Longitudinal Loyalty: Understanding The Barriers To Running Longitudinal Studies On Crowdsourcing Platforms

Longitudinal Loyalty: Understanding The Barriers To Running Longitudinal Studies On Crowdsourcing Platforms

MICHAEL SOPRANO, Department of Mathematics, Computer Science and Physics, University of Udine, Udine, Italy
KEVIN ROITERO, Department of Mathematics, Computer Science and Physics, University of Udine, Udine, Italy
UJWAL GADIRAJU, Delft University of Technology, Delft, Netherlands
EDDY MADDALENA, Department of Mathematics, Computer Science and Physics, University of Udine, Udine, Italy
GIANLUCA DEMARTINI, The University of Queensland, Saint Lucia, Australia

Crowdsourcing tasks have been widely used to collect a large number of human labels at scale. While some of these tasks are deployed by requesters and performed only once by crowd workers, others require the same worker to perform the same task or a variant of it more than once, thus participating in a so-called *longitudinal study*. Despite the prevalence of longitudinal studies in crowdsourcing, there is a limited understanding of factors that influence worker participation in them across different crowdsourcing marketplaces. We present results from a large-scale survey of 300 workers on 3 different micro-task crowdsourcing platforms: Amazon Mechanical Turk, Prolific and Toloka. The aim is to understand how longitudinal studies are performed using crowdsourcing. We collect answers about 547 experiences and we analyze them both quantitatively and qualitatively. We synthesize 17 take-home messages about longitudinal studies together with 8 recommendations for task requesters and 5 best practices for crowdsourcing platforms to adequately conduct and support such kinds of studies. We release the survey and the data at: https://osf.io/h4du9/.

CCS Concepts: • **General and reference → Surveys and overviews**; **Empirical studies**; • **Social and professional topics → User characteristics**; • **Human-centered computing → Collaborative and social computing systems and tools**; • **Software and its engineering → Designing software**.

Additional Key Words and Phrases: Longitudinal Studies, Crowdsourcing Platforms, Surveys, Online Sampling, Amazon Mechanical Turk, Prolific, Toloka

1 Introduction

In recent years, micro-task crowdsourcing has become a popular method for collecting human labels on a large scale. Typically, platforms host the tasks to be performed. These tasks are then allocated to a crowd of workers in a first-come, first-served approach. However, requesters sometimes need to conduct studies that require a specific worker to perform new chunks of work over multiple days, weeks, or even months — namely longitudinal studies.

Authors' Contact Information: Michael Soprano, Department of Mathematics, Computer Science and Physics, University of Udine, Udine, Italy; e-mail: michael.soprano@uniud.it; Kevin Roitero, Department of Mathematics, Computer Science and Physics, University of Udine, Udine, Italy; e-mail: kevin.roitero@uniud.it; Ujwal Gadiraju, Delft University of Technology, Delft, Zuid-Holland, Netherlands; e-mail: u.k.gadiraju@tudelft.nl; Eddy Maddalena, Department of Mathematics, Computer Science and Physics, University of Udine, Udine, Italy; e-mail: eddy.maddalena@uniud.it; Gianluca Demartini, The University of Queensland, Saint Lucia, Queensland, Australia; e-mail: demartini@acm.org.

ACM 2469-7826/2024/7-ART
https://doi.org/10.1145/3674884

Longitudinal studies aim to observe changes that may occur with respect to a chosen subject over a given or extended period of time.

Longitudinal studies can be defined as a series of single, self-contained virtual work unit allocated to and performed by a worker from the same requester which are published regularly over time and require the same workers to participate. A longitudinal study consists of a collection of subsequent sessions, each with a temporal delay between them. A session encompasses the entire set of virtual work units allocated to workers.

Running longitudinal studies on crowdsourcing platforms has become popular, as evidenced by Litman et al. [47], who introduced a tool for longitudinal study functions on top of Amazon Mechanical Turk. This popularity is largely attributed to the convenience and accessibility that crowdsourcing platforms offer for accessing potential study participants.

Despite the growing popularity of crowdsourcing-based research over traditional lab studies [28], there is limited understanding and several open questions around how workers perceive longitudinal studies. What motivates or deters worker participation in longitudinal studies? Why do workers drop out? Can insights from worker experiences enhance such studies? How can platforms better support longitudinal research?

In this paper, we address the aforementioned research gap by presenting results from a large-scale survey on online longitudinal studies. We surveyed workers from three platforms: Amazon Mechanical Turk, Prolific, and Toloka, aiming to understand their experiences and expectations. We recruited 300 (100 from each platform) who reported on 547 previous experiences, answering questions about their perception of such studies and factors influencing their participation in future studies. We analyzed their responses using a mixed-methods approach.

Our results show that workers with experience in longitudinal studies are readily available on platforms like Prolific, where studies typically have more sessions. Most reported experiences occurred within a year before the survey. Sessions usually lasted up to 2 hours, with intervals of 1 to 30 days between them. Partial rewards motivate workers, with monetary incentives being key. Most workers complete and wish to continue such studies, though commitment and insufficient rewards limit availability. On average, workers commit to 21 days of daily 15-minute sessions or 103-minute sessions. They prefer daily to weekly participation, allocating about 2.7 hours daily and suggesting $10.75 as acceptable hourly payment. Incentives for participating in new studies focus on rewards. Study length influences participation decisions. Benefits include increased productivity, but downsides include long-term commitment and inflexible rewards.

2 Related Work

We start by summarizing in Section 2.1 the studies that looked at the crowd worker experience. We focus also on those that address current barriers to a fruitful experience and propose tools and methods aiming at improving it. Then, in Section 2.2 we discuss previous work that has conducted longitudinal studies over crowdsourcing platforms and report their approaches and strategies.

2.1 Exploring And Improving Crowd Worker Experiences

Previous work has looked extensively at workers' needs and experience on crowdsourcing platforms wherein workers receive monetary compensation for successfully completing a micro-task [27].

Wu and Quinn [83] examined the impact of task design choices on worker experience and performance, while Hettiachchi et al. [36] studied task assignment methods that address plurality problems. Nouri et al. [53, 54] highlighted the importance of clear instructions and proposed computational tools to assist task requesters in designing clear tasks. Irani and Silberman [41] and Williams et al. [82] investigated the impact of using tools to support crowd work, demonstrating how they introduce task switching and multitasking while improving productivity. Another approach to enhancing crowd work experience is through coaching by fellow workers, as described by Chiang et al. [13]. Previous studies have suggested the concept of conversational crowdsourcing,

utilizing worker avatars and metaphors intelligently to enhance worker engagement and improve their overall experience [18, 42, 64, 65].

There have been several efforts to empower crowd workers and support their work experiences to overcome challenges related to fair wages, power asymmetry, and unfair rejections that have plagued different crowdsourcing marketplaces [22, 26]. Reputation systems have been proposed to help propagate high-quality work and safeguard worker interests [29]. Self-organization has been suggested to help crowd workers obtain stronger negotiation power with platforms and requesters [71].

Related to their experience and earnings, Hara et al. [34] adopted a quantitative lens to analyze earnings on crowdsourcing platforms, showing how workers are underpaid on average. Cantarella and Strozzi [11] explored the differences between the earnings of crowd workers based in Europe and the United States. Whiting et al. [81] proposed a method to ensure fair pay for workers on Amazon Mechanical Turk. Fan et al. [23] proposed a reward mechanism that allows workers to share these risks and rewards and achieve a standardized hourly wage equally split for all participating workers within cooperatives. Varanasi et al. [79] discussed the difficulties faced by low-income Indian women through a qualitative study. Toxtli et al. [78] analyzed the time spent by workers on non-rewarded activities, which further decrease hourly wages. Durward et al. [21] addressed both the nature of the task performed and the financial compensation from the worker's perspective.

Other individual and social factors influence workers' attitudes and behavior. Abbas and Gadiraju [1] explored the goal-setting practices of crowd workers on Amazon Mechanical Turk and Prolific and highlighted the challenges that workers face. Fulker and Riedl [24] focused on exploring factors that lead crowd workers to cooperative efforts towards completing the task, while Pfeiffer and Kawalec [62] study justice expectations of workers involved in different types of crowdsourcing platforms, showing that they perceive injustices in four areas: planning insecurity, lack of transparency in performance evaluation, lack of clarity in task instructions, and low remuneration.

Compared to this existing body of research, we address the crowd worker experience within longitudinal studies, which require sustained commitment compared to standard micro-tasks. We offer guidelines and recommendations for task designers and requesters on *how to design tasks and engage workers effectively in longitudinal studies on crowdsourcing platforms.*

2.2 Longitudinal Studies On Crowdsourcing Platforms

The original definition of longitudinal study [12] has been proposed in the past by researchers in the fields of psychology and medicine. Bauer [7] described various types of longitudinal designs along with practical considerations on how to conduct them.

Ployhart and Ward [63] proposed and answered a list of 12 questions that typically researchers must address when designing and conducting longitudinal studies. More recently, researchers ran longitudinal studies on crowdsourcing platforms, within different fields of study.

2.2.1 Perspectives And Fields Of Study. The research community has focused from a longitudinal perspective, for instance, on (mis)information assessment. Roitero et al. [68, 69] run a truthfulness labeling task repeated four times at a distance of one month each inviting both new and previously participating workers. They observe that returning workers spend more time on the task as compared to fresh workers who have not done the task before. Fan et al. [23] repeated the same crowdsourcing task multiple times inviting the same group of participating workers each day for 20 days observing a sharp decline in return rates over time. Mensio et al. [49] propose a tool for the longitudinal assessment of the misinformation shared by Twitter accounts.

Longitudinal studies often address health-related issues and challenges. Strickland and Stoops [75] conducted a study on alcohol use, using a weekly survey over 18 weeks. The study involved an initial task that took 21 minutes to complete, followed by regular 2-minute follow-up tasks. High response rates (64.1%-86.8%) were

observed across the 18 weeks. Active participation was incentivized through entry into a raffle for one of five $50 bonuses if participants completed 14 or more weekly surveys.

Mishra and Carleton [50] describe a study aimed at gathering data on gambling-related behaviors, tendencies, and traits. They conducted three crowdsourcing experiments and a fourth two-wave longitudinal study, which included 13.5% of Study 1 participants and 14.8% of Study 2 participants. This longitudinal study demonstrated acceptable test-retest reliability for the identified problem. Similarly, Brooks and Clark [9] conducted a longitudinal study involving 636 young adults to investigate the gambling-related issue of loot boxes in video games.

Strickland and Stoops [76] provide an overview of using Amazon Mechanical Turk to conduct longitudinal studies for addiction science. They show a fourfold increase in the number of papers utilizing this platform for participant recruitment from 2014 to 2017. Goodwin et al. [30], on the other hand, examine the potential of Reddit as a recruitment strategy for addiction science research, arguing that it could be useful for conducting longitudinal follow-up surveys.

Ogata et al. [55] explore the relationship between domestic pets and their owners during the COVID-19 pandemic through a four-staged longitudinal study involving 4,237 workers. In a related context, Dayton et al. [17] investigate testing hesitancy and disclosure stigma in a four-wave study with 355 workers, while Dang et al. [16] study COVID-19's progression characteristics and recovery patterns by collecting audio samples from 212 individuals. Mun et al. [52] conducted a two-year longitudinal study on 1453 adults with chronic pain, surveying them three times to explore pain severity, interference, emotional distress, and opioid misuse during the pandemic. Additionally, Mun et al. [51] investigated the impact of insomnia severity and evening chronotype on chronic pain in 884 adults over 21 months. They found that insomnia may be a stronger predictor of changes in pain and emotional distress.

The literature review by Cho et al. [14] examines crowdsourcing-based approaches in ophthalmology, analyzing 17 longitudinal studies. Schober et al. [72] investigates pollen allergies through a longitudinal study, analyzing approximately 25,000 crowdsourced search queries from citizens spanning 2017 to 2020. Rajamani et al. [67] utilize a longitudinal crowdsourcing approach to gather ideas and feedback for enhancing electronic health record systems, collecting 294 responses between 2019 and 2022.

2.2.2 Human Factors And Participation Dynamics. Other researchers address human-related aspects while employing longitudinal-based crowdsourcing approaches. Daly and Nataraajan [15] conducted three studies. The first focused on a two-month re-response rate among a US Amazon Mechanical Turk sample (n = 752; 75%). The second study (n = 373) explored four- and eight-month re-response rates among US immigrants (56% and 38%, respectively). The third study examined a thirteen-month re-response rate (47%), all involving a 23-minute task.

Qiu et al. [66] explored human memorability in the context of information retrieval on the web in a longitudinal study spanning 2 sessions across 7 days with at least a 3 day gap between the two sessions. The authors recruited participants from Amazon Mechanical Turk, and measured knowledge gain and long-term memorability of participants in their study. Tolmeijer et al. [77] investigated trust development in a house recommendation system through a Prolific study spanning three sessions within a week. Initially, 255 workers participated, with 83% returning for the second session two days later. Of those, 96% completed the third session, resulting in 203 participants who finished all three sessions, representing a nearly 80% retention rate throughout the study. Li et al. [46] conduct a large-scale longitudinal study about recruitment and retention in remote research. They recruit 10,000 workers across two phases, gathering 12 weeks of daily surveys and passive smartphone data, resulting in 330,000 days (equivalent to 900 years) of observation.

Wang et al. [80] introduce a two-week game with a purpose. Through longitudinal studies, they examine individuals' experiences with hedonic and social factors in early stages and expand to include hedonic, social, and usability-related factors in later stages. Leung et al. [45] surveyed 1000 Amazon Mechanical Turk workers

to uncover factors influencing continued participation. Their findings highlight two main triggers: external regulation, such as monetary rewards, and workers' intrinsic motivation.

Grant et al. [31] explore fairness in crowdsourcing through two theoretical lenses: organizational justice and institutional logic. They conduct a longitudinal netnographic study to understand workers' perceptions of fairness.

Aljohani and Jones [4] present initial findings from recruiting qualified yet anonymous workers for hacking experiments involving defensive cyber deception. These experiments are part of a longitudinal study examining malicious cybersecurity experiments on crowdsourcing platforms [3].

Gurung et al. [32] designed a crowdsourcing platform for a longitudinal study analyzing incorrect answers from 2015-2020 academic years across two mathematics courses, aiming to understand how to enhance student learning through remediation.

Sometimes, the specific (micro-task) commercial crowdsourcing platform chosen can hamper the overall worker experience. Peer et al. [60] show that Amazon Mechanical Turk shows a lower population replenishment rate and tends to have more dishonest workers compared to platforms like Prolific. In a subsequent study, Peer et al. [61] highlight Prolific's data quality across various measures relevant to behavioral research. Given the relevance of longitudinal studies to behavioral research [50, 75, 76], platform choice becomes a crucial consideration.

Hata et al. [35] analyzed longitudinal crowdsourcing platform data and found that work quality remains stable over time for the same worker, suggesting that long-term work quality can be predicted after the first five tasks. Additionally, Huang et al. [40] explored the motivations behind continued participation of crowd workers in crowd logistics platforms, confirming the importance of monetary incentives as well as workers' trust and cooperation.

2.2.3 Retention Rates And Strategies. Retention rates of workers vary significantly across longitudinal studies and decrease as time passes [10, 37, 44, 52, 73], starting from the 80% obtained by Shapiro et al. [73] after a week to the 56% over an year obtained by Mun et al. [52].

Various studies used different reward schemes and incentives to increase retention rates, with strategies predominantly revolving around payment schemes. A common approach involves incentivizing worker retention through supplementary payments.

Difallah et al. [19] show that offering a bonus to achieve a milestone is the most effective method for retaining workers up to a predefined milestone within a continuous series of tasks with no interruptions. Auer et al. [6] compared traditional work to crowd work in longitudinal studies regarding performance payment effects. They found no difference in performance but emphasized the importance of ethically rewarding workers due to their limited bargaining power. Pay significantly affects attrition (i.e., single task abandonment) but not retention in the second wave of longitudinal studies.

Benbunan-Fich [8] investigates the question of whether workers who quit a study before its completion should receive monetary compensation. They propose that determining an appropriate partial payment, especially for longitudinal studies, involves complex considerations beyond simple monetary compensation.

3 Aims And Motivations

In Section 3.1 we discuss the novelty of our study concerning other works that address longitudinal studies in crowdsourcing. Then, in Section 3.2 we list the three research questions addressed.

3.1 Research Contribution

Our study aims to address a research gap concerning worker perception in longitudinal studies. While previous research has primarily focused on short-term micro-task crowdsourcing, we provide a comprehensive exploration of longitudinal studies, which remains relatively under-explored. Although researchers have previously proposed

considerations and suggestions for designing and conducting such studies, there has been limited characterization and comprehensive exploration from the worker perspective.

Furthermore, the novelty of our research also lies in the experimental nature of its data. Through surveys conducted across three diverse crowdsourcing platforms, we aimed to capture a broad spectrum of personal experiences and perspectives regarding longitudinal studies. Lastly, our study digs deeply into the specific dynamics of longitudinal studies on crowdsourcing platforms by employing mixed-methods approach. While previous works have explored various aspects of crowdsourcing, our paper aims to provide new insights into the unique challenges and opportunities associated with conducting longitudinal studies.

3.2 Research Questions

Understanding key aspects of longitudinal study design would not only help identify barriers experienced by workers but also provide recommendations for practitioners and researchers conducting such studies on micro-task crowdsourcing platforms. Additionally, this enables the proposal of best practices for platforms supporting longitudinal research.

We remark that our research focuses on those who design and enable longitudinal studies. We base our considerations on both the worker perspective and our past experience as task requesters. The research questions we address are as follows:

RQ1 What is the current workers' perception of longitudinal studies on commercial micro-task crowdsourcing platforms? How did their previous experiences take place? What is workers' opinion about their participation and commitment to future longitudinal studies? Which are their preferred characteristics of a longitudinal study?

RQ2 What are the recommendations that researchers and practitioners who want to design and conduct longitudinal studies over commercial micro-task crowdsourcing platforms should follow?

RQ3 What are the best practices that commercial micro-task crowdsourcing platforms should employ to enable conducting longitudinal studies effectively and improve their support for such kind of studies in general?

4 Terminology

In this paper, we employ a specific set of nouns and technical terms that belong to the field of crowdsourcing. For readers' convenience, we provide a list of terms below that we will refer to throughout the paper, integrating the initial definition provided in Section 1. Some of these definitions were originally proposed by Howe [38] and Paolacci et al. [59], which we have further expanded upon.

- *Crowdsourcing*: the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. In the rest of this work, the term crowdsourcing refers to *microtask crowdsourcing*.
- *Platforms*: commercial micro-tasks marketplaces that allows individuals and businesses to outsource their processes and jobs to a distributed workforce who can perform these tasks virtually.
- *Human Intelligence Task (HIT)*: a single, self-contained, virtual work unit allocated to and performed by an individual.
- *Element*: item that a individual evaluates, uses, addresses within a HIT. A Human Intelligence Task is composed of a set of elements.
- *Batch*: a set composed of multiple HITs published by a single individual.
- *Requester*: an employer who recruits employees (usually called *workers* or *participants)* from an online labor marketplace for the execution of HITs in exchange for a wage (usually called *reward*).
- *Worker*: an individual who joins a crowdsourcing platform to perform and complete HITs published by requesters.

- *Session*: the whole set of HITs available, allocated to the same group of workers within a certain timespan.
- *Interval Between Sessions*: the time that elapses between the completion of a session and the beginning of the following one.
- *Session Duration*: time taken by a worker to complete a session.
- *Longitudinal Study (LS)*: a series of HITs from the same requester which are published regularly over time and require the same workers to participate. A longitudinal study is made of a collection of subsequent sessions, with some temporal delay between them. We thus define two more terms specific to the LS:
  - *Duration* (of the LS): the length of time required to complete a longitudinal study: from the beginning of the first session to the completion of the last one, including all the intervals.
  - *Frequency* (of the LS): the number of sessions that a longitudinal study requires a worker to complete over a timespan.

5 Methodology

We design a survey to characterize longitudinal studies from the perspective of crowd workers (Section 5.1) and we collect responses by conducting a crowdsourcing task (Section 5.2) on three popular commercial crowdsourcing platforms, namely Amazon Mechanical Turk,[1] Prolific [58],[2] and Toloka.[3] We analyze the responses collected by using quantitative and qualitative approaches (Section 5.3) and we perform statistical significance tests (Section 5.4).

The complete survey, along with the answers provided by workers and the dataset related to both quantitative and qualitative analysis of the answers, has been released and is available at: https://osf.io/h4du9/. The qualitative part presents a thematic analysis and includes a complete description of the coding scheme, codes, and themes. The whole survey is reported also in Appendix A.

5.1 Survey Design

The survey consists of two parts: P1 and P2. The first part of the survey (P1), reported in Appendix A.1, aims to explore the current perception of longitudinal studies in crowdsourcing. It focuses on workers' prior experience, the perceived suitability of platforms for hosting longitudinal studies, possible reasons limiting the popularity of longitudinal crowdsourcing studies and their availability on crowdsourcing platforms.

The second part (P2), reported in Appendix A.2, on the other hand, investigates workers' thoughts, opinions, and ideas about the design of, and their underlying motivations to participate in future longitudinal studies.

More specifically, the survey comprises 16 multiple-choice questions, 4 text-based questions (i.e., questions requiring a mandatory textual answer), and 6 numerical questions. Additionally, there are 11 questions that allowed workers to provide custom free-text responses to elaborate on their answers. Among the multiple-choice questions, 9 of them are implemented using radio buttons, as only a single answer was possible. In contrast, checkboxes are employed for the remaining 7 questions, as they allow for multiple answers, thus permitting a broader range of responses. The naming convention reported in Appendix A is used throughout the rest of this paper.

5.2 The Crowdsourcing Task

We designed and run the crowdsourcing task using Crowd_Frame,[4] a framework developed by Soprano et al. [74] which allows for setting up and deploy crowdsourcing experiments easily.

[1] https://www.mturk.com/
[2] https://www.prolific.co/
[3] https://toloka.ai/
[4] https://github.com/Miccighel/Crowd_Frame

The crowdsourcing task aimed to recruit 300 workers from three platforms: Amazon Mechanical Turk, Prolific, and Toloka, with 100 participants from each. Participation criteria required completing at least 4000 tasks on Amazon Mechanical Turk and 2000 tasks on Prolific. On Toloka, participants were directly asked about their prior experiences with longitudinal studies. Recruitment continued on each platform until 100 participants with at least one previous longitudinal study experience were obtained.

We initially recruited 50 workers from each of the three platforms. However, after analysis, we found that only a portion had previous experience. Therefore, we repeated the recruitment process four times on each platform until we obtained a total of 300 workers with at least one previous experience in longitudinal studies. This required 729 workers in total to successfully complete the task: 153 from Amazon Mechanical Turk, 160 from Prolific, and 412 from Toloka. This means that, for instance, on Amazon Mechanical Turk, we found the required 100 workers among the 153 recruited.

On Amazon Mechanical Turk, the task was published during the following periods: April 14-15, 2022; August 29-September 1, 2022; September 12, 2022; and March 10-13, 2023. On Prolific, the periods were: April 14, 2022; September 15, 2022; March 16-17, 2023; and April 11, 2023. On Toloka, the periods were: September 12-15, 2022; March 10, 2023; March 13, 2023; and March 15-17, 2023. In summary, the first iteration of the task was published on Amazon Mechanical Turk on April 14, 2022, while the last one was on April 11, 2023, on Prolific. Throughout the entire period, the task workflow and layout remained unchanged and were continuously available during the specified periods.

The task workflow proceeded as follows: workers were initially provided with general instructions and the study context, which included the definition of longitudinal studies introduced in Section 1. Subsequently, workers were asked to complete the first part of the survey (P1), followed by the second part (P2). In the P1 part, workers were asked to report their experiences with up to three longitudinal studies they had completed. We imposed this limit to ensure a reasonable completion time for the crowdsourcing task.

Each experience was reported and described by responding to a subset of 11-13 questions, with the total number of questions shown depending on the answer provided for question 1.1 (Appendix A.1). Conditional logic was used to determine whether certain sub-questions needed to be asked. Specifically, if a worker reported between $0 \leq X \leq 3$ experiences (denoted as X), the number of questions ranged from $1 + (11 * X) + 2$ to $1 + (13 * X) + 2$, as the block of questions 1.1.X was repeated X times, once for each experience. Additionally, only one question from either 1.1.X.9.1 or 1.1.X.9.2 was shown, depending on the answer provided for question 1.1.X.9. Conversely, the P2 part comprised 11 questions, asked only once. Thus, the total number of questions in the entire survey ranged from $1 + (11 * X) + 13$ to $1 + (13 * X) + 13$.

After completing P1 and P2, workers could submit their responses and receive payment. They also had the opportunity to provide final comments. To ensure response quality, a criterion required workers to spend a minimum of 3 seconds on each question. Workers received $2 USD for their participation, based on an hourly rate derived from the US minimum wage and task completion time. The median reward ranged from $10-13 per hour, with an average completion time of 700 seconds, a standard deviation of 593, and a median of 548 seconds.

5.3 Analysis Of Workers' Responses

We address each survey question from a quantitative or qualitative viewpoint, depending on the question type.

Initially, we provide some general remarks concerning the results obtained (Section 5.3.1). Then, we focus specifically on the quantitative analysis (Section 5.3.2) and on the qualitative approach we follow (Section 5.3.3).

5.3.1 General Remarks. To interpret our results correctly, it should be noted that some survey questions required multiple responses based on workers' past experiences with longitudinal studies, while others required only a single response, as described in Section 5.2.

Most questions in the P1 part require answers for each past experience, while questions in the P2 part and one question from the P1 part require only a single answer. Recruiting 300 workers, the maximum number of answers in the former case is 900 (assuming three experiences per worker). In the latter case, the maximum is 300. The results (Section 6.1.1) show that the number of reported experiences is 547.

In result analysis, we often break down results by the platform used to recruit workers who answered the survey. For instance, a worker recruited from Amazon Mechanical Turk but participating in a longitudinal study on Prolific would be included in the Amazon Mechanical Turk breakdown.

5.3.2 Quantitative Analysis. We use bar charts for closed-ended multiple-choice questions and univariate distribution charts for numerical questions in our quantitative analysis. Results are broken down by crowdsourcing platform to highlight differences visually. A color scheme (blue for Amazon Mechanical Turk, orange for Prolific, and green for Toloka) is introduced in the legend of the first figure and consistently applied in subsequent figures to prevent repetition and information overload.

In our bar charts, the x-axis shows available answers, and the second row shows their relative frequencies across platforms. The y-axis represents answer frequencies, with absolute frequencies displayed above each bar. Total absolute frequencies equal 547 or 300 based on question requirements, denoted with E or W, respectively. These values are shown in the chart's lower left corner. If a question allows for providing non-mutually-exclusive answers, the top chart is marked with A. Additionally, total answers and experiences/workers are reported in the lower left corner of the chart.

In univariate distribution charts, the y-axis represents the probability density function for three continuous random variables, representing the answers provided by workers across each considered crowdsourcing platform. The x-axis ranges from the minimum value to a cutoff, filtering out outliers. Dashed lines indicate mean values for each platform, using the established color scheme. Total data used is reported in the lower left corner, marked with a corresponding letter. In some cases, outliers are filtered out, noted by an additional label beneath the data count.

5.3.3 Qualitative Analysis. We used a conventional qualitative content analysis approach [39] to analyze open-ended responses. This inductive method describes phenomena with limited existing research or theory, unlike deductive qualitative analysis, which relies on predetermined themes from literature.

Two authors of this paper act as expert researchers, reviewing all responses to the open-ended mandatory questions and those allowing free-text input. For each response, they create a custom "code" by highlighting key phrases capturing significant insights using a predefined keyword. For instance, if a worker mentions participating in the longitudinal study because it was interesting and provided self-discovery, the initial code chosen by the authors might be the keyword *task_interest*. As analysis progresses, multiple core concepts emerge, forming the foundation of the initial overall coding scheme.

The qualitative analysis phase involved merging initially identified codes based on their inter-dependencies through multiple iterations and discussions. For instance, codes like *task_interest*, *task_payment*, and *task_easiness* were merged into the overarching theme of task_features. This process led to the emergence of seven themes, detailed in Table 1, with sample answers and initial codes. Due to expert involvement and iterative refinement, internal agreement is not reported here; interested readers can refer to McDonald et al. [48].

Table 2 details the distribution of additional free texts provided by workers while responding to each mandatory non-text-based question. For P1 part questions, the table reports both the total number of experiences with text and the number of workers providing it. For P2 part questions, only the latter is provided since workers are asked once per question. These texts augment the thematic analysis, offering additional insights to the quantitative analysis of provided answers.

Finally, it is important to note that if a question, such as question 1.1.X.9.2, is not included in the result analysis, it is because it specifically required text-based answers, and unfortunately, the workers did not provide any useful responses for analysis.

5.4 Statistical Testing

We conducted statistical significance tests on the survey responses for closed-ended question to investigate relationships across variables of interest.

In the following, we describe the approach followed for each type of such questions, beginning with those that required a numerical answer (Section 5.4.1), then moving to those that required choosing a mutually-exclusive answer (Section 5.4.2), and finally addressing those that required selecting a non-mutually-exclusive one (Section 5.4.3).

5.4.1 Numerical Answers. In the six cases where the answer provided by the workers was numeric, such as for question 1.1.X.1 of the P1 part (Section 6.1.2), we used ANOVA [56] to determine if there was a statistically significant difference ($p < 0.05$) between the means of the groups.

Specifically, we used a one-way ANOVA to compare the means of the three groups of workers (i.e. Amazon Mechanical Turk, Prolific, and Toloka). In the cases where we found a statistically significant difference at the $p < 0.05$ level, we performed posthoc tests using Tukey's HSD method [2] to determine which groups differed significantly from each other. Tukey's HSD is a multiple comparison test that controls for Type I error rate by adjusting the significance level based on the number of pairwise comparisons.

5.4.2 Mutually-Exclusive Answers. For the nine closed-ended questions, which required choosing a mutually-exclusive answer from a predefined set, such as question 1.1.X.5 of the P1 part (Section 6.1.6), we used chi-squared tests to determine if there were statistically significant differences between the groups.

Specifically, we calculated the observed contingency table of frequencies and used the chi-squared test to compare it to the expected contingency table under the null hypothesis of no difference between the groups. To account for and correct multiple comparisons, we used the false discovery rate (FDR) correction [70], which controls the expected proportion of false discoveries among the rejected null hypotheses. If we encountered zero expected frequencies while performing the chi-squared test, we excluded the comparison from the analysis.

5.4.3 Non-Mutually-Exclusive Answers. Similarly to the previous case, for the seven questions that allowed choosing multiple non-mutually-exclusive answers, such as question 7 of the P2 part (Section 6.1.21), we also used chi-squared tests to determine if there were statistically significant differences between the groups.

However, differently from the previous case, we had to address the situation where a respondent could select multiple options, resulting in overlapping categories. To accommodate this, we calculated the observed contingency table of frequencies using a modified approach that allowed for overlapping categories. Then, we then employed the chi-squared test and FDR correction, as in the previous case, to determine if there were significant differences between the groups.

6 Results

We analyze in Section 6.1 the answers provided for the questions of each survey part (RQ1). Then, in Section 6.2, we provide recommendations for practitioners and researchers who want to conduct longitudinal studies based on our study's insights (RQ2). Finally, in Section 6.3, we outline the best practices for crowdsourcing platforms to facilitate similar experiments (RQ3).

Table 1. Themes emerged while reading each text-based answer provided by the workers.

Theme	Description	Sample Answer	Initial Code
task_features	Aspects related to the task to be performed during a given session of the longitudinal study, such as its design, easiness, etc.	"It was easy to complete"	*task_easiness*
worker_features	Aspects related to workers' own beliefs and motivations, their satisfaction after participating in the longitudinal study, etc.	"It gave me the chance to be a part of change and real scientific study and know that my part contributed."	*worker_motivation*
requester_features	Aspects related to the requester who is publishing the longitudinal study, such as reliability, communicativeness, etc.	"Be reliable - offer a reasonable window during which the study can be completed and respond promptly to any messages from participants"	*requester_reliability*
ls_features	Aspects related to the longitudinal study as a whole, such as session scheduling, reward mechanism, etc.	"Performance rewards are a good way to maintain interest, as it feels like your time and effort are being rewarded"	*ls_progress*
platform_features	Aspects related to the crowdsourcing platform on which the longitudinal study is conducted such as its features, interface, general design, etc.	"Yes. I think there is a large enough pool to pull from and if set up properly and rewarded, people will respond"	*platform_adequacy*
no_suggestion	Answers provided by workers that acknowledge by explaining explicitly that they do not have any additional suggestions.	"Nothing comes to mind"	no_suggestion
answer_useless	Answers that do not convey anything related to the question proposed or that are made of random words and digits.	"Unique crowdsourcing business model"	*answer_useless*

Table 2. Distribution of the additional free texts provided by the workers while answering non-text-based questions.

Part	Section	Question	Experiences	Workers
P1	6.1.4	*Interval Between Sessions*	22 (4.02%)	18 (6.00%)
P1	6.1.5	*Session Duration*	22 (4.02%)	16 (5.33%)
P1	6.1.6	*Crowdsourcing Platform*	33 (6.10%)	26 (8.67%)
P1	6.1.7	*Payment Model*	35 (6.40%)	30 (10.00%)
P1	6.1.10	*Participation Incentives (In Prev. Experiences)*	27 (4.94%)	22 (7.33%)
P1	6.1.14	*Reasons That Limit Availability On Platforms*	48 (8.78%)	48 (7.67%)
P2	6.1.16	*Reasons For Declining Participation*	–	50 (16.67%)
P2	6.1.21	*Participation Incentives (In New Experiences)*	–	17 (5.67%)
P2	6.1.22	*Tasks Type*	–	14 (4.67%)
P2	6.1.24	*Involvement Downsides*	–	23 (7.67%)

6.1 RQ1: Analysis Of Workers' Answers

We begin by analyzing the answers provided by the workers for the P1 part of the survey, from Section 6.1.1 to Section 6.1.14, and those provided for the P2 part, from Section 6.1.15 to Section 6.1.25. We then summarize all our findings in Section 6.1.26.

6.1.1 Previous Experiences. To begin the investigation, we analyzed the previous experiences with longitudinal studies in which each worker reported having taken part, reported in Table 3. We recall that the charts shown in the following figures (Figure 1–Figure 22) should be interpreted as described in Section 5.3.2.

A total of 300 workers were recruited, with each platform contributing 100 workers. They reported 547 previous experiences with longitudinal studies, averaging 1.82 experiences per worker. Prolific workers reported the most experiences (193), followed by Amazon Mechanical Turk (187) and Toloka (167). Prolific had the highest proportion of workers with previous experience (35.28%), followed by Amazon Mechanical Turk (34.19%), while Toloka workers had less experience (30.53%). Additionally, 97 workers (32.3%) reported experiences from a different crowdsourcing platform than their recruitment platform (see also Figure 6).

Table 3. Previous experiences with longitudinal studies reported by the workers recruited.

Platform	Experiences	Percentage	Mean
Amazon Mechanical Turk	187	34.19%	1.85
Prolific	193	35.28%	1.89
Toloka	167	30.53%	1.67
Total	547	100%	1.82

Figure 1 details workers' previous experiences with longitudinal studies from Table 3. The analysis shows that 45% of workers reported one experience, while 27.67% and 27.33% reported two and three experiences, respectively.

These proportions varied across platforms. For Amazon Mechanical Turk, 42% reported one experience, 29% reported two, and 29% reported three. In Prolific, 43% reported one experience, 21% reported two, and 36% reported three. In Toloka, 50% reported one experience, 33% reported two, and 17% reported three. No statistically significant differences were observed across platforms.

The analysis suggests that workers on Prolific are more likely to report multiple previous experiences compared to those on other platforms, validating the recruitment criterion described in Section 5.2. Workers on Amazon Mechanical Turk and Toloka seem accustomed to longitudinal studies, indicating the need for a higher HIT completion threshold to recruit them effectively.

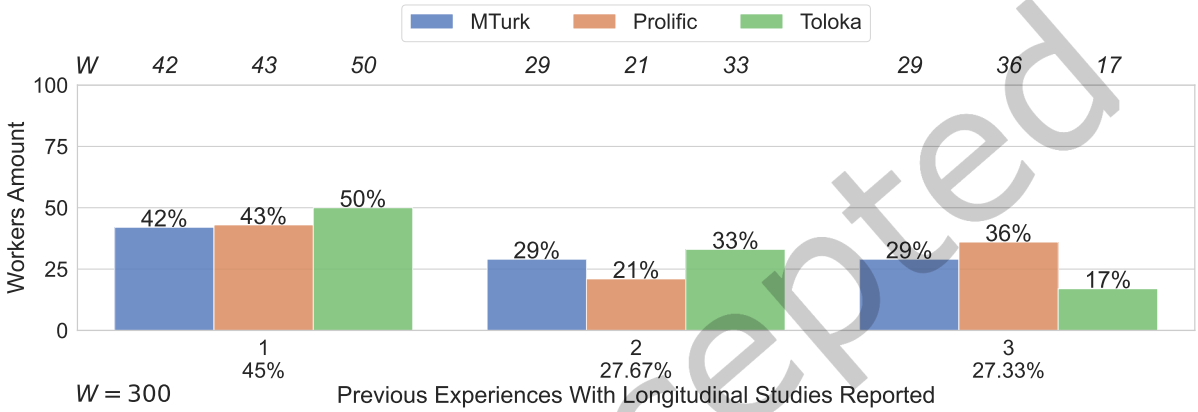

Fig. 1. Number of workers who report 1, 2, or 3 previous experiences with longitudinal studies.

6.1.2 Time Elapsed. Figure 2 describes the time elapsed in terms of months since each previous experience reported, with a particular focus on participation in longitudinal studies that occurred up to 12 months earlier.

The majority of the reported experiences (87%), indeed, occurred within the 12 months preceding participation in the survey, while the remaining 13% occurred earlier. The distribution of participation that took place within the previous year, however, is rather homogeneous, with roughly 13% of participation for each crowdsourcing platform occurring more than 12 months earlier. This indicates that on Amazon Mechanical Turk and Prolific, workers were able to commit to longitudinal studies throughout the whole year before participating in this survey, while on Toloka, the experiences reported have been more recent (Amazon Mechanical Turk vs. Toloka statistically significant, adjusted p-value < 0.05).

6.1.3 Number Of Sessions. Figure 3 details, for each previous experience with longitudinal studies reported, how many sessions composed the overall study referred.

The longitudinal studies in which workers participated on Amazon Mechanical Turk and Toloka have an average of about 6 sessions, while those on Prolific have 7 sessions on average. In general, it appears that task requesters tend to publish slightly longer longitudinal studies on Prolific, although we did not obtain statistically significant comparisons across platforms.

6.1.4 Interval Between Sessions. Figure 4 details the time elapsed, in terms of days, between the sessions of the longitudinal study to which the reported experiences refer, focusing on ranges from 1 day to more than 30 days.

The timespans ranging from 1 day to 9 days, encompass the majority of the longitudinal studies referred to by the reported experiences (63.45%). By extending the considered range up to 30 days, the vast majority of previous

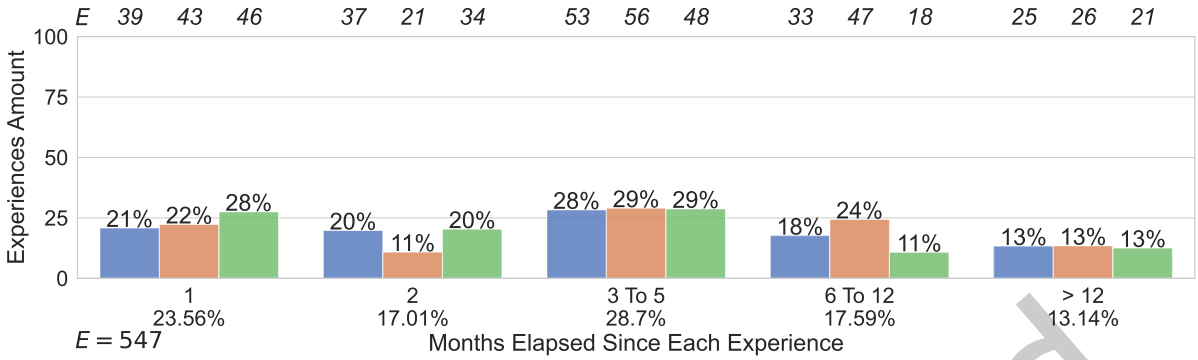

Fig. 2. Time elapsed in months since each previous experience with longitudinal studies reported.

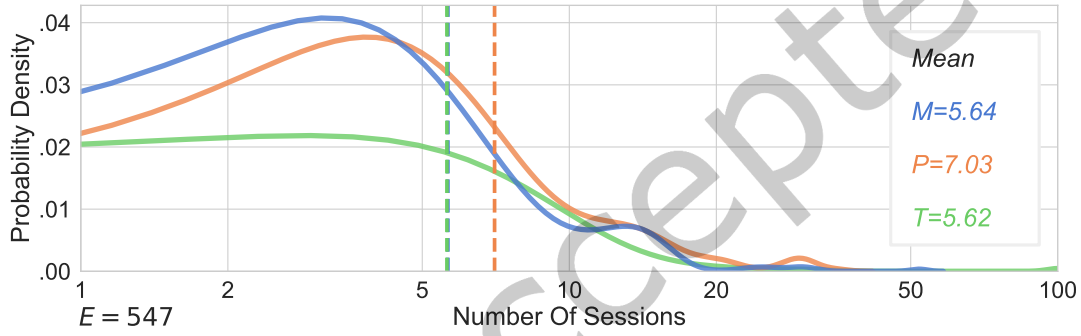

Fig. 3. Number of sessions of the longitudinal study to which each reported experience refers.

experiences (90%) are comprised. Summarizing, most requesters schedule the next session of a study starting from the following day up to a month later, with ten days being the most common timespan (Amazon Mechanical Turk vs. Toloka statistically significant, adjusted p-value < 0.01).

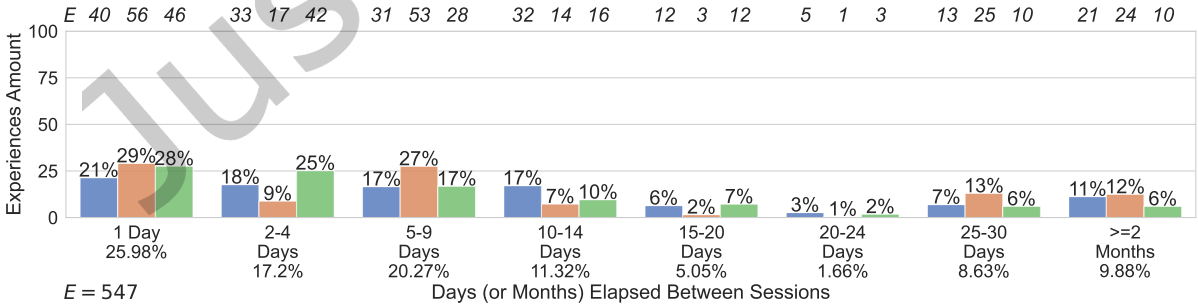

Fig. 4. Time elapsed in days or months between the sessions of the longitudinal study to which each reported experience refers.

6.1.5 Session Duration. Figure 5 details the duration of sessions in the longitudinal study to which the reported experiences refer, measured in minutes or hours.

Almost half of the longitudinal studies had sessions lasting 15 minutes (48.09%), while 22.89% lasted for 30 minutes, 12.72% for 45 minutes, and 12.41% for 60 minutes. The vast majority of sessions, thus, take place within an hour of work (96.11%). There is a small but not negligible number of sessions in longitudinal studies available on Toloka that last for two hours (13), along with 2 sessions on Amazon Mechanical Turk and a single session on Prolific. Furthermore, two workers reported Amazon Mechanical Turk sessions lasting three hours or more.

In general, the vast majority of task requesters on Prolific tend to publish longitudinal studies with shorter sessions, primarily 15 minutes (72%) or 30 minutes (20%), compared to other platforms. The answer distribution is more uniform when comparing Amazon Mechanical Turk and Toloka, although requesters on the latter platform tend to publish studies with longer sessions (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka and Toloka vs. Prolific statistically significant; adjusted p-value < 0.01).

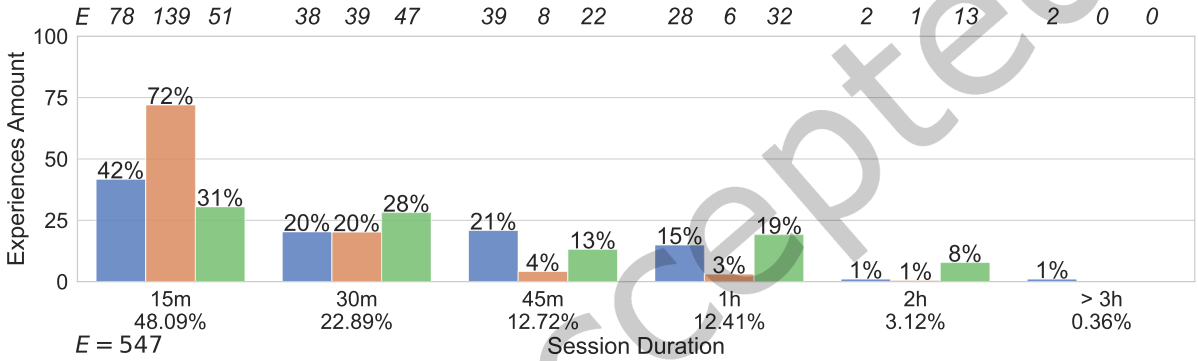

Fig. 5. Duration in minutes or hours of the sessions of the longitudinal study to which each reported experience refers.

6.1.6 Crowdsourcing Platform. Figure 6 describes on which previous experiences with longitudinal studies were conducted, as a worker recruited on a platform might have worked also elsewhere. Roughly the same number of experiences took place on Amazon Mechanical Turk (38.16%) and Prolific (39.47%), while fewer experiences (22.37%) happened on Toloka.

Breaking down the responses by platform, the majority of experiences reported by Amazon Mechanical Turk and Prolific workers occurred on their respective platforms (around 90%). However, there were instances of cross-platform participation: 9% of Amazon Mechanical Turk workers reported experiences on Prolific, while 6% of Prolific workers reported experiences on Amazon Mechanical Turk and 4% on Toloka. Additionally, although experiences reported by Toloka workers primarily occurred on Toloka (63%), a notable portion also occurred on Amazon Mechanical Turk (17%) and Prolific (19%).

Summarizing, the distribution of the collected answers shows that Toloka workers tend to work on other platforms more frequently than those recruited from Amazon Mechanical Turk and Prolific, particularly in the context of longitudinal studies. However, this trend can also be observed on the remaining platforms (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka, and Prolific vs. Toloka are statistically significant with an adjusted p-value < 0.01).

6.1.7 Payment Model. Figure 7 investigates the payment model adopted by the longitudinal studies in which the recruited workers reported participating.

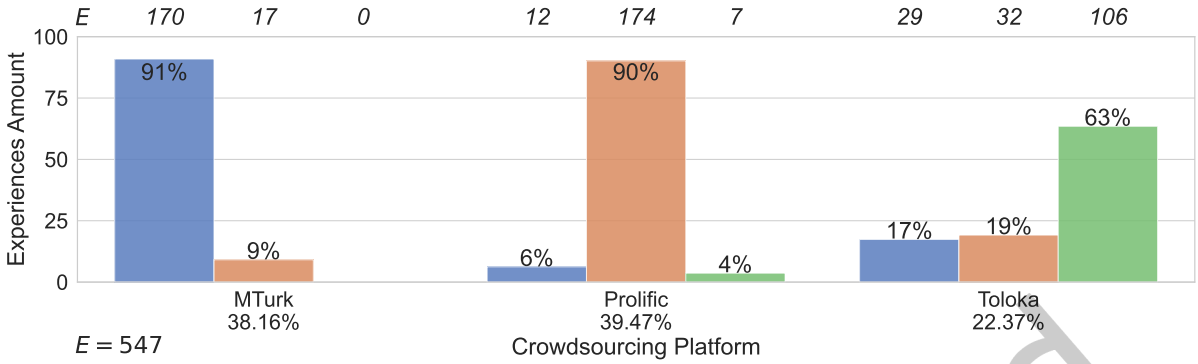

Fig. 6. Crowdsourcing platforms where the longitudinal study to which each reported experience refers took place.

The majority of reported previous experiences (70.31%) involved longitudinal studies where workers were paid after each session, while 21.84% reported experiences with a final reward as the only form of payment. Only 7.84% of the reported experiences described studies relying on a combination of both payment approaches.

The distribution of the answers collected shows that the majority of previous experiences reported were part of longitudinal studies in which the workers were paid after each session, particularly on Amazon Mechanical Turk (75%). Using a final reward is also a viable option, as in 25% of the experiences reported by workers recruited on Prolific and Toloka. Furthermore, 9% of the experiences reported by Amazon Mechanical Turk workers and 7% of those reported on the remaining platforms refer to longitudinal studies that employed a combination of both approaches (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01).

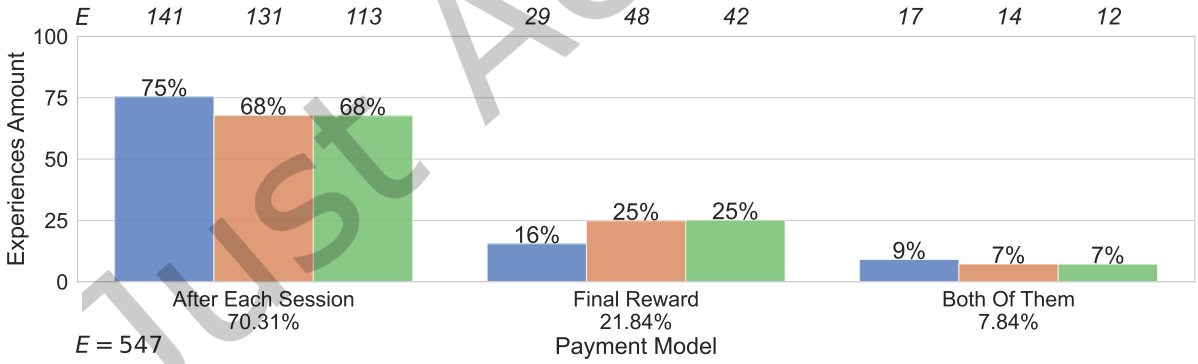

Fig. 7. Payment model of the longitudinal study to which each reported experience refers (i.e., when the reward was provided).

6.1.8 Participation In Same Study. Figure 8 investigates the workers' satisfaction after having participated in the longitudinal study referred to by each reported experience.

The vast majority of workers (91.59%) express their interest in participating again in the same longitudinal study. When breaking down the data across each platform, such opinion is consistent for both Prolific and Toloka workers, with a percentage of positive answers of 98% and 93%, respectively, while it lowers to 83% for Amazon

Mechanical Turk workers (Amazon Mechanical Turk vs. Prolific and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01).

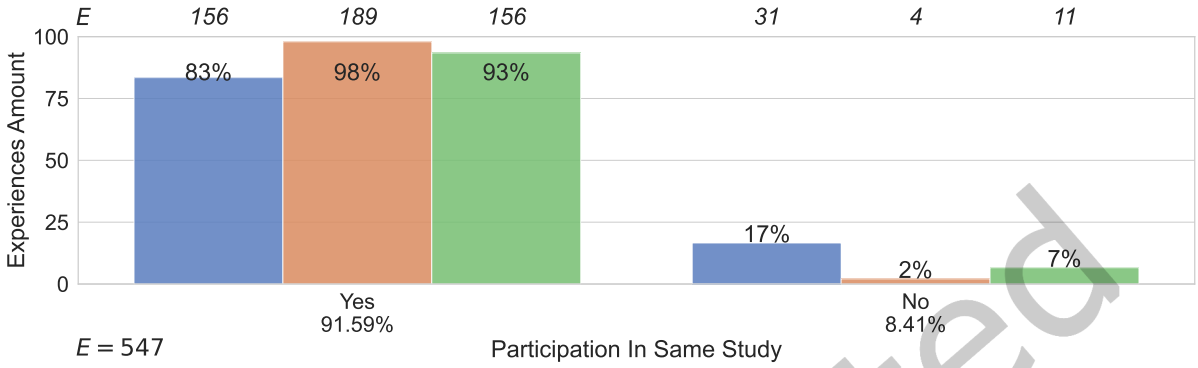

Fig. 8. Workers willingness to participate again in the longitudinal study to which each reported experience refers.

6.1.9 Loyalty And Commitment. The mandatory open question 1.1.X.7.2 (P1 part) asks workers to specify what drove them to return for a second session after completing the first one in the longitudinal study referred to by the reported experience. Also, the workers must explain why they would refuse to participate in the same study altogether.

The workers provided 485 answers among the 547 previous experiences with longitudinal studies reported (88.66%). The distribution of answers collected across different themes is as follows: 272 out of 485 (56.08%) addressed aspects related to the task performed (task_features), while 101 (20.82%) focused on workers' own beliefs and motivations (worker_features). Additionally, 10 (2.06%) were about the longitudinal study as a whole (ls_features), 9 (1.86%) about the requester (requester_features), and 2 (0.41%) about the platform (platform_features). Lastly, 91 (18.76%) answers were deemed unusable (answer_useless). Table 7 (Appendix B) shows a sample of such answers.

The majority of responses (272 out of 485, 56.08%) highlight how task attributes influence their decisions. Some workers find tasks interesting (100 out of 272, 36.76%), easy (54 out of 272, 19.85%), or well-paid (112 out of 272, 41.58%), which motivates their return. Others (15 out of 272, 5%) mention the perceived reliability of securing rewards in subsequent sessions as a driver to return. Many workers (58 out of 272, 41.58%) appreciate the task's agency for expressing their views and getting paid in return. Conversely, issues like low or unfair rewards, worker unavailability during follow-up sessions, or device-specific requirements are common reasons for abandonment or refusal to participate in longitudinal studies after a session. About 20.82% of responses (101 out of 485) come from workers who believe their preferences and attributes influence their decision to return for subsequent sessions in longitudinal studies.

A few workers (4 out of 101, 3.96%) mentioned the sunk costs of completing the first session as a motivating factor to return [5]. Additionally, 45 out of 101 workers (44.55%) expressed satisfaction with completing the initial session, citing the commitment required (12 out of 101, 11.88%), overall involvement, or the chance to gain insights, learn, and develop skills throughout the studies (15 out of 101, 15%).

A small number of workers (9 out of 485, 1.86%) discuss aspects and characteristics of the task requester that impact loyalty and commitment to the longitudinal study. They highlight communication with the requesters and their ability to remind participants of subsequent study sessions as crucial factors. Additionally, 10 workers

out of 485 (2%) touch on aspects of the longitudinal study as a whole. They describe the type of study they enjoy and explain how longitudinal studies provide guaranteed work without the need to compete for tasks.

6.1.10 Participation Incentives (In Previous Experiences). Figure 9 addresses the underlying motivations that drive workers' participation in the previous experiences with longitudinal studies reported.

Monetary aspects such as rewards and bonuses are the most important incentives for the participation in the majority of reported experiences (70.42%). Workers' personal interest in the task proposed by the requester in the longitudinal study is an incentive for roughly 19% of experiences. Roughly 6% of participation in the reported experiences occurred because the worker found the task proposed educative, while the workers' altruism, in terms of helping the overall research, has a lower but not negligible importance, considered by 4.71% of the respondents.

When considering each platform, it is interesting to note that 17% of Toloka workers found the task proposed in the reported experience with longitudinal study educative, while this component is almost absent from Amazon Mechanical Turk (1%) and Prolific. Furthermore, Prolific is the platform that published the majority of experiences that took place due to workers' personal interest (26%) or willingness to help the research (7%). This may be due to the fact that such platforms are mostly focused on academic-related research projects, and task requesters are often researchers [58].

Generally, even though monetary aspects are the most popular incentives that drove workers to participate in the previous experiences with longitudinal studies reported, the remaining factors should not be overlooked when designing the overall longitudinal study (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka, and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01).

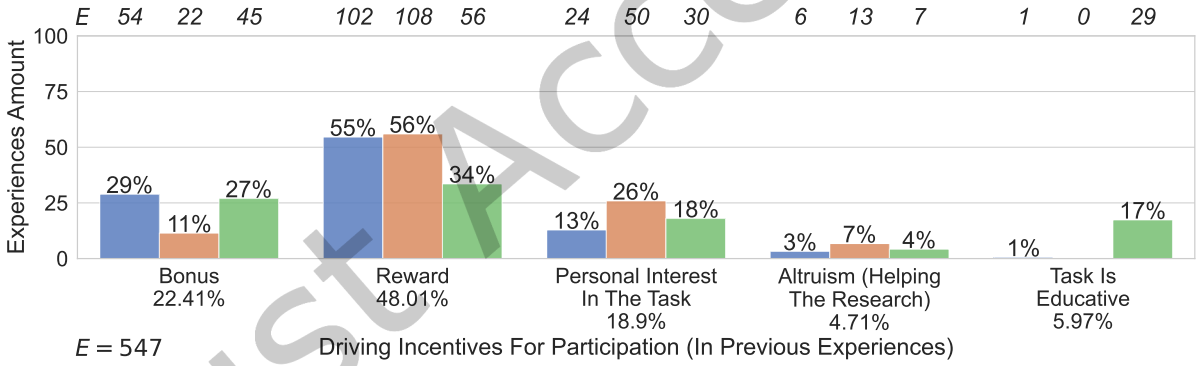

Fig. 9. Incentives that drive workers to participate in the longitudinal study to which each reported experience refers.

6.1.11 Study Completion. Figure 10 investigates whether workers completed the overall longitudinal study to which each reported experience refers. Specifically, they claim completion of almost every previous experience (97.65%), with only 13 experiences out of 547 (2.35%) dropped.

When considering each platform, workers claim completion of almost every experience on Prolific and Toloka (99%), while this amount is slightly lower for Amazon Mechanical Turk, particularly 95% (no statistically significant comparisons across platforms obtained). Even though the crowdsourcing platforms do not provide any means of verifying this data, we recall that we recruit workers with certain task completion rates (i.e., experienced workers), as described in Section 5.2. Thus, we argue that they have little incentive to provide inaccurate information about their previous completions.

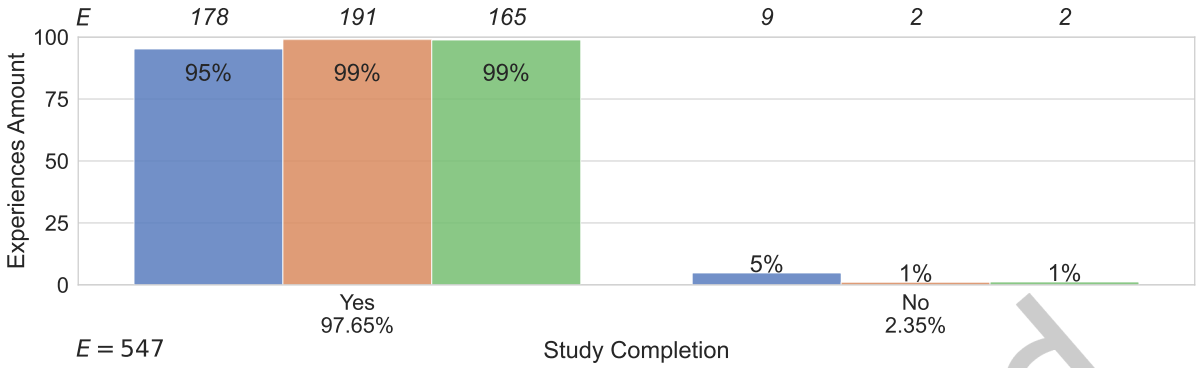

Fig. 10. Completion claimed by workers of the longitudinal study to which each reported experience refers.

6.1.12 Completion Incentives (In Previous Experiences). Figure 11 addresses the underlying motivations that drive workers to complete the previous experiences with longitudinal studies reported and should be compared with the answers provided for question 1.1.X.8, analyzed in Section 6.1.10, which focuses on the ones that drive workers to participate. Indeed, while the set of possible answers is the same, this question restricts the focus to completed experiences and attempts to grasp the changes in workers' perception of the overall experience. Thus, the 11 experiences from which workers dropped participation (i.e., those reported in the right half of Figure 10) are marked using a separate string, that is "Participation Dropped", to allow a direct comparison of the bar charts.

Monetary aspects such as rewards and bonuses remain the most important factors for the majority of the previous experiences reported (68.3%), with a slight decrease (2.12%). The impact of workers' personal interest in the task proposed by the requester (18.49%) remains almost unchanged, as does their opinion about the task being educative. Most of the answers that shift from monetary aspects, indeed, end up describing workers' willingness to help with the overall research, from 4.71% to 6.06%.

When considering each platform, the overall distribution of answers does not change in terms of relative comparisons. The most noticeable difference is found for Prolific, where workers' personal interest in the proposed task drops from 26% to 19%, becoming comparable with that of other platforms. A similar phenomenon occurs for Amazon Mechanical Turk, where interest in the final reward shifts from 49% to 55% (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01).

6.1.13 Crowdsourcing Platforms Suitability. The mandatory open-ended question 2 (P1 part) is used to ask workers about the adequacy and suitability of the crowdsourcing platform of provenance in the support they provide for longitudinal studies.

The majority of workers (273 out of 300, 91%) provided an answer that allows us to draw some kind of consideration. The distributions of the answers collected across different themes is as follows: 244 out of 273 (89.34%) addressed aspects related to the crowdsourcing platform (platform_features), while 11 (4.03%) focused on workers' own beliefs and motivations (worker_features). Lastly, 18 (6.59%) answers were deemed unusable (answer_useless). Table 8 (Appendix B) shows a sample of such answers.

The vast majority of answers directly relate to the crowdsourcing platform of origin (244 out of 273, 89.34%). Breaking down the respondents across each platform reveals 98 workers from Amazon Mechanical Turk, 100 from Prolific, and 76 from Toloka. The majority of Amazon Mechanical Turk workers (70 out of 98, 71.43%) believe the platform is generally adequate, with few providing additional details. Three of them (3.06%) specifically mention

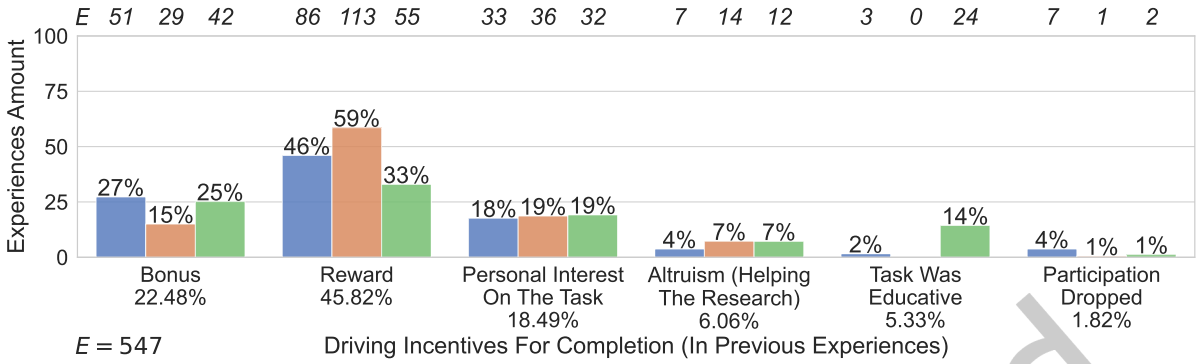

Fig. 11. Incentives that drive workers to completing the longitudinal study to which each reported experience refers.

the ease of sending reminders for upcoming longitudinal study sessions. Only seven (7.14%) find the platform inadequate in supporting longitudinal studies. One worker suggests that the platform needs design improvements to facilitate scheduling tasks for longitudinal studies, while another highlights the challenge for requesters to ensure worker honesty.

Nearly all Prolific workers (97 out of 100, 97%) consider the platform adequate for supporting longitudinal studies, with many providing detailed responses. Some mention the platform's detailed task reports, facilitating tracking throughout the study. Others (7 out of 100, 7%) highlight the diverse backgrounds and skills of available individuals. Factors such as ease of contacting or sending reminders to workers using their identifier are noted by 16 out of 100 workers (16%). Additionally, two workers (2 out of 100, 2%) emphasize worker motivation and reliability as important considerations for researchers. Notably, one worker mentions being recruited from the platform via a third-party application that relies on the platform's API.

The majority of Toloka workers (68 out of 76, 89.47%) consider the platform adequate overall, with few providing specific details. Two workers (2 out of 76, 2.63%) mention worker availability and the ease of contacting them using their identifier. One worker's response is notable; they believe the platform cannot adequately support a longitudinal study due to residing in a country with poor network infrastructure.

When considering workers who are uncertain or outright deny the adequacy of the platform, several cross-platform factors become apparent. These workers are more likely to drop out of longitudinal studies due to perceived inadequacies. They express difficulties in assessing requester honesty, which can lead to skepticism about participating in such studies. Additionally, respondents believe that workers typically do not actively seek out longitudinal studies, suggesting a need for platforms to better distinguish these studies from standard crowdsourcing tasks.

6.1.14 Reasons That Limit Availability On Platforms. Figure 12 investigates the reasons that limit the availability of longitudinal studies on crowdsourcing platforms according to workers' opinions. The most prevalent reasons, chosen roughly the same number of times, are that workers dislike the required commitment (32.85%) and that the provided rewards and incentives are insufficient. Several answers indicate that longitudinal studies are not optimally supported by current popular crowdsourcing platforms (24.85%), and 9.07% of answers point out that usually requesters do not need longitudinal participation since most tasks deal with static data to annotate.

The distribution of answers changes when considering each platform. Specifically, 44% of the answers provided by Prolific workers indicate their dislike of the required commitment, while this factor is less important for Amazon Mechanical Turk workers (29%) and Toloka workers (26%). The lack of adequate technical support is

prevalent among the answers provided by Toloka workers (35%), while for Prolific, this is reported by only 12% of the answers. The percentage of answers indicating that rewards and incentives are insufficient is slightly higher for Amazon Mechanical Turk (36%) compared to Toloka (33%), which in turn is slightly higher than Prolific (29%). Among the answers describing that often crowdsourcing tasks do not need longitudinal participation, those from Prolific are prevalent (15%).

Summarizing, workers indeed dislike the required commitment and find monetary aspects and related incentives insufficient. Also, they think that longitudinal studies are not adequately supported by crowdsourcing platforms (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka, and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01).

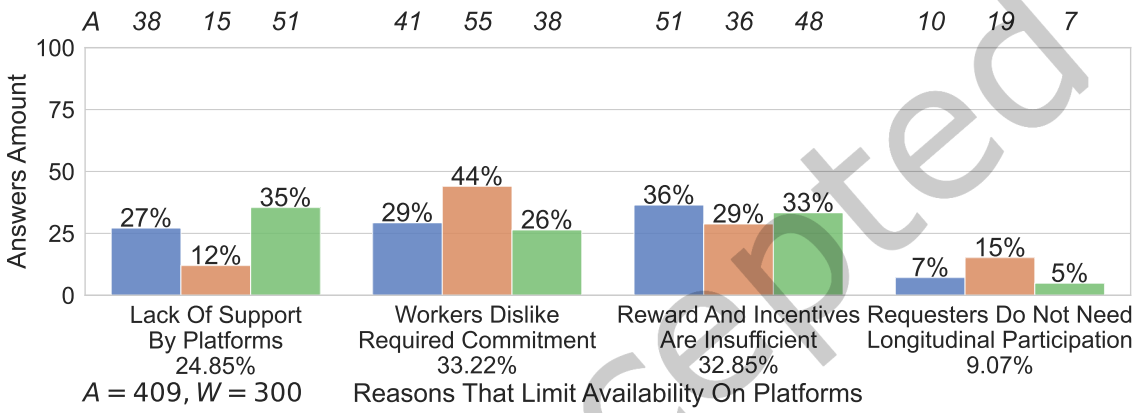

Fig. 12. Reasons that limit the availability of longitudinal studies on crowdsourcing platforms, according to workers.

6.1.15 Preferred Commitment Duration. Figure 13 investigates the number of days workers would be happy to commit to for a longitudinal study, hypothesizing a single session having a duration of 15 minutes per day.

By considering each platform, Amazon Mechanical Turk and Toloka workers show rather similar trends, with mean numbers of days around 19 and 17, respectively. Turning to Prolific, this number increases to an average of almost a month (29 days).

Generally, Prolific is the platform that allows for finding workers willing to commit to longitudinal studies for longer periods, at least when compared with Toloka (Prolific vs Toloka statistically significant with adjusted p-value < 0.05).

6.1.16 Reasons For Declining Participation. Figure 14 investigates which are the reasons that drive workers to decline participation in longitudinal studies.

The majority of the answers provided by workers indicate that the length of the longitudinal study, in terms of the number of sessions and thus the time elapsed in days or even months since its start, is the most important factor (70.79%). The remaining answers (29.03%) indicate that the frequency of the sessions of the longitudinal study is also a reason that can lead to declining participation and should not be overlooked.

By considering each platform, the vast majority of answers provided by Prolific workers (85%) consider study length as a major concern, and this holds also when considering Toloka, albeit to a lesser extent (71%). As for Amazon Mechanical Turk, the trend is more nuanced, since the gap between answers that consider study length (57%) and study frequency (43%) is smaller (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01).

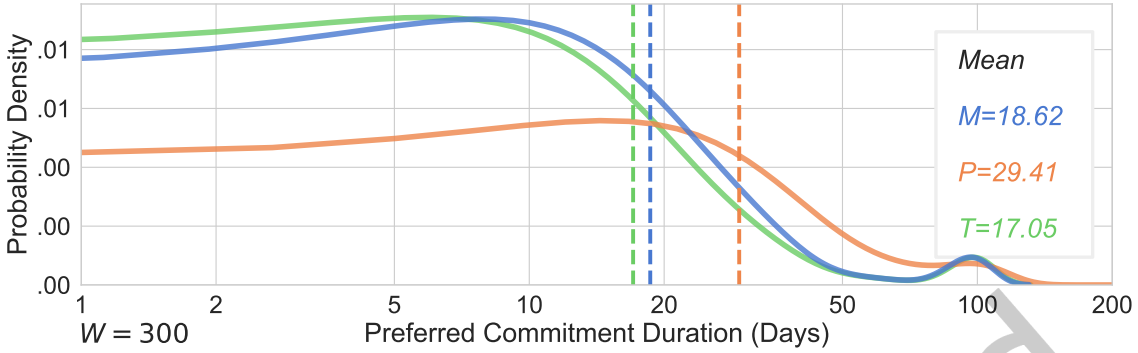

Fig. 13. Number of days workers would be happy to commit for a longitudinal study, hypothesizing a single session of 15 minutes per day.

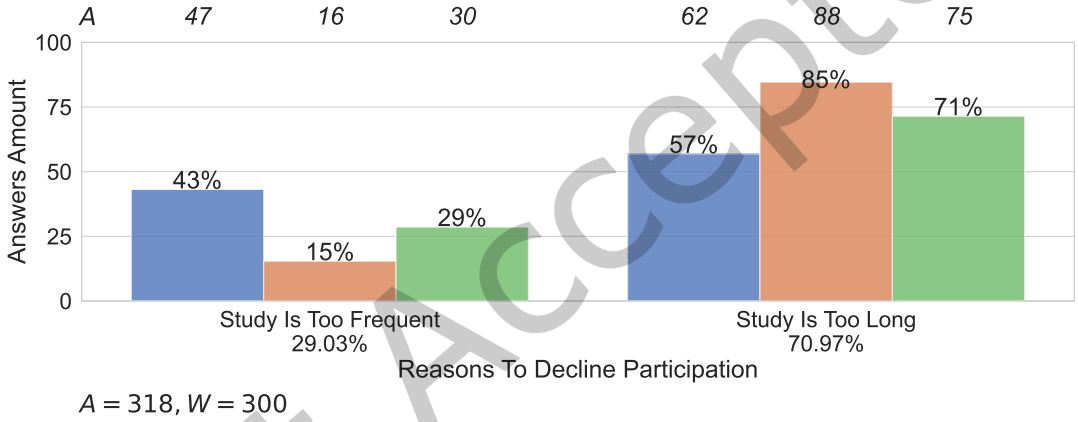

Fig. 14. Reasons that drive workers to decline participation in longitudinal studies.

6.1.17 Preferred Participation Frequency. Figure 15 investigates the preferred participation frequency in longitudinal studies according to the workers, in terms of time periods.

The vast majority of workers prefer frequent studies, having a daily to weekly participation commitment. Particularly, a daily participation is the most popular option overall (42.78%). Only a niche of 11 workers (6.68%) would prefer longer time periods.

There are some nuances among the preferences of the workers recruited from each platform. Particularly, Toloka workers prefer, for the most part, a daily participation frequency (53%). Prolific workers, on the other hand, have a slightly higher preference for a weekly frequency (40%), followed by a daily frequency (35%). For Amazon Mechanical Turk workers, the trend is the opposite, as they prefer a daily participation frequency (40%), shortly followed by a weekly frequency (38%). Regarding longer frequencies, it is worth noting that 6 Toloka workers (6%) prefer a biweekly frequency, and 5 Amazon Mechanical Turk workers (5%) along with 3 Prolific workers (3%) prefer a monthly frequency.

These findings can be aligned with those described in Figure 14, as indeed the study length is a major concern for workers (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01).

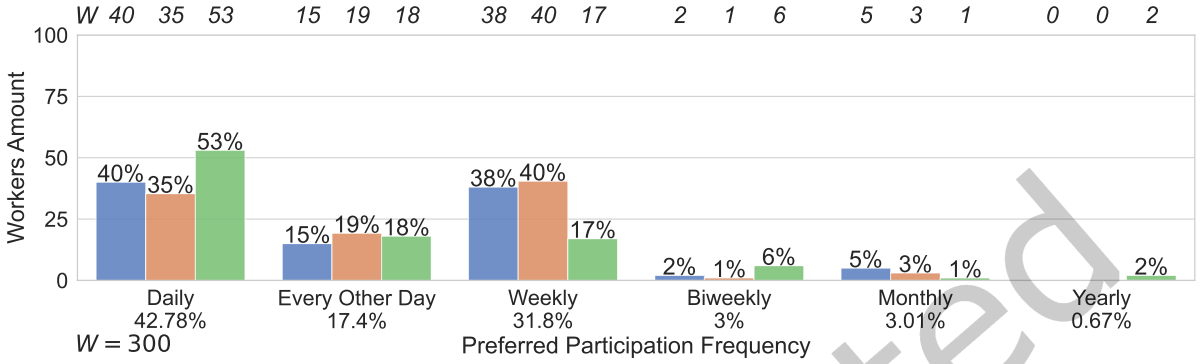

Fig. 15. Preferred participation frequency in a longitudinal study according to workers.

6.1.18 Preferred Session Duration. Figure 16 investigates the preferred session duration in hours for longitudinal studies according to workers.

Prolific workers prefer short sessions of less than 1 hour on average, while Amazon Mechanical Turk and Toloka workers share a more uniform preference, indicating an average of about two hours. The figure does not show 9 outliers who provide non-reasonable durations (i.e., between 15 and 50 hours), thus they are removed.

Generally speaking, Amazon Mechanical Turk and Toloka workers are thus keen to work for a longer time within a single session when compared with Prolific workers (Amazon Mechanical Turk vs. Prolific and Prolific vs. Toloka statistically significant; adjusted p-value < 0.05).

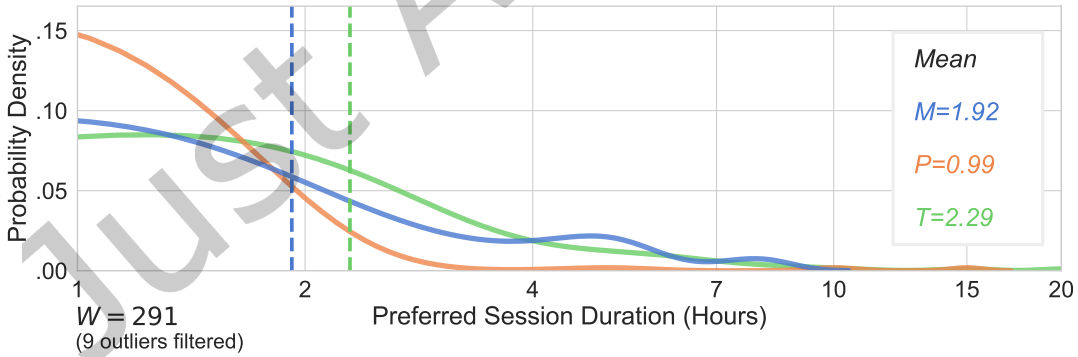

Fig. 16. Preferred session duration in hours for longitudinal studies according to workers.

6.1.19 Acceptable Hourly Payment. Figure 17 investigates the acceptable hourly payment rate in USD$ for participating in longitudinal studies on the recruitment platform, as reported by the workers.

Amazon Mechanical Turk workers aim to receive the highest hourly payment on average (about $13), while for Prolific workers, this amount lowers to about $10.50. On the other hand, Toloka workers indicate the lowest

acceptable amount of money (about $8.5). The figure does not include 8 outliers who provided unreasonable amounts (i.e., amounts ranging between $80 and $100) and were thus removed.

To interpret the provided answers, one must consider that the payment models of Amazon Mechanical Turk and Toloka differ from that of Prolific. On the first two platforms, a task requester proposes a unitary amount of money for each work unit performed, which can be arbitrarily high. On the other hand, the Prolific platform requires requesters to estimate the task completion time and propose, instead of a unitary amount, a minimum amount of money based on the hourly estimate. Thus, this difference may impact the workers' perception of the acceptable payment amount (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka statistically significant; adjusted p-value < 0.05).

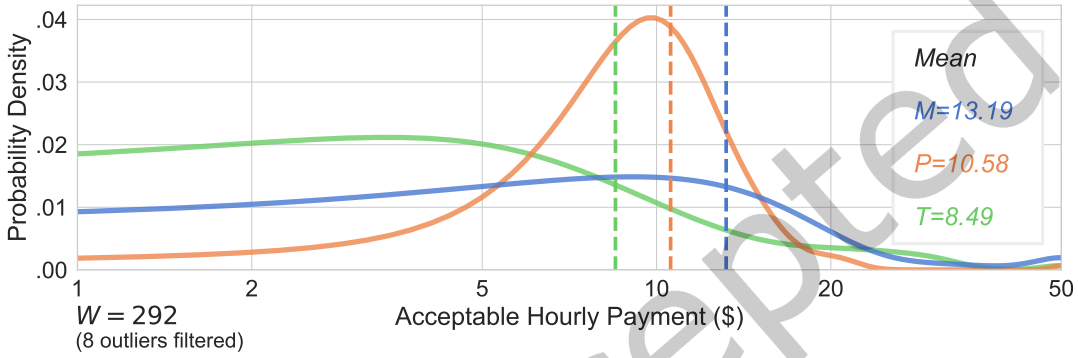

Fig. 17. Acceptable hourly payment in USD$ for participation in longitudinal studies according to the workers.

6.1.20 Preferred Time To Allocate Daily. Figure 18 investigates the preferred amount of time in hours that workers are available to allocate for participating in longitudinal studies on a daily basis.

The workers recruited on Toloka are those keen to work more per day, being available to allocate up to almost four hours on average (3.81). Then, Amazon Mechanical Turk workers prefer working up to almost three hours (2.85), while Prolific ones expect to work less, with roughly an hour and a half (1.66). The figure does not show 18 outliers who provided non-reasonable amounts of hours per day (i.e., between 20 and 25), and were thus removed.

In general, Toloka workers are those who are keen to work more within a day and expect to be rewarded less. This is evident not only in the time they allocate daily for participation, as shown in Figure 18, but also when asked about their preferred session duration (Figure 16) or their ideal daily payment (Figure 17). As for Amazon Mechanical Turk and Prolific workers, they expect to work less on average, particularly the latter ones (Amazon Mechanical Turk vs. Prolific, Prolific vs. Toloka statistically significant; adjusted p-value < 0.05).

6.1.21 Participation Incentives (In New Experiences). Figure 19 investigates the underlying motivations that drive participation in new longitudinal studies.

In general, the type of reward/payment mechanism is the most important incentive, according to the vast majority of answers provided by workers (81.86%). Among them, the preferred alternative is providing payment after each session (32.07%). As for the remaining ones, 24.22% indicate a final bonus to be awarded after the last session, while 20.38% prefer a progressive incremental payment after each session. A progressive decremental payment (2.51%) or eventual penalization for skipping one or more sessions (2.43%) have a small but not negligible influence on participation chances in new studies.

Beyond the reward/payment mechanism, 12.04% of answers indicate working on different task types to increase engagement diversity, while 6.18% suggest experimental variants of the same tasks to reduce repeatability. When

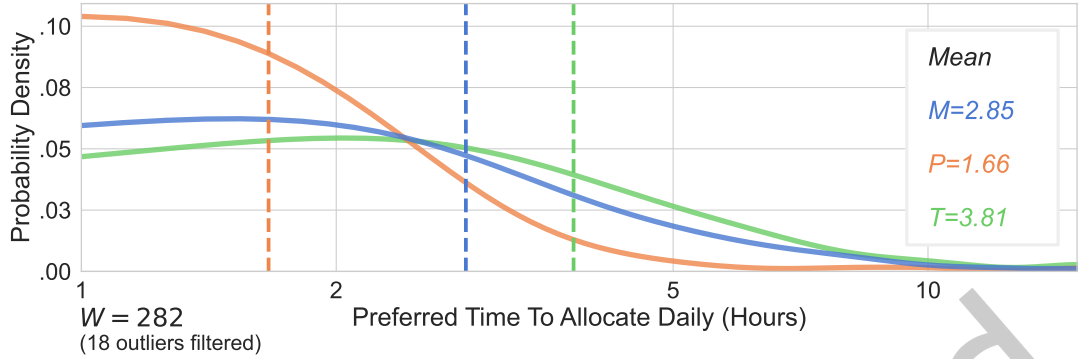

Fig. 18. Preferred amount of time in hours to allocate on a daily basis for participating in longitudinal studies according the workers.

considering each crowdsourcing platform, no particular trends emerge (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka, and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01).

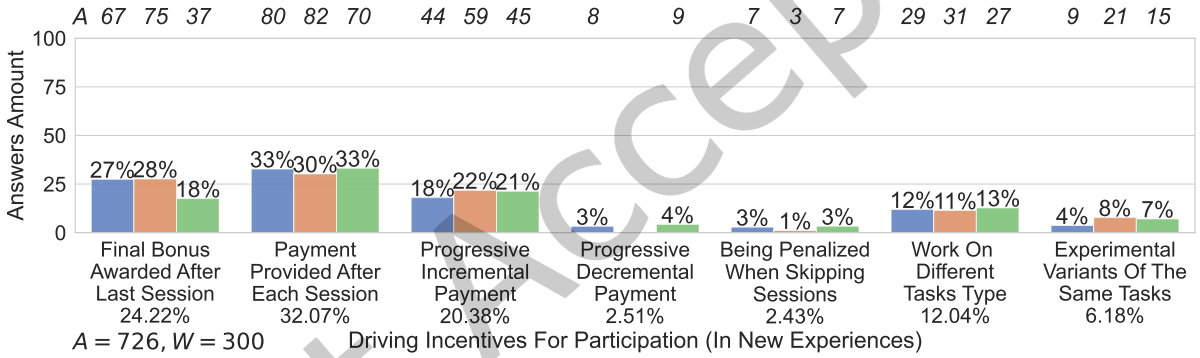

Fig. 19. Incentives that drive workers to participate in new longitudinal studies according to those who answered the survey.

6.1.22 Tasks Type. Figure 20 investigates the tasks that workers would like to perform in a longitudinal study. We acknowledge that the predefined set of answers we provided might not have been perceived as exhaustive. Indeed, they were given the opportunity to provide a free-text response to further elaborate.

By surveys, we refer to surveys about various aspects that are usually crowdsourced, like demographics (22.71%). Verification and validation tasks require workers in the crowd to either verify certain aspects as per the given instructions, or confirm the validity of various kinds of content (17.99%). Interpretation and analysis tasks rely on the wisdom of the crowd to use their interpretation skills during task completion (17.92%). Information finding tasks delegate the process of searching to satisfy one's information need to the workers in the crowd (16.51%). Content access tasks require the crowd workers to simply access some content (14.59%) and content creation tasks require the workers to generate new content for a document or website (10.28%).

It is worth noting that two workers mentioned in their free text responses other types of tasks, namely gamified tasks and content editing, which indeed is an option that we did not consider along with content access and content creation.

Summarizing, workers are willing to perform any of the task types proposed across each platform, with a rather homogeneous answer distribution. However, this distribution still accounts for statistical significance (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka, and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01).

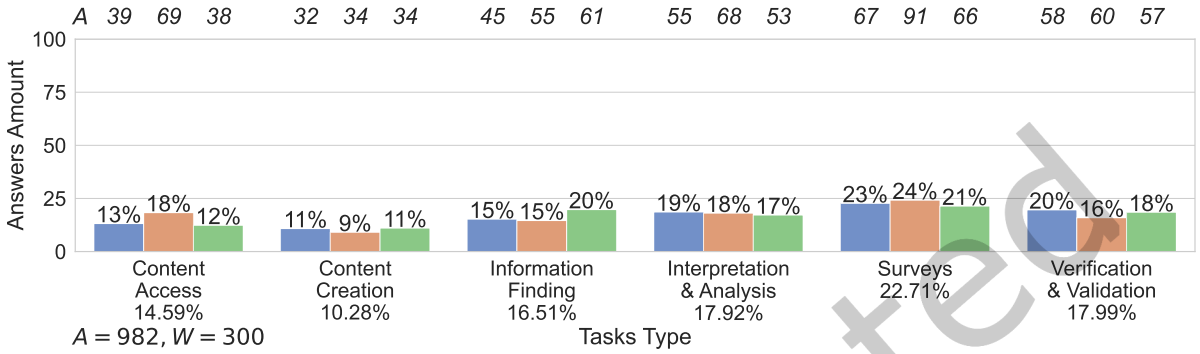

Fig. 20. Tasks type that workers would like to perform in a longitudinal studies according to those who answered the survey.

6.1.23 Involvement Benefits. Figure 21 investigates which are the benefits of being involved in longitudinal studies according to workers.

In general, workers think that the most important benefit characterizing longitudinal studies is increased productivity due to their more operational nature (32.1%). They also appreciate the time-saving aspect, as longitudinal studies eliminate the need for regular task searching (26.64%). Furthermore, workers think that receiving intermediate payments, after each session of the longitudinal study, would increase trust in the requester (25.81%). Additionally, some workers find value in avoiding the need to re-learn tasks when participating in longitudinal studies (15.45%).

The trends are homogeneous across all platforms, with no factor considered more important than others. However, the only exception is increased productivity, which is more prominent for Amazon Mechanical Turk workers (36%) and Toloka workers (37%) compared to Prolific (24%). Nonetheless, this distribution still accounts for statistical significance (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka, and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01).

6.1.24 Involvement Downsides. Figure 22 investigates which are the downsides of being involved in longitudinal studies according to workers.

The answers provided by workers indicate that a reward provided only at the end of the longitudinal study is the most important downside (30.87%). The lack of flexibility in the study schedule and the long term commitment required have roughly are indicated by roughly the same amount of answers, namely 27.48% and 27.63%. The lack of diversity in terms of the work to be performed during each session of the overall study plays a minor role (14.02%).

By considering each platform, the trends are rather homogeneous for Amazon Mechanical Turk and Prolific. However, it is interesting to notice how the lack of diversity is a more prominent downside for Toloka workers (20%), while at the same time, the long-term commitment is less of an issue (21%) when compared to the remaining platforms (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka, and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01).

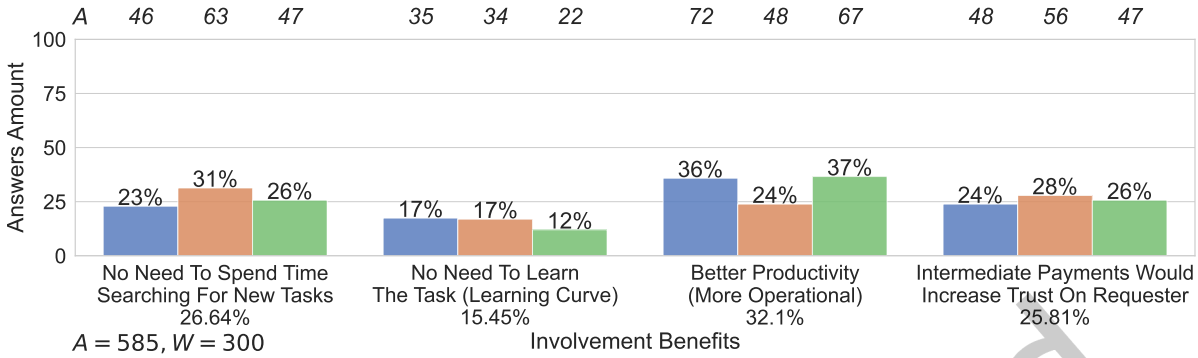

Fig. 21. Benefits of being involved in longitudinal studies according to the workers.

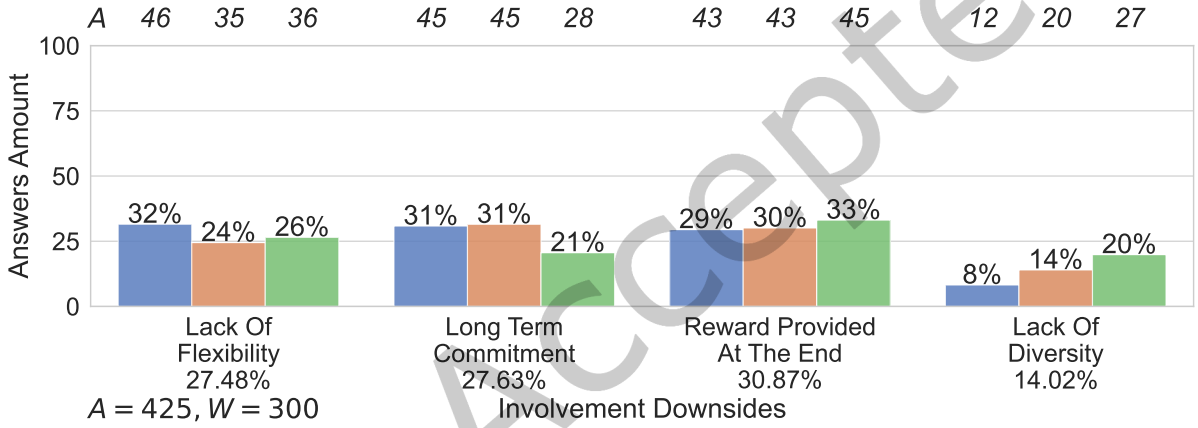

Fig. 22. Downsides of being involved in longitudinal studies according to the workers.

6.1.25 Suggestions About Longitudinal Study Design. The last and optional question 11 (P2 part) asked workers to provide any suggestions to requesters that aim to design a longitudinal study.

There are 201 out of 300 (67%) workers who provide some kind of answer. The distribution of answers collected across different themes is as follows: 139 out of 201 (69.15%) addressed aspects related to the task performed (task_features), 9 (4.48%) focused on requesters (requester_features), and 7 (3.48%) focused on workers' own beliefs and motivations (worker_features). Additionally, 5 (2.49%) were about the longitudinal study as a whole (ls_features), and 2 (1%) were about the platform (platform_features). Lastly, 2 (1%) answers were deemed unusable (answer_useless). Furthermore, 37 (18.41%) workers explicitly stated that they did not have any suggestions (no_suggestion). Table 9 (Appendix B) shows a sample of such answers.

The majority of workers (139 out of 201, 69.15%) suggest improvements related to the features of the task to be performed within each session of the longitudinal study, including its design, scheduling, and participant filtering. Six out of 139 workers (4.32%) propose allowing a reasonable window for completion, considering other activities in workers' schedules. One worker suggests the option to skip a session if unable to commit occasionally. Additionally, a few workers (3 out of 139, 2.16%) emphasize the importance of conducting pilot tests for the tasks, which can help both requesters find suitable participants and retain workers throughout the study. A

worker suggests offering different systems for participating in the study (e.g., desktop devices, smartphones) and another worker advises against requiring downloads. This resonates with prior work that has revealed diverse work environments that workers are embedded in [25]. Workers emphasize the need for clear instructions and user interface, an understandable sequence of events, identifying changes over time, and providing insight into cause-and-effect relationships. Some believe variability could help maintain interest in the study.

Regarding the overall structure of a longitudinal study (5 out of 201, 2.49%), workers suggest planning all sessions in advance while remaining flexible with the schedule, especially when involving multiple geographic time zones. They also recommend establishing a sense of progression, such as highlighting differences in previous responses at the end of each session.

A few workers (7 out of 201, 3.48%) provide personal insights. One worker notes that many are self-employed and must pay taxes on their earnings from crowdsourcing platforms, so rewards should reflect this. Another worker prefers small payments with a bonus for completing all sessions.

Considering aspects related to the task requesters (9 out of 201, 4.48%), workers think regular feedback from requesters is important. They suggest that requesters should be communicative and friendly, leave spaces for feedback in each study, send reminders when needed, and provide clear upfront information.

6.1.26 Summary. The workers' answers for the P1 part of the survey are summarized in Table 4, while those provided for P2 in Table 5. Both tables provide a detailed summary of the answers, along with the code used to classify each question and a breakdown of responses across each crowdsourcing platform considered.

Table 6 shows the outcome of statistical tests performed by comparing the groups of answers provided across each platform. The table includes the name and answer type of each question. A checkmark (✓) indicates a statistically significant comparison with the adjusted p-value provided, while its absence indicates that a given comparison was not statistically significant.

Finally, we summarize the key findings with a list of take-home messages, starting from the perception of longitudinal studies' according to workers' previous experiences (messages 1-9, P1 part questions), then moving to workers' opinions about future longitudinal studies (messages 10-17, P2 part questions). For each message, we report a reference to the corresponding section where the analysis is reported.

1. Workers with more experience with longitudinal studies can be found more easily on the Prolific platform (Section 6.1.1), and the available studies on this platform tend to have more sessions compared to other platforms (Section 6.1.3).
2. Most of the experiences reported by the workers took place up to one year before their participation in the survey (Section 6.1.2).
3. Most of the sessions of the reported longitudinal studies lasted up to 2 hours, with roughly half of them lasting for only 15 minutes (Section 6.1.5).
4. Most of the time intervals between sessions in the reported longitudinal studies range from 1 to 30 days (Section 6.1.4).
5. Most of the longitudinal studies reported provide partial rewards after each session (Section 6.1.7).
6. The main motivation that drove workers to participate in and complete the reported longitudinal studies is the monetary aspect (Section 6.1.10 and Section 6.1.12).
7. Almost every worker claims completion of the reported longitudinal studies (Section 6.1.11).
8. Most of the workers want to continue participating in the longitudinal studies reported in the future (Section 6.1.8).
9. The main reasons that limit the availability of longitudinal studies on crowdsourcing platforms are workers' dislike for the required commitment and the insufficiency of the provided rewards (Section 6.1.14).
10. In a hypothetical longitudinal study where workers are asked to engage in a single session for 15 minutes each day, workers are willing to commit to participating for an average of 21 days (Section 6.1.15). However,

when considering session duration, workers are generally willing to work for up to an average of 103 minutes per session (Section 6.1.18).

11. Most of the workers prefer a daily to weekly participation frequency for longitudinal studies (Section 6.1.17).
12. The workers prefer to allocate daily for participating in longitudinal studies about 2.7 hours on average (Section 6.1.20).
13. The workers think that the acceptable hourly payment for participating in longitudinal studies is about $10.75 on average (Section 6.1.19). It must be noted that such an amount should be adjusted for inflation.
14. Workers report that the main incentives driving participation in new longitudinal studies are related to the reward provided (Section 6.1.21).
15. Most of the workers believe that the length of a longitudinal study is critical in influencing their decision to refuse participation (Section 6.1.16).
16. Workers report that the main benefits of being involved in longitudinal studies are increased productivity due to their operational nature and the elimination of the need for regular task searching (Section 6.1.23).
17. Workers report that the main downsides of being involved in longitudinal studies are the long-term commitment required, the lack of flexibility, and the reward provided only at their completion (Section 6.1.24).

Table 4. Summary of the key findings for the P1 part of the survey presented in the quantitative analysis.

Section	Question	Amazon Mechanical Turk	Prolific	Toloka
6.1.1	*Previous Experiences*	42% 1 experience, 29% 2 experiences, 29% 3 experiences	43% 1 experience, 21% 2 experiences, 36% 3 experiences	50% 1 experience, 33% 2 experiences, 17% 3 experiences
6.1.2	*Time Elapsed*	87% up to 1 year before, 13% later	87% up to 1 year before, 13% later	87% up to 1 year before, 13% later
6.1.3	*Sessions*	∼6 on average	∼7 on average	∼6 on average
6.1.4	*Interval Between Sessions*	89% up to 1 month, 11% later	88% up to 1 month, 12% later	97% up to 1 month, 6% later
6.1.5	*Session Duration*	98% up to 1 hour, 3% more	99% up to 1 hour, 1% more	91% up to 1 hour, 8% more
6.1.6	*Crowdsourcing Platform*	91% MTurk, 9% Prolific, 0% Toloka	6% MTurk, 90% Prolific, 4% Toloka	17% MTurk, 19% Prolific, 63% Toloka
6.1.7	*Payment Model*	75% after each session, 16% final reward, 9% both	68% after each session, 25% final reward, 7% both	68% after each session, 25% final reward, 7% both
6.1.8	*Participation In Same Study*	83% yes, 17% no	98% yes, 2% no	93% yes, 7% no
6.1.10	*Participation Incentives (In Previous Experiences)*	29% bonus, 55% reward, 13% personal interest, 3% altruism, 1% educative task	11% bonus, 56% reward, 26% personal interest, 7% altruism, 0% educative task	27% bonus, 34% reward, 18% personal interest, 4% altruism, 17% educative task
6.1.11	*Study Completion*	95% yes, 5% no	99% yes, 1% no	99% yes, 1% no
6.1.12	*Completion Incentives (In Previous Experiences)*	27% bonus, 46% reward, 18% personal interest, 4% altruism, 2% educative task, 4% participation dropped	15% bonus, 59% reward, 19% personal interest, 7% altruism, 0% educative task, 1% participation dropped	25% bonus, 33% reward, 19% personal interest, 7% altruism, 14% educative task, 1% participation dropped
6.1.14	*Reasons That Limit Availability On Platforms*	27% lack of support, 29% dislike commitment, 36% reward and incentives insufficient, 7% no need longitudinal participation	12% lack of support, 44% dislike commitment, 29% reward and incentives insufficient, 15% no need longitudinal participation	35% lack of support, 26% dislike commitment, 33% reward and incentives insufficient, 5% no need longitudinal participation

Table 5. Summary of the key findings for the P2 part of the survey presented in the quantitative analysis.

Section	Question	Amazon Mechanical Turk	Prolific	Toloka
6.1.15	*Preferred Commitment Duration*	~19 days on average	~29 days on average	~17 days on average
6.1.16	*Reasons For Declining Participation*	42% study is too frequent, 58% study is too long	15% study is too frequent, 85% study is too long	29% study is too frequent, 71% study is too long
6.1.17	*Preferred Participation Frequency*	92% up to 1 week, 3% biweekly, 5% monthly, 0% yearly	95% up to 1 week, 1% biweekly, 2% monthly, 0% yearly	88% up to 1 week, 5% biweekly, 1% monthly, 2% yearly
6.1.18	*Preferred Session Duration*	~115 minutes on average	~60 minutes on average	~137 minutes on average
6.1.19	*Acceptable Hourly Payment*	13.19 USD$ on average	10.58 USD$ on average	8.49 USD$ on average
6.1.20	*Preferred Time To Allocate Daily*	~171 minutes on average	~100 minutes on average	~228 minutes on average
6.1.21	*Participation Incentives (In New Experiences)*	27% final bonus, 33% pay after each session, 18% prog. incr. payment, 3% progr. decr. payment, 3% penalization for skipping, 12% different task types, 4% experimental variants	28% final bonus, 30% pay after each session, 22% prog. incr. payment, 0% prog. decr. payment, 1% penalization for skipping, 11% different task types, 8% experimental variants	18% final bonus, 33% pay after each session, 21% progr. incr. payment, 4% progr. decr. payment, 3% penalization for skipping, 13% different task types, 7% experimental variants
6.1.22	*Tasks Type*	13% content access, 11% content creation, 15% information finding, 18% interpretation and analysis, 23% surveys, 19% verification and validation	18% content access, 9% content creation, 15% information finding, 18% interpretation and analysis, 24% surveys, 16% verification and validation	12% content access, 9% content creation, 20% information finding, 18% interpretation and analysis, 24% surveys, 16% verification and validation
6.1.23	*Involvement Benefits*	23% no need to search, 17% no need to learn, 36% better productivity, 24% increase trust	31% no need to search, 17% no need to learn, 34% better productivity, 28% increase trust	26% no need to search, 12% no need to learn, 37% better productivity, 26% increase trust
6.1.24	*Involvement Downsides*	32% lack of flexibility, 31% long term commitment, 29% reward at the end, 8% lack of diversity	24% lack of flexibility, 31% long term commitment, 30% reward at the end, 14% lack of diversity	26% lack of flexibility, 21% long term commitment, 33% reward at the end, 20% lack of diversity

Table 6. Summary of statistical tests comparing answer groups of each platform. Questions without any statistically significant comparisons are not reported. Statistical significance is computed using adjusted p-values according to Section 5.4

Part	Section	Question	Type	MTurk Vs. Prolific	MTurk Vs. Toloka	Prolific Vs. Toloka	Significance Level
P1	6.1.2	*Time Elapsed*	mcq		✓		$p \leq 0.05$
P1	6.1.4	*Interval Between Sessions*	mcq		✓		$p \leq 0.01$
P1	6.1.5	*Session Duration*	mcq	✓	✓	✓	$p \leq 0.01$
P1	6.1.6	*Crowdsourcing Platform*	mcq	✓	✓	✓	$p \leq 0.01$
P1	6.1.7	*Payment Model*	list	✓	✓	✓	$p \leq 0.01$
P1	6.1.8	*Participation In Same Study*	mcq	✓	✓	✓	$p \leq 0.01$
P1	6.1.10	*Participation Incentives (In Previous Experience)*	mcq	✓		✓	$p \leq 0.01$
P1	6.1.12	*Completion Incentives (In Previous Experience)*	mcq	✓	✓	✓	$p \leq 0.01$
P1	6.1.14	*Reasons That Limit Availability On Platforms*	mcq	✓	✓	✓	$p \leq 0.01$
P2	6.1.15	*Preferred Commitment Duration*	number			✓	$p \leq 0.05$
P2	6.1.16	*Reasons For Declining Participation*	list	✓	✓	✓	$p \leq 0.01$
P2	6.1.17	*Preferred Participation Frequency*	mcq	✓	✓	✓	$p \leq 0.01$
P2	6.1.18	*Preferred Session Duration*	number		✓		$p \leq 0.05$
P2	6.1.19	*Acceptable Hourly Payment*	number	✓	✓		$p \leq 0.05$
P2	6.1.20	*Preferred Time To Allocate Daily*	number	✓		✓	$p \leq 0.05$
P2	6.1.21	*Participation Incentives (In New Experiences)*	list	✓	✓	✓	$p \leq 0.01$
P2	6.1.22	*Tasks Type*	list	✓	✓	✓	$p \leq 0.01$
P2	6.1.23	*Involvement Benefits*	list	✓	✓	✓	$p \leq 0.01$
P2	6.1.24	*Involvement Downsides*	list	✓	✓	✓	$p \leq 0.01$

6.2 RQ2: Recommendations For Researchers And Practitioners

Although there is no standard approach for designing and conducting longitudinal studies on a crowdsourcing platform, the quantitative and qualitative analyses of workers' responses, together with our experience in deploying a crowdsourcing task, have allowed us to develop 8 recommendations that could serve as a framework (Section 6.2.1–6.2.8).

We believe that our recommendations should be considered by task requesters when designing longitudinal studies, as they provide useful guidelines and address workers' fears and needs that emerged during our study.

6.2.1 R1: *Be Communicative And Provide Feedback.* Communication is a critical factor in encouraging workers' retention and decreasing the abandonment rate, as emerges by their answers about what drove them towards returning to longitudinal studies and their suggestions to task requesters, described in Section 6.1.9.

According to the workers, task requesters should inform them about upcoming sessions, the progress made throughout the study, and, eventually, contact the workers explicitly to invite them to participate in newly published studies, also considering that several days can pass, as shown in Figure 4. Requesters should also provide information about the overall progress of the longitudinal study and feedback concerning the quality of the work performed up to the current session.

When asked to provide additional suggestions, as reported in Section 6.1.25, they also point out that alerts, emails, or notifications should be sent according to a regular schedule. Sending them randomly could be detrimental to the worker experience. Furthermore, they note that platforms like Prolific provide only an internal notification system, without any way to send a standard email to the worker.

6.2.2 R2: *Schedule Each Session Mindfully.* Workers have free time to dedicate to participation in crowdsourcing tasks during different days of their working week, as reported in Section 6.1.25, and generally, it consists of roughly 2 hours on average (Figure 18). Scheduling the work required properly, thus, is particularly important for longitudinal studies, also considering that they can be composed of a potentially high number of sessions, as shown in Figure 3.

Determining a priori and explicitly stating the overall number of sessions is useful since it would allow the worker to estimate the amount of commitment required, especially given that they are available to commit for up to a month, as shown in Figure 13. Communicating when the subsequent session is going to happen will provide some flexibility to the workers.

Furthermore, task requesters should be careful when recruiting workers from multiple geographic time zones. For one worker, the session might start in the morning, while for another, it may be during the night. It could be beneficial to split the work required in multiple batches spread across the whole 24-hour timespan. Alternatively, requesters could provide a high enough time frame for workers to complete a session, with some of them suggesting 24 to 48 hours. Also, it should not pass too much time between each session, as workers prefer a daily participation frequency (Figure 15). We thus argue that they may become bored or not recall the overall study, and thus drop participation halfway through.

The requester could also consider allowing workers to skip one or more sessions to provide additional flexibility, especially considering that the presence of eventual penalizations is not an aspect to further motivate workers in participating, according to them (Figure 19).

6.2.3 R3: *Workers Fear Performance Measurement.* Crowdsourcing platforms measure worker performances and quality using various metrics and indicators, such as the time elapsed between accepting a given HIT and its successful submission and the overall completion rate. These indicators can be used by task requesters to filter the pool of available workers as, indeed, we ourselves have done (Section 5.2).

Workers suggest that they might avoid participating in longitudinal studies because they somehow believe that this could increase the odds of being rejected at any time after a given session, once completed, thus impacting

the completion rates and performance as measured by the platform. In other words, workers fear performance measurement, especially in the context of longitudinal studies (Section 6.1.9).

A way to address such an issue is by disclosing and clarifying the whole study's workflow, having a particular focus on the rejection criteria. They should be described accurately along with the behaviors and causes that may trigger them.

6.2.4 R4: *Longitudinal Studies Boost Reliability And Trustworthiness.* Even though longitudinal studies might increase the fear of performance indicators, task requesters should remember that workers find such kinds of studies more reliable than other types of crowdsourcing-based studies, as they point out when asked about their loyalty and commitment in Section 6.1.9.

Such reliability refers to the fact that workers find longitudinal studies to be more operational, as the same work is repeated over time, allowing for better productivity. Also, they think that longitudinal studies allow for avoiding spending time searching for new tasks, as shown by Figure 21.

They also suggest, in the answers analyzed in Section 6.1.25, that a successful longitudinal study demonstrates researcher honesty, increasing the overall trustworthiness. Hence, task requesters should employ a well-documented task design which is as consistent as possible across sessions, having a sound and understandable sequence of events. Turning back to Figure 21, it can be seen that several workers also find that a way to increase trust on requester is planning intermediate payments.

6.2.5 R5: *Worker Provenance Affects Their Availability.* Crowdsourcing platforms allow task requesters to recruit people from all over the world. This may include workers from countries characterized by not adequate network infrastructure. For instance, when considering the Toloka platform it is rather easy to find people from CIS countries [43] (Commonwealth of Independent States), as reported by a worker.

Task requesters should carefully consider where to recruit each worker since their provenance can affect profoundly their availability, loyalty, and commitment. For instance, when asked about the suitability of the platform for longitudinal studies in Section 6.1.13, a worker specifically points out that it needs further improvements for this specific studies such as for scheduling sessions, and infrastructural factors may further exacerbate issues which are platform intrinsic.

6.2.6 R6: *Design Cross-Device Layouts And Avoid Requiring Additional Software.* Workers may use various devices to perform crowdsourcing tasks. For instance, the Prolific platform offers task requesters a user interface control to explicitly allow the usage of certain device classes. Moreover, a worker could start working on a given device and then switch to another one, at a later time. This can be particularly true for longitudinal studies since they are made of different sessions that can be performed over an arbitrary amount of days.

The requester should thus design and build a layout as cross-platform as possible, thus offering the possibility of using different devices. However, workers do not necessarily agree with being required to download additional software to perform a crowdsourcing task, as they point out both in Section 6.1.9 and Section 6.1.25. Task requesters should aim to provide a single (and possibly web-based) interface where the workers can perform the work required, whenever possible.

6.2.7 R7: *Provide Partial Payments And Consider Bonuses.* The most important incentives to foster longitudinal studies' availability on crowdsourcing platforms and motivate workers in participating and completing them are those related to monetary aspects such as reward and bonuses, as shown in Figure 19.

While in a crowdsourcing setting, indeed, both terms refer to some kind of monetary compensation, the differences lie in the modalities by which they are provided. Usually, the reward is the payment planned at the task's completion, while a bonus might be implicit or provided based on workers' performance, among other factors. When further narrowing the focus to crowdsourcing-based longitudinal studies, the bonus is provided after completing all the sessions of a longitudinal study, or parts thereof, as done by Strickland and Stoops [75].

Task requesters should thus consider planning a reward after each individual session and one or more bonuses scattered throughout the study to minimize worker drop-off. The partial reward could be a fixed amount or initially low, increasing as the study progresses, as an incentive for consistent participation. Such a decision might help reduce the workers' abandonment rate by further motivating them, and using an incremental form of payment helps contain the expenses during the initial stages.

6.2.8 R8: *Consider Deploying Pilot And Training Versions.* Piloting a task to be performed helps reduce worker attrition due to errors and unexpected scenarios within its business logic, and longitudinal studies do not make an exception. In Section 6.1.25, the workers suggest that using good screeners can both help requesters find participants that fit the needs of the study, as well as participants that are less likely to quit part-way through.

Related to this, longitudinal studies may involve recruiting novice workers during subsequent sessions, as done by Roitero et al. [68]. Task requesters may consider deploying a lightweight training version of the task to be performed. This will help first-timers and prepare them to perform the overall study as expected.

6.3 RQ3: Best Practices For Crowdsourcing Platforms

In the past, researchers have conducted longitudinal studies on crowdsourcing platforms to a certain extent. However, the support for such studies by commercial platforms is not as straightforward as it may seem.

Through our analysis of workers' responses and our experience in deploying a crowdsourcing task, we have discovered that even simple goals, such as tracking the overall progress of the study for requesters and workers, are not easily achievable. As a result, we have synthesized a list of 5 best practices that we believe the designers of crowdsourcing platforms should adopt and prioritize to adequately support longitudinal studies (Section 6.3.1–6.3.5).

6.3.1 BP1: *Allow Requesters Sending Reminders To Workers.* One of the most pressing issues reported by the workers is the need of being reminded of an upcoming session when committing to a longitudinal study (Section 6.1.13 and Section 6.1.25). For instance, a worker answered by reporting that they enjoyed participating because he had been reminded daily. Several workers also believe that longitudinal studies are not optimally supported in general, and this could be part of the problem. Hence, the crowdsourcing platform should allow task requesters to remind workers somehow.

A solution could involve allowing automatic reminders to be scheduled. These reminders could be scheduled after each session or after a fixed amount of time, and they should include a customizable message if needed. The reminders could be sent as notifications on the platform's user interface or as simple email messages.

6.3.2 BP2: *Report To Workers The Overall Progress.* Workers often express a desire to perceive and understand their progress within a longitudinal study (Section 6.1.25). This desire is further motivated by the fact that some of them feel incentivized by their personal interest in the task, both in participating, as shown in Figure 9, and in completing it (Figure 11).

Similarly to reminding workers, allowing them to understand their progress within a longitudinal study seems a reasonable requirement at a first glance, yet it is hardly achievable on the platforms considered, as they generally provide feedback to the worker only within a single crowdsourcing task (i.e., a single session of the overall study).

One solution to provide feedback to the worker and build a sense of progress could be allowing requesters to display in advance the number of sessions of the whole study. Also, workers reported that they enjoy participating in longitudinal studies to monitor the changes in answers over time. This could thus be another interesting piece of information to be summarized and shown as a performance indicator.

6.3.3 BP3: *Support More Advanced Worker Recruitment Strategies.* Roitero et al. [68] designed and conducted a longitudinal study that involved asking workers to fact-check statements related to the COVID-19 pandemic

delivered by public figures, such as politicians. A particular aspect of their study is that they republished a fixed set of HITs four times. Each time, they contacted the workers who previously participated, asking them to repeat the fact-checking activity. They also recruited novice workers to compare the work of the two groups.

Given that workers are willing to commit to a longitudinal study for roughly 22 days on average, as shown in Figure 13, it is natural to assume that many of them will drop out of participation, also considering that they consider study length as a major reason for doing so (Figure 14), and even though almost everyone claim completion of previous longitudinal studies (Figure 10). Indeed, Roitero et al. [68] measured task abandonment [33], reporting a 50% abandonment ratio on average.

In light of the case considered, the crowdsourcing platform should, first and foremost, offer a simple method to facilitate the recruitment of workers based not only on demographic criteria but also on their previous participation in the study. Additionally, it should provide a way to compensate for the reduced number of returning workers by explicitly asking the requester whether they want to recruit novice workers as well. As of today, Prolific somehow mitigates this by allowing for saving lists of worker groups that can be used to select the exact same participants for new studies.

6.3.4 `BP4`: *Add Adequate User Interface Filters For The Workers.* When designing and publishing a study on a crowdsourcing platform, it is not possible to indicate that it will be conducted in a longitudinal fashion, by publishing additional sessions over time. Given that Figure 21 shows that several workers believe longitudinal studies enable them to avoid spending regular time searching for new tasks and allow them to be more productive, we suggest that platforms provide workers with a user interface filter to separate longitudinal studies from standard tasks. Consequently, platforms should offer requesters the option to choose whether their studies will be longitudinal or not.

While the idea of adding adequate user interface filters may seem obvious, the workers on every platform considered can only guess or rely on the study descriptions provided by the requester to understand whether they are going to participate in some kind of longitudinal study. Implementing this change will thus raise their awareness and facilitate participation, allowing task requesters to optimize the time needed to recruit the required number of workers.

6.3.5 `BP5`: *Provide Support For Non-Desktop Devices.* Workers use a multitude of devices to participate in crowdsourcing tasks of any kind. Among the platforms considered, only Prolific allows task requesters to indicate the type of device class (i.e., mobile, desktop, or tablet) required to perform a given task, and workers can filter the available tasks accordingly. Specifically, a worker reported participating in a longitudinal study that involved maintaining a log on an Android device, along with collecting certain health data (e.g., heartbeat, etc.), which indeed took place on Prolific, while answering about platform suitability (Section 6.1.13).

Thus, crowdsourcing platforms should provide task requesters with a way to design a layout suitable for each device class. This could be achieved by offering a set of predefined and responsive user interface components, as done to some extent by MTurk with its Crowd HTML elements,[5] or by Toloka with its template builder.[6] The issue with these two approaches, however, is that they require considerable web development skills. Prolific, on the other hand, started moving in October 2023 towards such a direction by rolling out a survey builder that can be used to design simple polls consisting of 1-5 questions as of today.[7]

To further improve support for as many devices as possible, the platform could provide a way to design different layouts for the same task, one for each device class supported. Then, the workers should be allowed to choose studies compatible with a certain device class using an appropriate filter, similar to the choice between

[5] https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMturkAPI/ApiReference_HTMLCustomElementsArticle.html
[6] https://toloka.ai/knowledgebase/interface/
[7] https://researcher-help.prolific.com/hc/en-gb/articles/5484164151836-Survey-builder

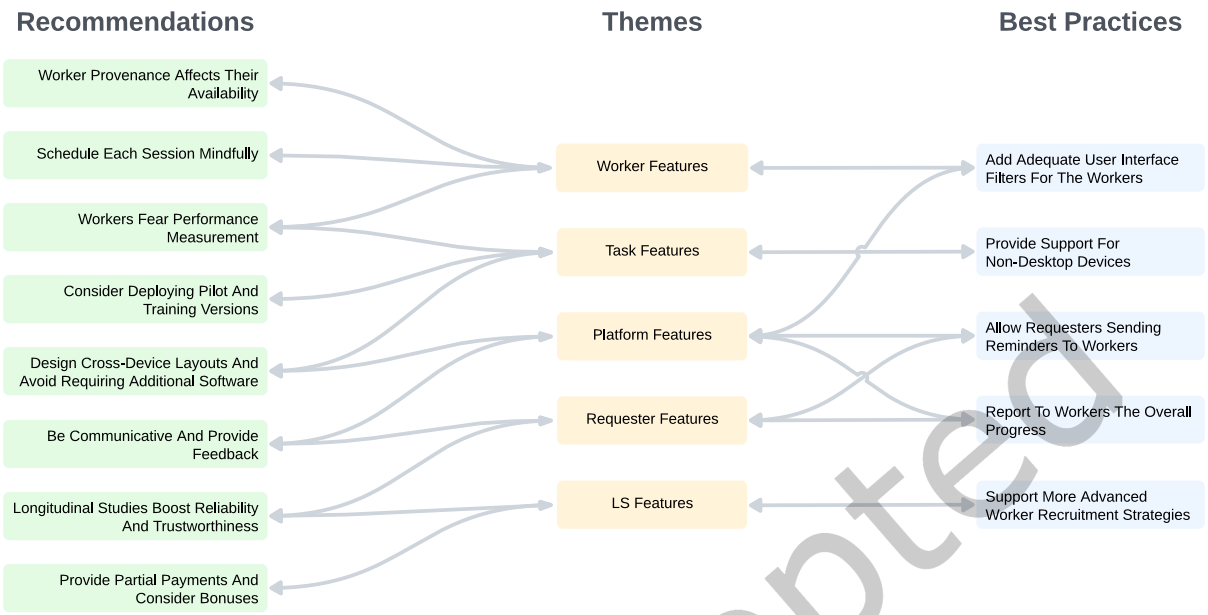

Fig. 23. Summary of the barriers emerged from our analyses, along with 8 recommendations for researchers and practitioners, 5 best practices for crowdsourcing platforms, and their interconnections.

participating in a longitudinal or standard study. This best practice is general and not limited to longitudinal study design.

7 Discussion

We recall the scope of our work and summarize our findings in light of the research questions in Section 7.1. Then, we acknowledge its limitations in Section 7.2 and sketch out future work in Section 7.3.

7.1 Summary

In this paper, we explored the barriers faced in conducting longitudinal tasks on crowdsourcing platforms, focusing on the perspective of workers. Through a large-scale survey across three major platforms, we examined the current perception, popularity, motivational factors, strengths, and weaknesses of longitudinal studies on these platforms.

We used both quantitative and qualitative analyses to gain insights, relying on an inductive thematic analysis for qualitative data. Integrating our findings with our experience as task requesters, we present an overview of results and their interconnections in Figure 23.

Our findings identified several barriers to longitudinal studies on crowdsourcing platforms (RQ1). These barriers fall into five main themes, outlined in the central part of Figure 23. Workers' activities are shaped by their needs, expectations, motivations, and fears. Task characteristics, platform design, and requester influence can also impact study outcomes. Difficulties may emerge from the longitudinal nature of the studies (Table 1).

For instance, our study found that workers were more likely to engage in longitudinal studies with higher payments, improved communication, and clear progress tracking. However, challenges arose from the lack of effective quality control mechanisms and transparent communication channels between requesters and workers.

Conducting successful crowdsourcing-based longitudinal studies poses significant challenges, with reported worker abandonment rates ranging from 50% to 80% [10, 37, 44, 52, 68, 73]. Existing literature lacks unified platform support or established best practices. Nonetheless, our recommendations and best practices aim to guide task requesters in improving the likelihood of success.

We provide 8 recommendations (RQ2) for researchers and practitioners to effectively design and conduct longitudinal studies on commercial micro-task crowdsourcing platforms, summarized in the leftmost part of Figure 23. Additionally, we propose 5 best practices (RQ3) for platforms to support successful longitudinal studies conducted via crowdsourcing, outlined in the rightmost part of Figure 23.

By following these recommendations, researchers and practitioners can overcome barriers to conducting successful longitudinal studies and leverage the benefits of crowdsourcing platforms. Implementing the suggested best practices would enhance the experience for both task requesters and workers.

7.2 Limitations

A limitation of our work is that we set parameters on crowdsourcing platforms (i.e., Amazon Mechanical Turk and Prolific) or directly ask workers (Toloka) to ensure the recruitment of workers with previous experience in longitudinal studies, as described in Section 5.1. However, we acknowledge that recruiting a sufficient number of experienced workers alone may not provide a complete understanding of longitudinal studies and the dynamics of workers in this context.

The survey design had two limitations. First, some questions would have yielded more insights with a Likert scale rather than binary responses (i.e., Section 6.1.11). Secondly, certain questions used single-choice radio buttons where multiple-choice options would have been more appropriate, potentially biasing responses (i.e., Section 6.1.10). It is noteworthy that cognitive biases may have played a role in shaping certain responses of workers [20].

Another limitation is due to the relatively small sample size recruited from three platforms, possibly not fully representing community heterogeneity despite achieving statistical significance in several survey questions. Additionally, the absence of behavioral data makes it difficult to assess whether implementing the survey results would achieve desired outcomes.

We argue that workers' backgrounds and demographics may impact their participation in longitudinal studies. For example, younger workers might have more time and respond differently to survey questions compared to older workers. While platforms like Prolific and Toloka provide demographic data, Amazon Mechanical Turk does not. Future studies could include questions to gather such information or use platform-specific criteria. For instance, recruiting workers from various age groups is feasible across all platforms, unlike other characteristics.

7.3 Future Work

In our future work, we aim to expand our findings through individual interviews with crowd workers. These interviews will help us better comprehend the motivations behind workers' engagement in longitudinal studies on crowdsourcing platforms. Additionally, we intend to interview task requesters to explore the obstacles they encounter in designing and conducting successful longitudinal studies, seeking potential solutions.

We also plan to conduct intervention studies to test new features and experimental settings on crowdsourcing platforms. Our goal is to enhance worker retention and satisfaction for both participants and requesters by assessing the effectiveness of these interventions.

Since we collect non-behavioral data, an interesting avenue for future work involves replicating our current setup and comparing new workers recruited from each platform. This aims to assess the robustness of our findings.

Another potential direction for future work to enhance our insights involves testing different combinations of our recommendations and best practices. By estimating the marginal effect of each practice, we can help the research community understand if perceptions of longitudinal studies among workers can be improved. Future work can also explore how platforms can better support longitudinal pilot studies [57].

The outlined future work will help us create a more robust process for conducting longitudinal studies on crowdsourcing platforms. This will benefit both workers and requesters involved.

8 Conclusions

Crowdsourcing platforms have gained increasing attention in the academic and business circles as valuable tools for data collection and analysis. Nevertheless, conducting longitudinal studies on these platforms poses significant challenges due to various factors.

Our contribution provides different practical implications. Theoretically, it enriches the crowdsourcing literature by exploring diverse worker motivations in longitudinal studies, extending beyond mere remuneration. Practically, it provides guidelines for optimizing longitudinal study design and management on crowdsourcing platforms, enhancing engagement and effectiveness.

By pursuing this line of research, we aim to contribute significantly to the expanding knowledge on crowdsourcing platforms and offer valuable insights for researchers, practitioners, and the platforms themselves. Ultimately, we seek to enhance the success of longitudinal studies on these platforms, benefiting both workers and requesters.

Acknowledgments

We thank all crowd workers who participated in our study. This research is supported by the European Union's NextGenerationEU PNRR M4.C2.1.1 – PRIN 2022 project "20227F2ZN3 MoT–The Measure of Truth: An Evaluation-Centered Machine-Human Hybrid Framework for Assessing Information Truthfulness" (20227F2ZN3_001, CUP G53D23002800006), and by the Strategic Plan of the University of Udine–Interdepartmental Project on Artificial Intelligence (2020-2025). This work is partially supported by the TU Delft AI Initiative and the Delft Design@Scale AI Lab.

References

[1] Tahir Abbas and Ujwal Gadiraju. 2022. Goal-Setting Behavior of Workers on Crowdsourcing Platforms: An Exploratory Study on MTurk and Prolific. In *Proceedings of the Tenth AAAI Conference on Human Computation and Crowdsourcing*, Jane Hsu and Ming Yin (Eds.), Vol. 10. AAAI Press, Washington, DC, 2–13. https://doi.org/10.1609/hcomp.v10i1.21983

[2] Hervé Abdi and Lynne J. Williams. 2010. Tukey's Honestly Significant Difference (HSD) Test. *Encyclopedia of Research Design* 3, 1 (2010), 1–5. https://personal.utdallas.edu/~Herve/abdi-HSD2010-pretty.pdf

[3] Asmaa Aljohani and James Jones. 2021. Conducting Malicious Cybersecurity Experiments on Crowdsourcing Platforms. In *The 2021 3rd International Conference on Big Data Engineering* (Shanghai, China) *(BDE 2021)*. Association for Computing Machinery, New York, NY, USA, 150–161. https://doi.org/10.1145/3468920.3468942

[4] Asmaa Aljohani and James Jones. 2022. The Pitfalls of Evaluating Cyber Defense Techniques by an Anonymous Population. In *HCI for Cybersecurity, Privacy and Trust*. Springer International Publishing, Cham, 307–325.

[5] Hal R. Arkes and Catherine Blumer. 1985. The Psychology of Sunk Cost. *Organizational Behavior And Human Decision Processes* 35, 1 (1985), 124–140. https://doi.org/10.1016/0749-5978(85)90049-4

[6] Elena M. Auer, Tara S. Behrend, Andrew B. Collmus, Richard N. Landers, and Ahleah F. Miles. 2021. Pay For Performance, Satisfaction And Retention In Longitudinal Crowdsourced Research. *PLOS ONE* 16, 1 (1 2021), 1–17. https://doi.org/10.1371/journal.pone.0245460

[7] Karen W. Bauer. 2004. Conducting Longitudinal Studies. *New Directions for Institutional Research* 2004, 121 (2004), 75–90. https://doi.org/10.1002/ir.102

[8] Raquel Benbunan-Fich. 2023. To Pay or Not to Pay? Handling Crowdsourced Participants Who Drop Out From a Research Study. *Ethics and Information Technology* 25, 3 (30 Jun 2023), 34. https://doi.org/10.1007/s10676-023-09708-8

[9] Gabriel A. Brooks and Luke Clark. 2023. The Gamblers of The Future? Migration From Loot Boxes to Gambling in a Longitudinal Study of Young Adults. *Computers in Human Behavior* 141 (2023), 107605. https://doi.org/10.1016/j.chb.2022.107605

[10] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science* 6, 1 (2011), 3–5. https://doi.org/10.1177/1745691610393980

[11] Michele Cantarella and Chiara Strozzi. 2021. Workers in The Crowd: The Labor Market Impact of The Online Platform Economy. *Industrial and Corporate Change* 30, 6 (7 2021), 1429–1458. https://doi.org/10.1093/icc/dtab022

[12] Edward Joseph Caruana, Marius Roman, Jules Hernández-Sánchez, and Piergiorgio Solli. 2015. Longitudinal Studies. *Journal of Thoracic Disease* 7, 11 (2015), 537–540. https://doi.org/10.3978/j.issn.2072-1439.2015.10.63

[13] Chun-Wei Chiang, Anna Kasunic, and Saiph Savage. 2018. Crowd Coach: Peer Coaching for Crowd Workers' Skill Growth. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW, Article 37 (11 2018), 17 pages. https://doi.org/10.1145/3274306

[14] Junsang Cho, Grayson W. Zhang, Wendy Armstrong, Daniel Cho, and Susan M. Culican. 2023. Crowdsourcing and Its Applications to Ophthalmology. *Expert Review of Ophthalmology* 18, 2 (2023), 113–119. https://doi.org/10.1080/17469899.2023.2200935

[15] Timothy M Daly and Rajan Nataraajan. 2015. Swapping Bricks For Clicks: Crowdsourcing Longitudinal Data on Amazon Turk. *Journal of Business Research* 68, 12 (2015), 2603–2609. https://doi.org/10.1016/j.jbusres.2015.05.001

[16] Ting Dang, Jing Han, Tong Xia, Dimitris Spathis, Erika Bondareva, Chloë Siegele-Brown, Jagmohan Chauhan, Andreas Grammenos, Apinan Hasthanasombat, R Andres Floto, Pietro Cicuta, and Cecilia Mascolo. 2022. Exploring Longitudinal Cough, Breath, and Voice Data for COVID-19 Progression Prediction via Sequential Deep Learning: Model Development and Validation. *Journal of Medical Internet Research* 24, 6 (21 6 2022), e37004. https://doi.org/10.2196/37004

[17] L. Dayton, W. Song, I. Kaloustian, E.L. Eschliman, J.C. Strickland, and C. Latkin. 2022. A Longitudinal Study of COVID-19 Disclosure Stigma and COVID-19 Testing Hesitancy in the United States. *Public Health* 212 (2022), 14–21. https://doi.org/10.1016/j.puhe.2022.08.003

[18] Esra Cemre Su de Groot and Ujwal Gadiraju. 2024. "Are we all in the same boat?" Customizable and Evolving Avatars to Improve Worker Engagement and Foster a Sense of Community in Online Crowd Work. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI</state>, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 640, 26 pages. https://doi.org/10.1145/3613904.3642429

[19] Djellel Difallah, Michele Catasta, Gianluca Demartini, and Philippe Cudré-Mauroux. 2014. Scaling-Up the Crowd: Micro-Task Pricing Schemes for Worker Retention and Latency Improvement. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 2. AAAI Press, Washington, DC, 50–58. https://doi.org/10.1609/hcomp.v2i1.13154

[20] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. 2021. A Checklist to Combat Cognitive Biases in Crowdsourcing. *Proceedings of the Ninth AAAI Conference on Human Computation and Crowdsourcing* 9, 1 (10 2021), 48–59. https://doi.org/10.1609/hcomp.v9i1.18939

[21] David Durward, Ivo Blohm, and Jan Marco Leimeister. 2020. The Nature of Crowd Work and its Effects on Individuals' Work Perception. *Journal of Management Information Systems* 37, 1 (2020), 66–95. https://doi.org/10.1080/07421222.2019.1705506

[22] Tom Edixhoven, Sihang Qiu, Lucie Kuiper, Olivier Dikken, Gwennan Smitskamp, and Ujwal Gadiraju. 2021. Improving Reactions to Rejection in Crowdsourcing Through Self-Reflection. In *Proceedings of the 13th ACM Web Science Conference 2021* (Virtual Event, United Kingdom) *(WebSci '21)*. Association for Computing Machinery, New York, NY, USA, 74–83. https://doi.org/10.1145/3447535.3462482

[23] Shaoyang Fan, Ujwal Gadiraju, Alessandro Checco, and Gianluca Demartini. 2020. CrowdCO-OP: Sharing Risks and Rewards in Crowdsourcing. *Proceedings of ACM Human-Computer Interaction* 4, CSCW2, Article 132 (oct 2020), 24 pages. https://doi.org/10.1145/3415203

[24] Zachary Fulker and Christoph Riedl. 2023. Cooperation in Crowd Work: Attitude and Perception of Freelancers on a Knowledge Work Platform. https://doi.org/10.48550/arXiv.2301.08808

[25] Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. 2017. Modus Operandi of Crowd Workers: The Invisible Role of Microtask Work Environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–29.

[26] Ujwal Gadiraju and Gianluca Demartini. 2019. Understanding Worker Moods and Reactions to Rejection in Crowdsourcing. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media* (Hof, Germany) *(HT '19)*. Association for Computing Machinery, New York, NY, USA, 211–220. https://doi.org/10.1145/3342220.3343644

[27] Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. 2015. Human Beyond the Machine: Challenges and Opportunities of Microtask Crowdsourcing. *IEEE Intelligent Systems* 30, 4 (2015), 81–85. https://doi.org/10.1109/MIS.2015.66

[28] Ujwal Gadiraju, Sebastian Möller, Martin Nöllenburg, Dietmar Saupe, Sebastian Egger-Lampl, Daniel Archambault, and Brian Fisher. 2017. Crowdsourcing Versus the Laboratory: Towards Human-Centered Experiments Using the Crowd. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*, Daniel Archambault, Helen Purchase, and Tobias Hoßfeld (Eds.). Springer International Publishing, Cham, 6–26. https://doi.org/10.1007/978-3-319-66435-4_2

[29] Snehalkumar (Neil) S. Gaikwad, Durim Morina, Adam Ginzberg, Catherine Mullings, Shirish Goyal, Dilrukshi Gamage, Christopher Diemert, Mathias Burton, Sharon Zhou, Mark Whiting, Karolina Ziulkoski, Alipta Ballav, Aaron Gilbee, Senadhipathige S. Niranga, Vibhor Sehgal, Jasmine Lin, Leonardy Kristianto, Angela Richmond-Fuller, Jeff Regino, Nalin Chhibber, Dinesh Majeti, Sachin Sharma, Kamila Mananova, Dinesh Dhakal, William Dai, Victoria Purynova, Samarth Sandeep, Varshine Chandrakanthan, Tejas Sarma, Sekandar Matin, Ahmed Nasser, Rohit Nistala, Alexander Stolzoff, Kristy Milland, Vinayak Mathur, Rajan Vaish, and Michael S. Bernstein. 2016. Boomerang: Rebounding the Consequences of Reputation Feedback on Crowdsourcing Platforms. In *Proceedings of the 29th Annual*

Symposium on User Interface Software and Technology (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 625–637. https://doi.org/10.1145/2984511.2984542

[30] S.R. Goodwin, M.J. Dwyer, S.L. Caliva, C.A. Burrows, and B.R. Raiff. 2023. Using Reddit as a Recruitment Strategy for Addiction Science Research. *Journal of Substance Use and Addiction Treatment* 148 (2023), 209011. https://doi.org/10.1016/j.josat.2023.209011

[31] Annetta Grant, Henri Weijo, and Peter A. Dacin. 2023. How Institutional Logics Shape Fairness in Crowdsourcing: The Case of Threadless. *International Journal of Research in Marketing* 40, 2 (2023), 378–397. https://doi.org/10.1016/j.ijresmar.2022.10.002

[32] Ashish Gurung, Sami Baral, Morgan P. Lee, Adam C. Sales, Aaron Haim, Kirk P. Vanacore, Andrew A. McReynolds, Hilary Kreisberg, Cristina Heffernan, and Neil T. Heffernan. 2023. How Common Are Common Wrong Answers? Crowdsourcing Remediation at Scale. In *Proceedings of the Tenth ACM Conference on Learning @ Scale* (Copenhagen, Denmark) *(L@S '23)*. Association for Computing Machinery, New York, NY, USA, 70–80. https://doi.org/10.1145/3573051.3593390

[33] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. All Those Wasted Hours: On Task Abandonment in Crowdsourcing. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne VIC, Australia) *(WSDM '19)*. Association for Computing Machinery, New York, NY, USA, 321–329. https://doi.org/10.1145/3289600.3291035

[34] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174023

[35] Kenji Hata, Ranjay Krishna, Li Fei-Fei, and Michael S. Bernstein. 2017. A Glimpse Far into the Future: Understanding Long-term Crowd Worker Quality. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, USA, 889–901. https://doi.org/10.1145/2998181.2998248

[36] Danula Hettiachchi, Vassilis Kostakos, and Jorge Goncalves. 2022. A Survey on Task Assignment in Crowdsourcing. *Comput. Surveys* 55, 3, Article 49 (feb 2022), 35 pages. https://doi.org/10.1145/3494522

[37] Christopher J. Holden, Trevor Dennie, and Adam D. Hicks. 2013. Assessing The Reliability of The M5-120 on Amazon's Mechanical Turk. *Computers in Human Behavior* 29, 4 (2013), 1749–1754. https://doi.org/10.1016/j.chb.2013.02.020

[38] Jeff Howe. 2006. The Rise of Crowdsourcing. *Wired Magazine* 14, 6 (2006), 1–4. https://www.wired.com/2006/06/crowds/

[39] Hsiu-Fang Hsieh and Sarah E. Shannon. 2005. Three Approaches to Qualitative Content Analysis. *Qualitative Health Research* 15, 9 (2005), 1277–1288. https://doi.org/10.1177/1049732305276687

[40] Lijuan Huang, Guojie Xie, John Blenkinsopp, Raoyi Huang, and Hou Bin. 2020. Crowdsourcing for Sustainable Urban Logistics: Exploring the Factors Influencing Crowd Workers' Participative Behavior. *Sustainability* 12, 8 (4 2020), 1–20. https://doi.org/10.3390/su12083091

[41] Lilly C. Irani and M. Six Silberman. 2013. Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 611–620. https://doi.org/10.1145/2470654.2470742

[42] Ji-Youn Jung, Sihang Qiu, Alessandro Bozzon, and Ujwal Gadiraju. 2022. Great Chain of Agents: The Role of Metaphorical Representation of Agents in Conversational Crowdsourcing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 57, 22 pages. https://doi.org/10.1145/3491102.3517653

[43] Paul Kubicek. 2009. The Commonwealth of Independent States: An Example of Failed Regionalism? *Review of International Studies* 35, S1 (2009), 237–256. https://doi.org/10.1017/S026021050900850X

[44] Klodiana Lanaj, Russell E. Johnson, and Christopher M. Barnes. 2014. Beginning The Workday Yet Already Depleted? Consequences of Late-night Smartphone Use and Sleep. *Organizational Behavior and Human Decision Processes* 124, 1 (2014), 11–23. https://doi.org/10.1016/j.obhdp.2014.01.001

[45] Gabriel Shing-Koon Leung, Vincent Cho, and C. H. Wu. 2021. Crowd Workers' Continued Participation Intention in Crowdsourcing Platforms: An Empirical Study in Compensation-Based Micro-Task Crowdsourcing. *Journal of Global Information Management (JGIM)* 29, 6 (2021), 1–28. https://doi.org/10.4018/JGIM.20211101.oa13

[46] Sophia Xueying Li, Ramzi Halabi, Rahavi Selvarajan, Molly Woerner, Isabell Griffith Fillipo, Sreya Banerjee, Brittany Mosser, Felipe Jain, Patricia Arean, and Abhishek Pratap. 2022. Recruitment And Retention In Remote Research: Learnings From a Large, Decentralized Real-world Study. *JMIR Formative Research* 6, 11 (11 2022), e40765. https://doi.org/10.2196/40765

[47] Leib Litman, Jonathan Robinson, and Tzvi Abberbock. 2017. TurkPrime.com: A Versatile Crowdsourcing Data Acquisition Platform For The Behavioral Sciences. *Behavior Research Methods* 49, 2 (2017), 433–442. https://doi.org/10.3758/s13428-016-0727-z

[48] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of ACM Human-Computer Interaction* 3, CSCW, Article 72 (11 2019), 23 pages. https://doi.org/10.1145/3359174

[49] Martino Mensio, Gregoire Burel, Tracie Farrell, and Harith Alani. 2023. MisinfoMe: A Tool for Longitudinal Assessment of Twitter Accounts' Sharing of Misinformation. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*

(Limassol, Cyprus) *(UMAP '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 72–75. https://doi.org/10.1145/3563359.3597396

[50] Sandeep Mishra and R. Nicholas Carleton. 2017. Use of Online Crowdsourcing Platforms For Gambling Research. *International Gambling Studies* 17, 1 (2017), 125–143. https://doi.org/10.1080/14459795.2017.1284250

[51] Chung Jung Mun, Nina Winsick, Stephen T. Wegener, Shawn Youngsted, Claudia M. Campbell, and Rachel V. Aaron. 2023. Longitudinal Effects Of Insomnia And Evening Chronotype On Pain And Emotional Distress Among Individuals With Chronic Pain. *The Journal of Pain* 24, 4, Supplement (2023), 106–107. https://doi.org/10.1016/j.jpain.2023.02.302

[52] Chung Jung Mun Mun, Claudia M. Campbell, Lakeya S. McGill, Stephen T. Wegener, and Rachel V. Aaron. 2022. Trajectories and Individual Differences in Pain, Emotional Distress, and Prescription Opioid Misuse During the COVID-19 Pandemic: A One-Year Longitudinal Study. *The Journal of Pain* 23, 7 (2022), 1234–1244. https://doi.org/10.1016/j.jpain.2022.02.005

[53] Zahra Nouri, Ujwal Gadiraju, Gregor Engels, and Henning Wachsmuth. 2021. What Is Unclear? Computational Assessment of Task Clarity in Crowdsourcing. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media* (Virtual Event, USA) *(HT '21)*. Association for Computing Machinery, New York, NY, USA, 165–175. https://doi.org/10.1145/3465336.3475109

[54] Zahra Nouri, Nikhil Prakash, Ujwal Gadiraju, and Henning Wachsmuth. 2023. Supporting Requesters in Writing Clear Crowdsourcing Task Descriptions Through Computational Flaw Assessment. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 737–749. https://doi.org/10.1145/3581641.3584039

[55] Niwako Ogata, Hsin-Yi Weng, and Locksley Messam. 2023. Temporal Patterns of Owner-pet Relationship, Stress, and Loneliness During the COVID-19 Pandemic, and the Effect of Pet Ownership on Mental Health: A Longitudinal Survey. *PLOS ONE* 18, 4 (4 2023), 1–18. https://doi.org/10.1371/journal.pone.0284101

[56] Stephen F. Olejnik and James Algina. 2004. Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. *Psychological Methods* 8 (1 2004), 434–47. https://doi.org/10.1037/1082-989X.8.4.434

[57] Jonas Oppenlaender, Tahir Abbas, and Ujwal Gadiraju. 2024. The State of Pilot Study Reporting in Crowdsourcing: A Reflection on Best Practices and Guidelines. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–45.

[58] Stefan Palan and Christian Schitter. 2018. Prolific.ac—A Subject Pool For Online Experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

[59] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5 (2010), 411–419. https://doi.org/10.1017/S1930297500002205

[60] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond The Turk: Alternative Platforms For Crowdsourcing Behavioral Research. *Journal of Experimental Social Psychology* 70 (2017), 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

[61] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2022. Data quality of platforms and panels for online behavioral research. *Behavioral Research Methods* 54, 4 (9 2022), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

[62] Sabine Pfeiffer and Sandra Kawalec. 2020. Justice Expectations In Crowd And Platform-mediated Work. *The Economic and Labour Relations Review* 31, 4 (2020), 483–501. https://doi.org/10.1177/1035304620959750

[63] Robert E. Ployhart and Anna-Katherine Ward. 2011. The "Quick Start Guide" for Conducting and Publishing Longitudinal Research. *Journal of Business and Psychology* 26, 4 (1 12 2011), 413–422. https://doi.org/10.1007/s10869-011-9209-6

[64] Sihang Qiu, Alessandro Bozzon, Max V. Birk, and Ujwal Gadiraju. 2021. Using Worker Avatars to Improve Microtask Crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2, Article 322 (10 2021), 28 pages. https://doi.org/10.1145/3476063

[65] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Improving Worker Engagement Through Conversational Microtask Crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376403

[66] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Towards Memorable Information Retrieval. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval* (Virtual Event, Norway) *(ICTIR '20)*. Association for Computing Machinery, New York, NY, USA, 69–76. https://doi.org/10.1145/3409256.3409830

[67] Geetanjali Rajamani, Molly Diethelm, Melissa A. Gunderson, Venkata S. M. Talluri, Patricia Motz, Jennifer M. Steinhaus, Anne E. La Flamme, Bryan Jarabek, Tori Christiaansen, Jeffrey T. Blade, Sameer Badlani, and Genevieve B. Melton. 2023. Crowdsourcing Electronic Health Record Improvements at Scale across an Integrated Health Care Delivery System. *Applied Clinical Informatics* 14, 02 (10 5 2023), 356–364. https://doi.org/10.1055/s-0043-1767684

[68] Kevin Roitero, Michael Soprano, Beatrice Portelli, Massimiliano De Luise, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. 2021. Can The Crowd Judge Truthfulness? A Longitudinal Study on Recent Misinformation About COVID-19. *Personal and Ubiquitous Computing* 2021, 9 (16 9 2021), 1–31. https://doi.org/10.1007/s00779-021-01604-6

[69] Kevin Roitero, Michael Soprano, Beatrice Portelli, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. 2020. The COVID-19 Infodemic: Can the Crowd Judge Recent Misinformation Objectively?. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 1305–1314. https://doi.org/10.1145/3340531.3412048

[70] Sigrid Rouam. 2013. *False Discovery Rate (FDR).* Springer New York, New York, NY, 731–732. https://doi.org/10.1007/978-1-4419-9863-7_223

[71] Niloufar Salehi, Lilly C. Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. 2015. We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1621–1630. https://doi.org/10.1145/2702123.2702508

[72] Anna Schober, Linda Tizek, Emma K. Johansson, Agneta Ekebom, Jan-Erik Wallin, Jeroen Buters, Simon Schneider, and Alexander Zink. 2022. Monitoring Disease Activity of Pollen Allergies: What Crowdsourced Data Are Telling Us. *World Allergy Organization Journal* 15, 12 (2022), 100718. https://doi.org/10.1016/j.waojou.2022.100718

[73] Danielle N. Shapiro, Jesse Chandler, and Pam A. Mueller. 2013. Using Mechanical Turk to Study Clinical Populations. *Clinical Psychological Science* 1, 2 (2013), 213–220. https://doi.org/10.1177/2167702612469015

[74] Michael Soprano, Kevin Roitero, Francesco Bombassei De Bona, and Stefano Mizzaro. 2022. Crowd_Frame: A Simple and Complete Framework to Deploy Complex Crowdsourcing Tasks Off-the-Shelf. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) *(WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 1605–1608. https://doi.org/10.1145/3488560.3502182

[75] Justin C. Strickland and William W. Stoops. 2018. Feasibility, Acceptability, and Validity of Crowdsourcing for Collecting Longitudinal Alcohol Use Data. *Journal of the Experimental Analysis of Behavior* 110, 1 (2018), 136–153. https://doi.org/10.1002/jeab.445

[76] Justin C. Strickland and William W. Stoops. 2019. The Use of Crowdsourcing in Addiction Science Research: Amazon Mechanical Turk. *Experimental and Clinical Psychopharmacology* 27, 1 (2019), 1. https://doi.org/10.1037/pha0000235

[77] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second Chance for a First Impression? Trust Development in Intelligent System Interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 77–87. https://doi.org/10.1145/3450613.3456817

[78] Carlos Toxtli, Siddharth Suri, and Saiph Savage. 2021. Quantifying the Invisible Labor in Crowd Work. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2, Article 319 (oct 2021), 26 pages. https://doi.org/10.1145/3476060

[79] Rama Adithya Varanasi, Divya Siddarth, Vivek Seshadri, Kalika Bali, and Aditya Vashistha. 2022. Feeling Proud, Feeling Embarrassed: Experiences of Low-Income Women with Crowd Work. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 298, 18 pages. https://doi.org/10.1145/3491102.3501834

[80] Xiaohui Wang, Dion Hoe-Lian Goh, and Ee-Peng Lim. 2020. Understanding Continuance Intention toward Crowdsourcing Games: A Longitudinal Investigation. *International Journal of Human–Computer Interaction* 36, 12 (2020), 1168–1177. https://doi.org/10.1080/10447318.2020.1724010

[81] Mark E. Whiting, Grant Hugh, and Michael S. Bernstein. 2019. Fair Work: Crowd Work Minimum Wage with One Line of Code. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing*, Edith Law and Jennifer Wortman Vaughan (Eds.). AAAI Press, Washington, DC, 197–206. https://doi.org/10.1609/hcomp.v7i1.5283

[82] Alex C. Williams, Gloria Mark, Kristy Milland, Edward Lank, and Edith Law. 2019. The Perpetual Work Life of Crowdworkers: How Tooling Practices Increase Fragmentation in Crowdwork. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW, Article 24 (11 2019), 28 pages. https://doi.org/10.1145/3359126

[83] Meng-Han Wu and Alexander Quinn. 2017. Confusing The Crowd: Task Instruction Quality on Amazon Mechanical Turk. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 5. AAAI, USA, 206–215. https://doi.org/10.1609/hcomp.v5i1.13317

A Survey Questions

This appendix provides each question of the survey employed to investigate the barriers to running longitudinal tasks on crowdsourcing platforms. Section 5.1 and Section 5.2 provide the details concerning the overall design of the survey and the crowdsourcing task.

The questions are shown in order, as they were presented to the recruited workers. Each question is labeled with the corresponding survey part and index. When a question is labeled using sub-indexes, it means that it is nested in the survey. The text of each question is reported in *italics*, together with the expected answer type using normal font and additional details written using `monospaced` font. In Appendix A.1, several questions are labeled with the letter X. This labeling is a result of an explicit design choice in our crowdsourcing task, as described in Section 5.2.

A.1 P1: Current Perception Of Longitudinal Studies

1: *Have you ever participated in a longitudinal study in the past, even if on other platforms?*

1.1: *How many?*

- Integer number (X) in the interval $[0, 3]$, such as $0 \leq X \leq 3$

`numerical field, free text not allowed`

1.1.X: *Describe your experience with the longitudinal study nr. X*

1.1.X.1: *When was the study performed?*

- 1 month ago
- 2 months ago
- 3 to 5 months ago
- 6 to 12 months ago
- More than 1 year ago

`closed-ended, radio button, free text not allowed`

1.1.X.2: *How many sessions did the longitudinal study have?*

- Positive integer number

`numerical field, free text not allowed`

1.1.X.3: *Which was the time interval between each session?*

- 1 day
- 2 to 4 days
- 5 to 9 days
- 10 to 14 days
- 15 to 20 days
- 20 to 24 days
- 25 to 1 month
- 2 months
- 3 months
- 4 months
- 5 to 6 months
- 7 to 12 months
- More than 1 year
- Other (please, specify)

`closed-ended, radio button, free text allowed`

1.1.X.4: *What was the duration of each session?*

- 15 minutes
- 30 minutes
- 45 minutes
- 60 minutes
- 1 hour
- 2 hours
- 3 hours
- More than 3 hours
- Other (please, specify)

`closed-ended, radio button, free text allowed`

1.1.X.5: *Which was the crowdsourcing platform?*

- Amazon Mechanical Turk

- Prolific
- Toloka
- Other (please, specify)

`closed-ended, radio button, free text allowed`

1.1.X.6: *Which was the payment model?*

- Payment after each session
- Single final reward
- Other (please, specify)

`closed-ended, checkbox, free text allowed`

1.1.X.7: *How was your general satisfaction:*

1.1.X.7.1: *Would you participate in the same study again?*

- Yes
- No

`closed-ended, radio button, free text not allowed`

1.1.X.7.2: *Please, tell us why*

- Non-empty text

`textual field`

`question group`

1.1.X.8: *What was the main incentives that convince you into participating in the longitudinal study?*

- Bonus
- Reward
- Interest on task
- Altruism (to help the research)
- Because the task was educative
- Other (please, specify)

`closed-ended, radio-button, free text allowed`

1.1.X.9: *Did you complete the task?*

- *Yes*

1.1.X.9.1: *What were the main incentives that convinced you in completing the longitudinal study?*

- Bonus
- Reward
- Interest on task
- Altruism (to help the research)
- Because the task was educative

`closed-ended, radio-button, free text not allowed`

- *No*

1.1.X.9.2: *What are the reasons that made you dropout?*

- Non-empty text

`textual field`

`closed-ended, radio-button, free text allowed`

`question group, repeated X times`

`question group`

2: *Do you think this crowdsourcing platform is suitable to carry out longitudinal studies? Please, elaborate your answer*

– Non-empty text
`textual field`

3: *Longitudinal studies are not very common in crowdsourcing yet. Which of these statements do you agree with?*
– Longitudinal studies are not optimally supported by current popular crowdsourcing platforms
– Workers do not like to commit on daily effort
– Reward and incentives are insufficient
– Requesters do not need longitudinal participation since most of the tasks work with static data to annotate
– Other (please, specify)
`closed-ended, checkbox, free text allowed`

A.2 P2: Your Possible Participation And Commitment To Longitudinal Studies

1: *How many days would you be happy to commit to a longitudinal study (imagine a session of about 15 min per day)*
– Positive integer number
`numerical field, free text not allowed`

2: *Which of the following would make you refuse participation in a longitudinal study?*
– Too frequent
– Too long
– Other (please, specify)
`closed-ended, checkbox, free text allowed`

3: *What's your preferred frequency of participation in a longitudinal study?*
– Daily
– Every other day
– Weekly
– Biweekly
– Monthly
– Every six months
– Yearly
`closed-ended, radio button, free text not allowed`

4: *What is your preferred session duration (in hours)?*
– Positive integer number
`numerical field, free text not allowed`

5: *What do you consider an acceptable hourly payment for your work on this platform (in USD$ dollars)?*
– Positive integer number
`numerical field, free text not allowed`

6: *How much time would you be happy to allocate per day to work on longitudinal studies (in hours)?*
– Positive integer number
`numerical field, free text not allowed`

7: *Which incentives would most motivate you to participate and engage in longitudinal studies?*
– Final bonus to be awarded after the last contribution
– Payment after each session
– Progressive increment of payment
– Progressive decrement of payment
– Being penalized when skipping working sessions

- Work on different tasks type to increase engagement diversity
- Experimental variants of the same tasks to reduce repeatability
- Other (please, specify)

`closed-ended, checkbox, free text allowed`

8: *What types of tasks would you like to perform in a longitudinal study?*
- Information Finding - Such tasks delegate the process of searching to satisfy one's information need to the workers in the crowd. For example, "Find information about a company in the UK".
- Verification and Validation - These are tasks that require workers in the crowd to either verify certain aspects as per the given instructions, or confirm he validity of various kinds of content. For example, "Match the names of personal computers and verify corresponding information".
- Interpretation and Analysis - Such tasks rely on the wisdom of the crowd to use their interpretation skills during task completion. For example, "Choose the most suitable category for each URL".
- Content Creation - Such tasks usually require the workers to generate new content for a document or website. They include authoring product descriptions or producing question-answer pair. For example, "Suggest names for a new product".
- Surveys - Surveys about a multitude of aspects ranging from demographics to customer satisfaction are crowdsourced. For example, "Mother's Day and Father's Day Survey (18-29 year olds only!)".
- Content Access - These tasks require the workers to simply access some content. For example, "Click on the link and watch the video".
- Other (please, specify)

`closed-ended, checkbox, free text allowed`

9: *What do you think are the benefits of being involved in longitudinal studies?*
- No need to spend time regularly searching for new tasks to perform
- No need to learn how to do the job (Learning curve)
- Better productivity (more operationale)
- Intermediate payments would increase trust on requester
- Other (please, specify)

`closed-ended, checkbox, free text allowed`

10: *What do you think are the downsides that limit your interest in participating in longitudinal studies?*
- Lack of flexibility
- Long term commitment
- Reward assigned at the end
- Lack of diversity
- Other (please, specify)

`closed-ended, checkbox, free text allowed`

11: *Do you have any additional suggestions for a requester who plans to design an attractive longitudinal study?*
- Non-empty text

`textual field`

B Examples Of Workers' Responses

This appendix provides examples of the responses provided by the workers recruited for the three questions analyzed qualitatively, whose findings are reported in Section 6.1.

Table 7 reports examples for question 1.1.X.7.2 (P1 part of the survey), which addresses workers' loyalty and commitment to the reported longitudinal studies. Then, Table 8 provides examples for question 2 (P2 part), which is about the suitability of the platform of provenance in supporting longitudinal studies. Lastly, Table 9 reports examples for question 11 (P2 part) which asks workers to provide general suggestions and considerations.

Table 7. Sample of answers provided by workers concerning loyalty to longitudinal studies.

Worker Responses
It was a well-designed study and the requester was very specific about when the follow-up tasks would be posted, and they sent reminders as well.
I felt the study was interesting and the reward was excellent so happy to do it again
It was very well organized and efficient. I didn't have to wait much between sessions.
Because I find interesting seeing how differently sometimes my answers can be just after a few days due to changes in the circumstances.
I dont likes that participating in same studies again because of im afraid of getting rejected
As long as the daily tasks are short and do not require an app download of any sort, I'll do them. I don't like downloading software or committing much time. I also don't like time windows. I like doing studies when I have free time, not during required blocks of time.
The individual studies were well-compensated and there was a generous bonus for completing all sessions of the study. Other than that, the study itself was quite unique and enjoyable to complete.
It's interesting to participate in longitudinal studies because it's pleasant to help with a research that monitors our learning/evolution over time in a given subject. This particular study was a monitored study that checked my performance on a repetitive memory task over the weeks. Also, the reward was excellent.
There would be random alerts on my phone (the study work took place within an app but was paid via Prolific) and I really struggled over the course of the fortnight duration - I was effectively a slave to my phone.
I don't find them any different to normal single part studies other than they can be more repetitive but so long as they meet the minimum payment reward on Prolific then I don't have any issue and I don't even care about bonuses for completing all parts because I complete all studies that I am invited to anyway and with Prolific I get instant alerts but you also get e-mail invitations when you aren't available so you can always complete them later on, it is really impossible to miss them and because each part is paid separately and approved individually it is more trustworthy for both participant and researcher.

Table 8. Sample of answers provided by workers concerning the adequacy of crowdsourcing platforms in supporting longitudinal studies.

Worker Responses	**Platform**
I think that this platform is good for longitudinal studies, especially when a Requester can send email reminders to the Workers about when the follow-up tasks are available to be completed.	MTurk
Yes, I have done tasks like that on this platform before and it went well for me.	MTurk
I don't think so because everything that gets released gets snatched up quickly. Also, the requesters on this platform don't respond much. Before, yes but not most likely not.	MTurk
Yes but it need further improvements for this specific type of tasks such as scheduling improvements etc.	MTurk
Yes, I think it is perfectly suitable given its nature. I do think coordinating longer studies can be more difficult on mturk compared to other platforms, as there are many other studies constantly on the platform and remembering longitudinal studies can be difficult while also keeping up with regular studies. To remedy this, requestors must often use e-mail reminders and other types of reminders, which I have no issues with at all.	MTurk
Yes, I believe it is. This platform is the host of many other studies all of which provide for a professional and safe environment (on both sides, for the requester and surveyee with full disclosure of all procedures. I've had previous experience with a longitudinal study on this platform and I have zero complaints.	Prolific
Yes. The messaging system on Prolific is very useful in this regard, the platform itself can easily be tailored to longitudinal studies, and both the researcher and the participant can rely on Prolific for any support required around the task.	Prolific
Yes I think Prolific works very well, I have Prolific Assistant so get the alerts if I'm on my PC so usually I start them just like any other study but even if you don't then you would be sent an e-mail invitation to remind you so you are very unlikely to ever miss any part of a study and I have completed all parts of any longitudinal studies that I have been part of. I think so long as all of the details are explained in the first part and the participant agrees to complete all of the following parts then they should have very high success rates and if anyone does drop out or has any reason to you can also communicate this via Prolific messaging.	Prolific
Not really, there should be an option to separate normal from longitudinal studies.	Prolific
Yes, but Prolific does not email you outside of itself. This can be a problem if the study requires out-of-band responses. With Mechanical Turk your requests hit email so I get message reminders when I am not at my desk.	Prolific
Yes, it's a nice platform to work, to earn rewards and to learn some new things so it would be a great platform for longitudinal studies too.	Toloka

Continues in the next page

Table 8. Sample of answers provided by workers concerning the adequacy of crowdsourcing platforms in supporting longitudinal studies (cont.)

Worker Responses	Platform
Yes it fits. I think there is a large number of participants, which makes the study more accurate.	Toloka
I have had good experiences with tasks offered by Toloka. Proper instructions are provided.	Toloka
Yes, it has participants which login every or almost every day, they are interested in completing tasks they are already acquainted with.	Toloka
Yes, it is suitable because most people in this platform work more than five hours everyday	Toloka

Table 9. Sample of suggestions provided by workers concerning longitudinal studies.

Worker Responses
Establish the correct sequence of events, identify changes over time, and provide insight into cause-and-effect relationships.
Plan each session in a way that it makes the surveyee feel like the're making progress. Maybe at the end of each session highlight the differences in their previous answer to accentuate that feeling of progression.
Beside all of the aspects regarding time and money, fast communication between requester and worker and also regular feedbacks regarding workers task quality would be great to increase their (our :)) commitment.
Maybe offer different platforms on which to take the study (ie android, PC, mac, etc)
Just don't require downloads. Keep tasks short. No time frames.
A lot of us work from home and are self employed so we have to pay tax on these earnings. As long as it pays a decent amount for the time taken (at least £6 per hour), I would be more than happy to take part.
It is useful to allow one or two sessions to be skipped if the responder can't commit to absolutely every session.
Be reasonable with what you expect people to do. People who work full time and have caring responsibilities won't necessarily have the capacity/flexibility to do daily tasks that last an hour or more. If your study makes those demands then you're going to only be getting a certain kind of participant (e.g. unemployed).
Keep them to the point, don't give long, fatigued instructions, try not to ask the same question fifty different ways. Also, if you have a game, games are very attractive for me; I'd be interested in longitudinal studies where we have to play a game and collect something, like points, or something. And gives a good bonus! Good base pay, as well. At least 12 dollars an hour.

Continues in the next page

Table 9. Sample of suggestions provided by workers concerning longitudinal studies (cont.)

Worker Responses
Ensure the timings are not onerous when considering participants from multiple geographic zones - they need adequate time to complete. A final bonus payment completion incentive helps reduce attrition - and on that note, keep the study shorter (say 2 weeks) to minimise participant drop-off.
I think you have to be as revealing as possible in the first part of the study so the participant knows in advance what they are signing up for, it would help if the participant gets a good idea or sampling of the task in full so there are no surprises if that is possible so it would be good to have them complete the worst part of it if there is one and if it is repetitive and hard to complete over a longer period then to explain that so they can make a judgement. So long as they know what is involved and what is expected of them in advance before they then agree to take part because then so long as they understand the commitment they are making and the schedule and timing they should be able to complete it.
Using good screeners can both help requesters find participants that fit the needs of the study, as well as participants that are less likely to quit part-way through. Also, compensation schemes that reward consistent participation are likely to increase the odds that participants complete all required sessions of the study.

Received 7 September 2023; revised 7 June 2024; accepted 10 June 2024