8 displays the instructions for this classic attention check. The second page also includes questions about participants' demographic information, shown in Figures G.9 and G.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 G.11 shows the instructions for the video attention check.

Figure G.1: Participant ID



Please enter your SONA ID:

Figure G.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.

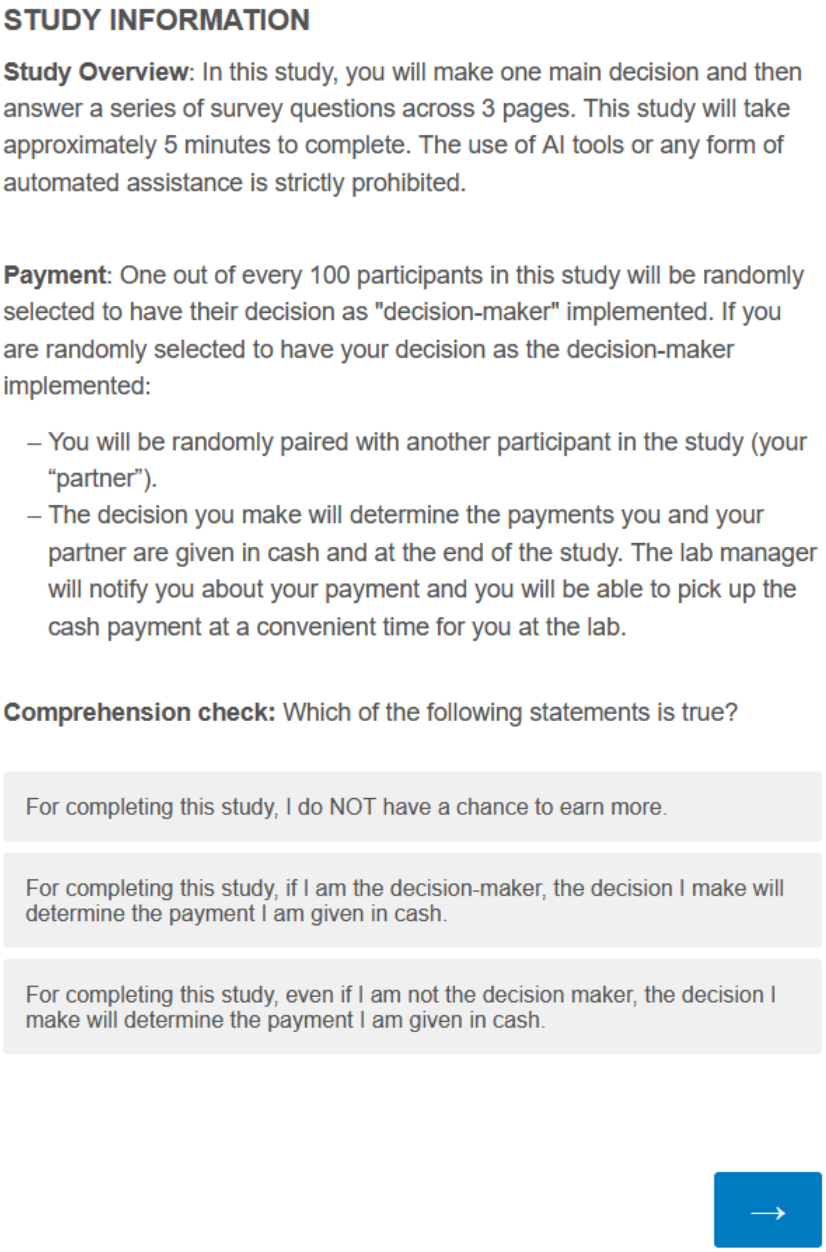
What should I know about a research study?

- 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 G.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.

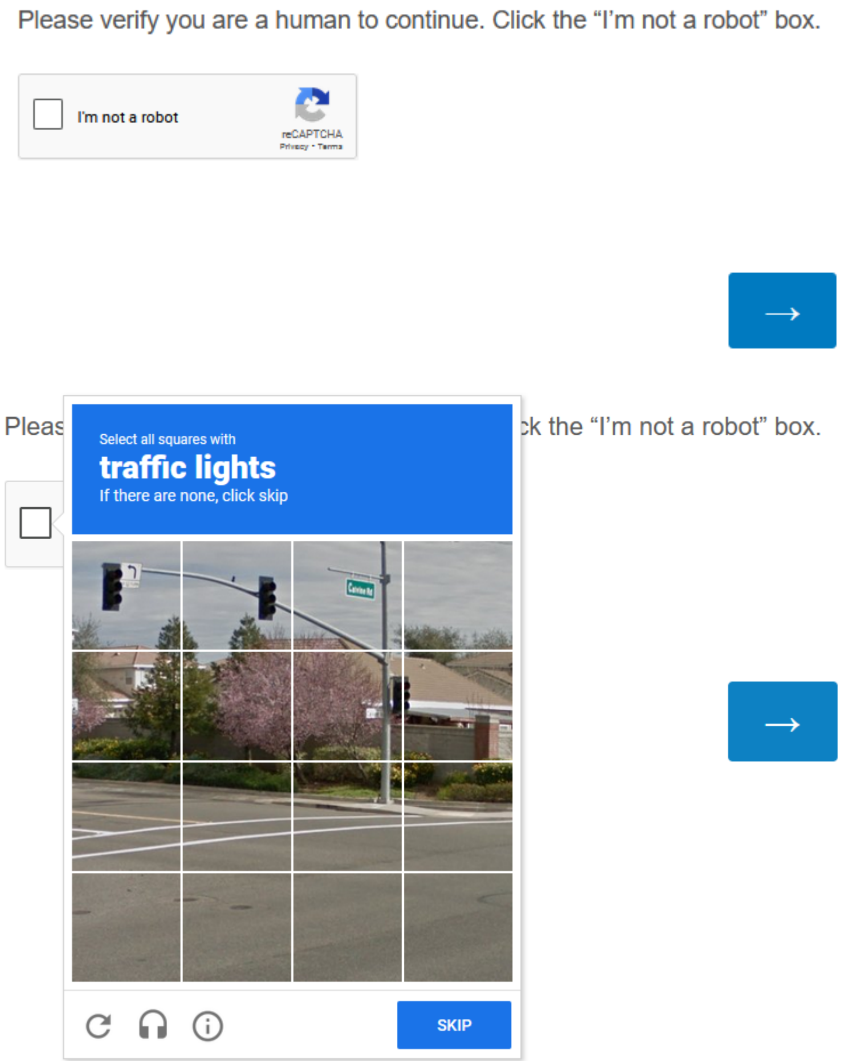
Comprehension check: Which of the following statements is true?

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 G.4: ReCAPTCHA



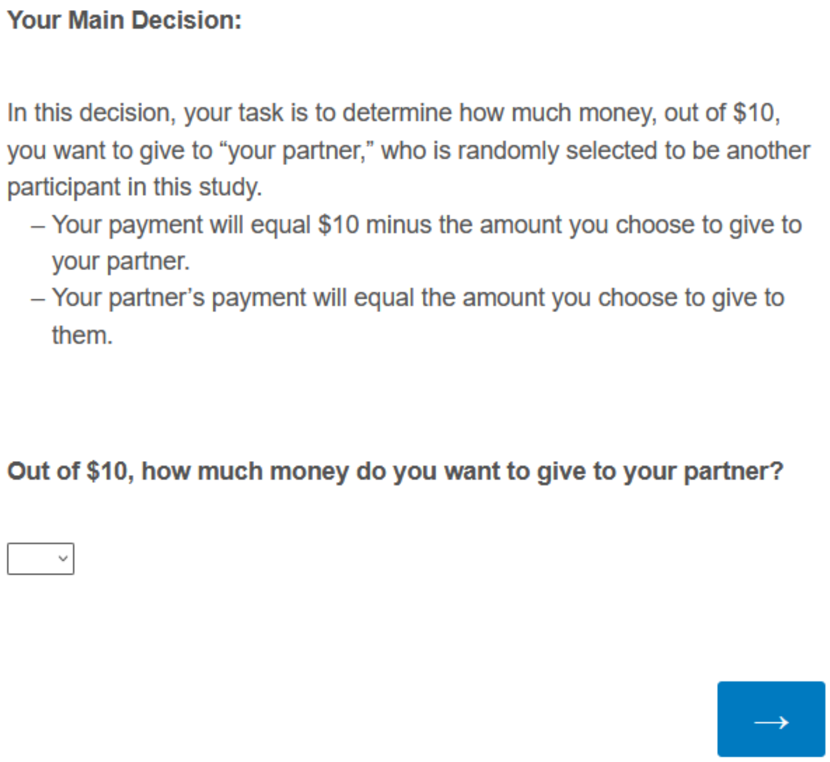
Please verify you are a human to continue. Click the “I’m not a robot” box.

□ 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 G.5: Dictator Game Decision



Your Main Decision:

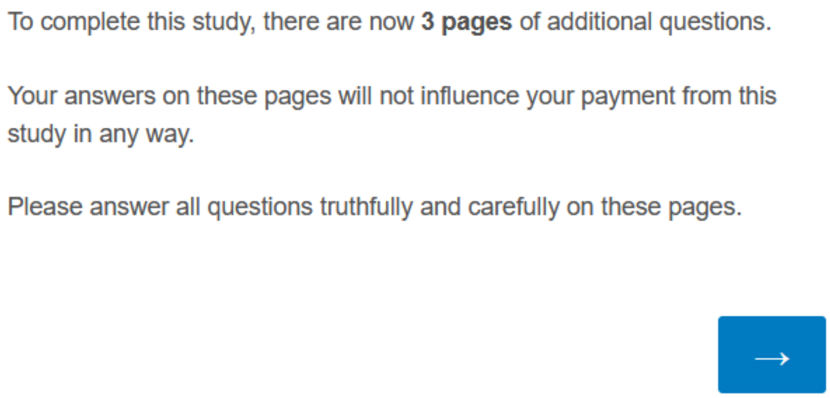
In this decision, your task is to determine how much money, out of $10, you want to give to “your partner,” who is randomly selected to be another participant in this study.

– 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 G.6: Transition Page

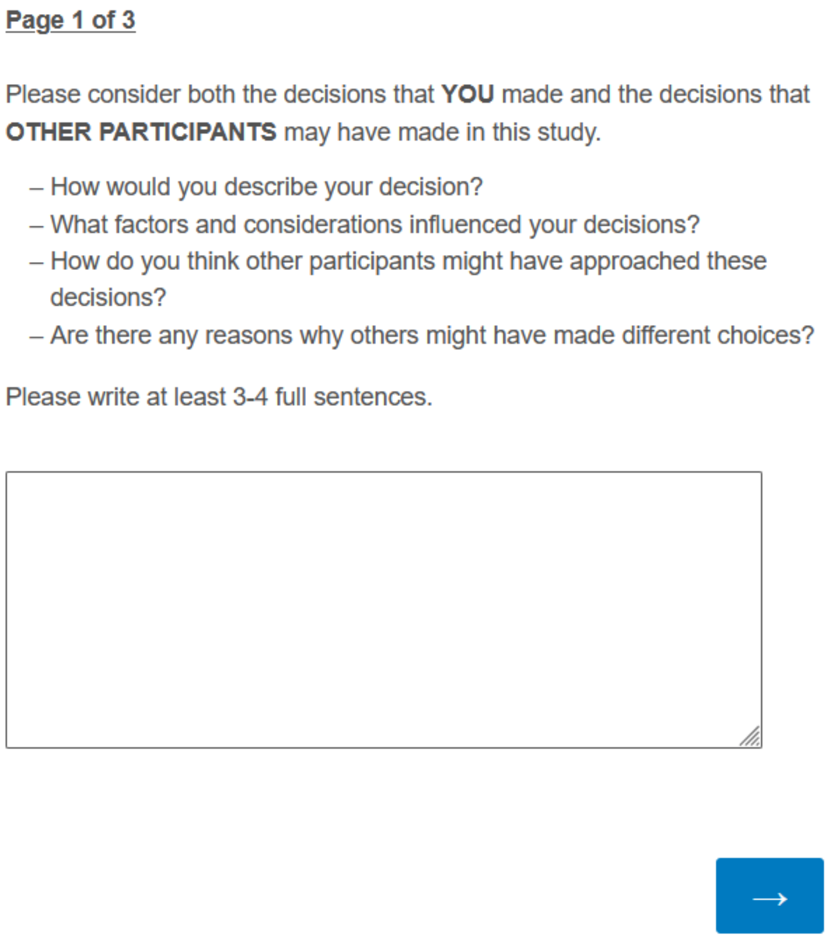


To complete this study, there are now **3 pages** of additional questions.

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 G.7: Open-Ended Question



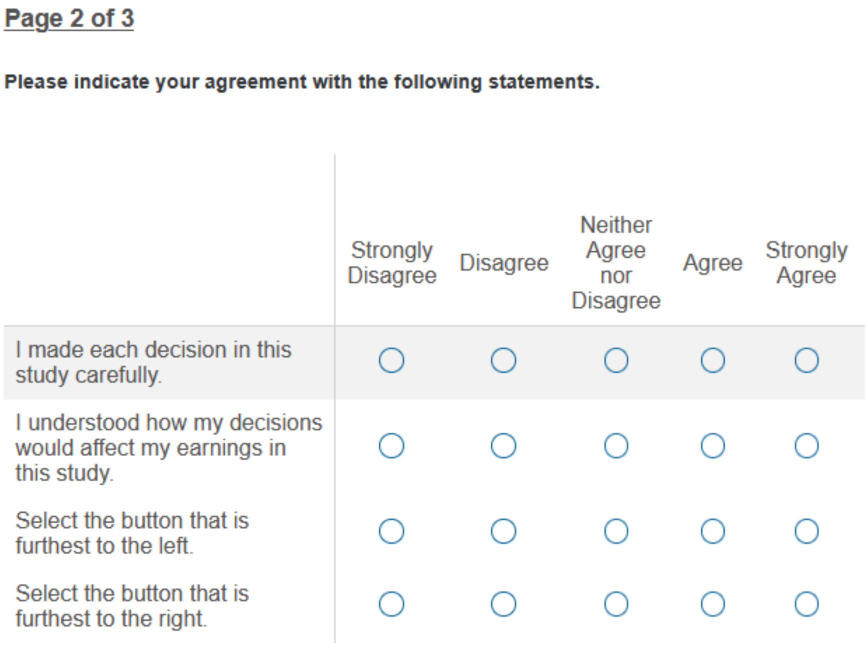
Page 1 of 3

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 G.8: Classic Attention Check

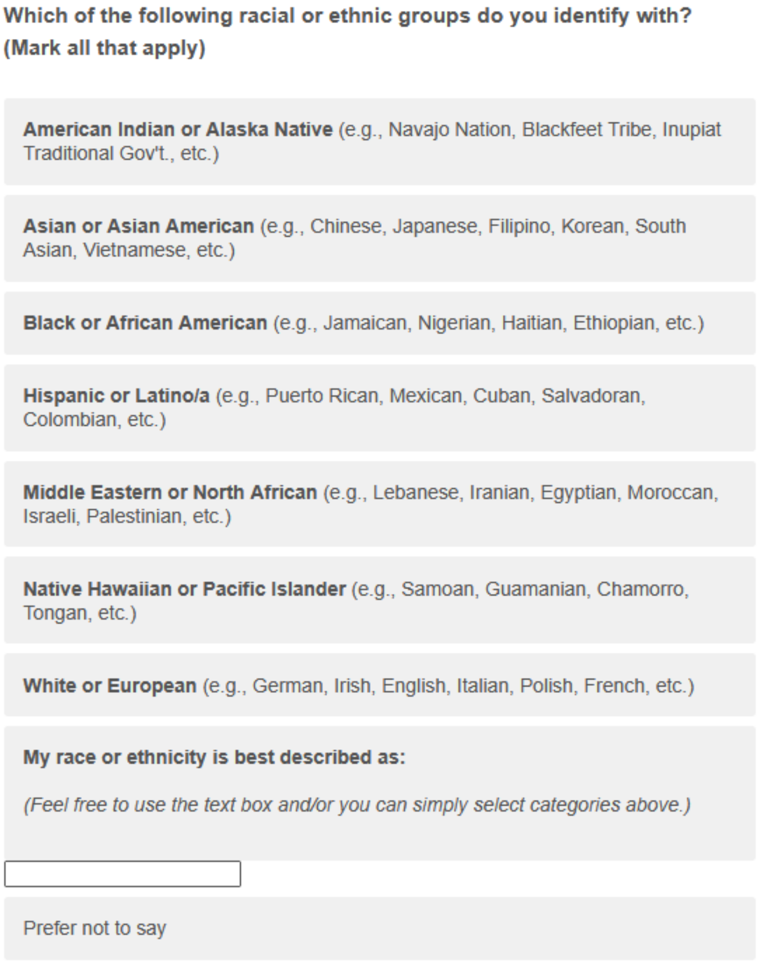


Page 2 of 3

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 G.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.)

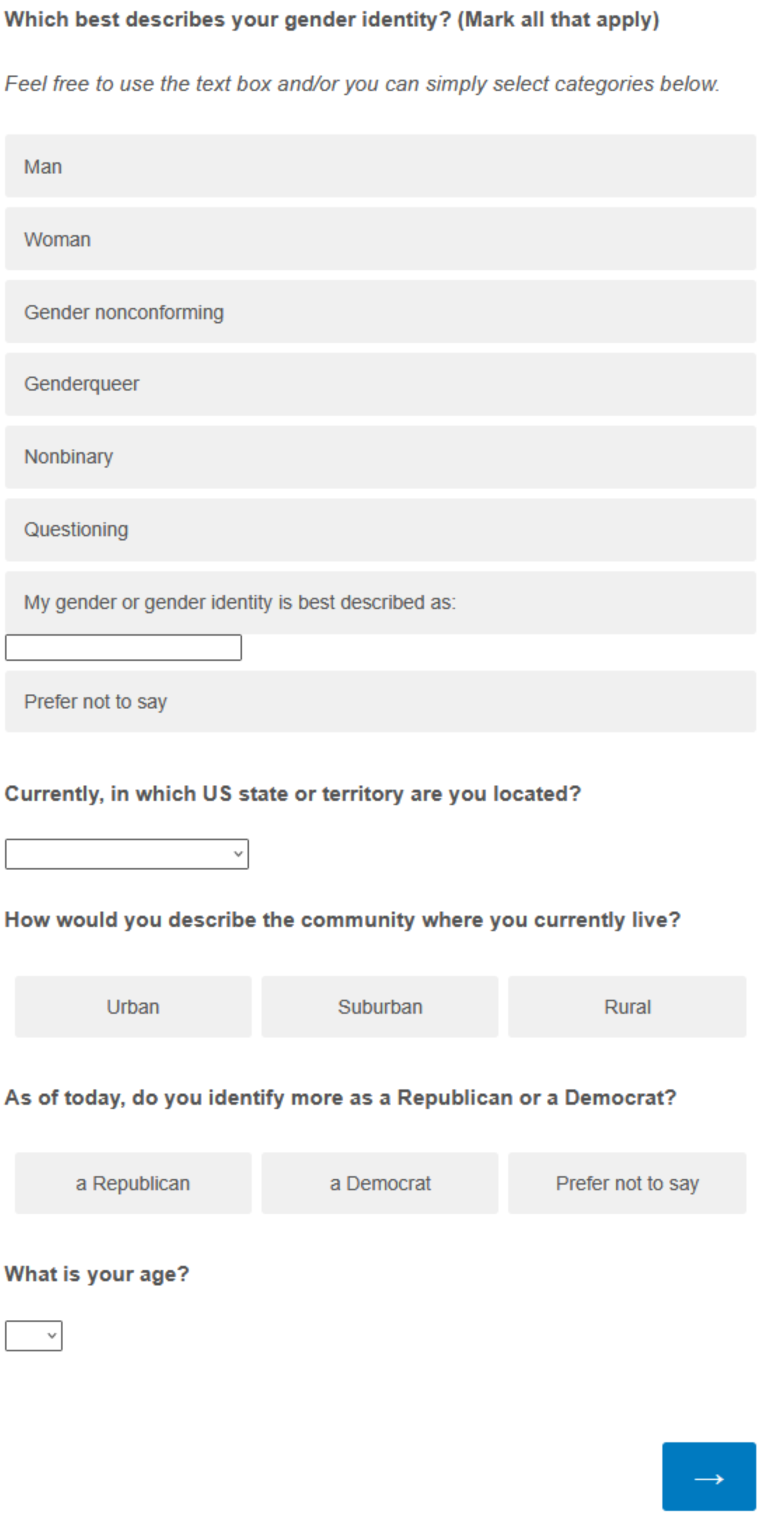
White or European (e.g., German, Irish, English, Italian, Polish, French, etc.)

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 G.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?

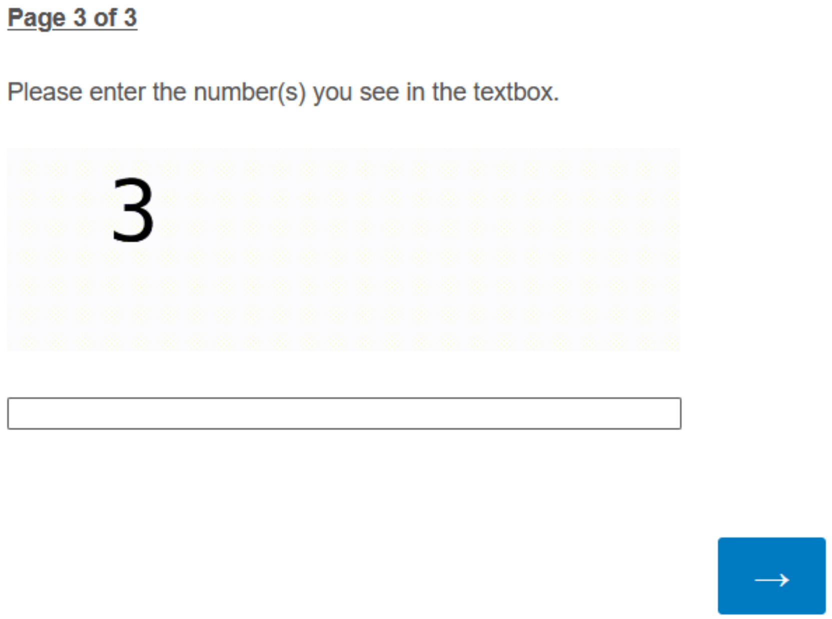
Urban | Suburban | Rural

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 G.11: Video Attention Check



Page 3 of 3

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

Note: You can watch the video here.