Tragedy of the Commons - A Critical Study of Data Quality and Validity Issues in Crowd Work-Based Research

Huichuan Xia
Department of Information Management, Peking University
huichuanxia@pku.edu.cn

Abstract

Academic scholars have leveraged crowd work platforms such as MTurk to conduct research and collect data. Though prior studies have discussed data quality and validity issues in crowd work via surveys and experiments, they kind of neglected to explore the scholars' and particularly the IRB's ethical concerns in these respects. In this study, we interviewed 17 scholars from six disciplines and 15 IRB directors and analysts in the U.S. to fill this research gap. We identified common themes among our respondents but also discovered distinctive and even opposing views regarding the approval rate, rejection, internal and external research validity. Based on the findings, we discussed a potential Tragedy of the Commons regarding the data quality deterioration and the disciplinary differences regarding validity in crowd work-based research. Finally, we advocated that the IRB's ethical concerns in crowd work-based research should be heard and respected further.

1. Introduction

Crowd work is a social, technical, and economic mode of production that harnesses collective intelligence from an undefined network of people [1,2] . Amazon Mechanical Turk (MTurk) is the most popular and well-known crowd work platform since its public launch in 2005. For over a decade, scholars in different disciplines from computer science to psychology have been leveraging crowd work platforms, particularly MTurk, to recruit participants and collect data for various research purposes, such as photo tagging, audio transcription, surveys, and experiments. In this paper, we call this type of academic research crowd work-based research. Previously, scholars have termed it more broadly as crowdsourcing research [3], the crowdsourcing model of research [4], or crowdsourced research [5]. We prefer a narrow term since crowdsourcing does not necessarily include a monetary incentive as research compensation. We refer to the academic research on MTurk as the epic representation of crowd work-based research because to date, MTurk is still the most popular crowd work platform for scholars [2,6] even though Prolific has obtained growing attention and traction in academia [7,8].

Crowd work-based research is popular for three main reasons. First, a crowd work platform is fast and convenient for scholars to recruit from a diverse population of crowd workers around the globe, almost without any time or geographical limit [9]. Second, a crowd work platform provides relatively affordable samples for scholars [10,11], particularly before any "minimum wage" was popularized as a rule of thumb for crowd work-based research compensation. Third, since MTurk's inception, many scholars have argued and defended that the data quality and validity in crowd work-based research are comparable or even superior to similar research conducted on the other venues [10,12,13]. This third merit of crowd work-based research is controversial because some scholars have also observed frequent cheaters on crowd work platforms on MTurk and Prolific [14,15]. We also had conducted crowd work-based research but still had some ethical concerns and doubts about crowd work platforms' "merits." However, we found little prior research about the ethical issues with data quality and validity in crowd work-based research from the scholars' and the IRB's perspectives. Hence, to fill this research gap, we conducted 32 semi-structured in-depth interviews with both scholars and IRB directors and analysts who were experienced in conducting and reviewing crowd work-based research to probe their ethical perceptions and practices.

We summarized the following themes from our research findings regarding data quality and validity in crowd work-based research. First, most scholars did not want to bother or waste their time arguing with crowd workers about data quality issues in their task submission. Some scholars in our interviews would approve crowd workers unconditionally even though they knew that some crowd workers were cheating or spamming. Only one scholar raised his concern about the consequence if every scholar avoided rejecting crowd workers. Second, two scholars from different research disciplines held contrasting views of data validity in crowd work-based research, particularly between internal and external validity. Third, the IRB

URI: https://hdl.handle.net/10125/79885
978-0-9981331-5-7

directors and analysts in our interviews usually held a suspicious if not critical attitude toward data quality and validity issues in crowd work-based research. Some of them expressed a resignation that crowd work-based research had been so popular and powerful in academia that their ethical concerns would not impact scholars' flocking to this trend. One IRB director even felt threatened when he helped the scholars in his institution to negotiate with some crowd workers about their data quality and rejection.

Based on these findings, we conducted a critical analysis. First, echoing one scholar's perspective during our interview, we argue that it becomes a "Tragedy of the Commons" for all the scholars in crowd work-based research if every scholar continues to avoid rejecting any crowd workers with cheating behaviors. Because consequently, no crowd worker needs to care about their data quality, and the ecosystem of crowd work-based research will collapse. Second, we argue that the validity of crowd work-based research should not be cross-referenced between different research disciplines or focused only on a crowd work platform such as MTurk. Instead, the validity of crowd work-based research should be assessed with a research discipline's particular characteristics or expectations. Finally, we argue that the IRB's ethical concerns in crowd work-based research should be voiced and respected. We felt that many IRB directors and analysts were in an embarrassing or even inferior position. Their ethical doubts or critiques of crowd work-based research had little influence on many scholars' favor of recruiting crowd workers. Also, they had to help scholars deal with "aggressive" complaints and menace from a few crowd workers, as one IRB director put it.

Our study contributes to the crowdsourcing and digital workforce in the gig economy minitrack as follows. First, both ethics and crowd work-based research are popular and prominent topics in academia but have not been cross-investigated extensively in HICSS and beyond. Our study paves a path into this merging and underdeveloped research area. Second, existing scholarship lacks an exploration from the IRB's perspectives about data and validity issues in crowd work-based research. Our study is a pioneering effort to uncover their perceptions and practices. Third, our study provided multiple comparative lenses into crowd work-based research ethics between scholars in different research disciplines and between scholars and IRBs. Finally, prior research on ethics in the gig economy focused primarily on the digital workforce's vulnerability. However, our study took an alternative view and identified a Tragedy of Commons for scholars doing crowd work-based research if they continue to refuse or avoid rejecting unaccountable crowd workers that provide poor quality data.

2. Related Work

2.1. Data quality issues in crowd work

Data quality issues have been in concern and dispute since the nascence of crowd work. Many scholars in various research disciplines have claimed that the data collected from crowd workers were of comparable and even superior quality to that from the other venues [10,12,13,16]. Nonetheless, some scholars also observed that crowd workers would deliberately cheat in their responses [14] or use virtual private servers (VPS) to conceal their real IP address to take tasks [17] and even use automatic scripts to answer surveys [18]. Due to these paradoxical findings, research is ongoing about what factors impact data quality in crowd work. Scholars have analyzed the correlation between the monetary incentive and data quality but could not reach a consensus. For example, Buhrmester et al. and Mason & Watts suggested that increasing the payment rate would lead to more contribution quantity but not necessarily better data quality from crowd workers [16,19]. Yet, Litman et al. and Yin et al. have found it otherwise, at least to some specific crowd worker populations and types of crowd work tasks [20,21].

Scholars also investigated the data quality variance related to crowd workers' reputations and crowd work fairness. For example, Peer et al. found that the worker's reputation is a sufficient condition for their data quality, and those with high approval rates tend to provide high-quality input [22]. More recently, however, Lovett et al. revealed that some MTurk workers with high approval rates confessed that they did not always provide reliable data [23]. In addition, Whiting et al. argued that it would be inherently inappropriate to associate crowd workers' reputation with their input quality because crowd workers are distributed and decentralized, who undercut accountable behaviors and institutions that are essential to work quality [24]. Thus, Whiting et al. proposed a worker-led reputation prototype to facilitate peer-review among crowd workers to produce quality work [24].

Moreover, scholars have attributed data quality issues to the unfair and unbalanced power dynamics between crowd workers and requesters. Specifically, they argued that the unfairness and exploitation on MTurk had discouraged crowd workers from participating and producing good work [25-27]. Some scholars proposed migrating MTurk to alternative crowd work platforms, most notably Prolific since

Prolific exhibits several sampling and fairness advantages over MTurk [7,8]. Nonetheless, a recent study has disclosed that cheating behaviors and data quality issues exist in Prolific as they do on MTurk [15]. To summarize, scholars still face ethical challenges to ensure data quality while protecting crowd workers' benefit in research. Still, a gap in the prior literature is that little do we know about how IRBs perceive data quality issues in crowd work-based research and how their perceptions compare with scholars' ethical perspectives and practices.

2.2. Research validity issues in crowd work

Apart from data quality, scholars also debate the validity of leveraging crowd work for academic purposes. Research validity can be further divided into internal validity and external validity. Internal validity means that certain conditions are shown to lead to other conditions, and their relationships are not due to spurious factors in a specific research context [28]. External validity means a study's findings can be generalized to a more extensive research context and population [28]. The scholarly debate over the validity of crowd work-based research is evident in an exemplary "academic fight." To start, Kees et al. published a paper to advocate for MTurk as a valid research venue to collect data besides the professional panels and student subject pools [13]. Then, Ford wrote a comment and critiqued Kees et al.'s advocacy, and pointed out that the "spammers and speeders" on MTurk would severely ruin the validity of research conducted on this platform [29]. Subsequently, Kees et al. published another paper rebuking Ford's critique again and argued that spammers and speeders were not "unique" to MTurk and could engender a similar level of validity problem in the other pools [30].

More specific into the concerns, some scholars were dubious about research validity on MTurk due to MTurk workers' prior knowledge or acquittance with frequently asked survey questions and experimental manipulations [31,32]. Additionally, some scholars were concerned that crowd workers were not as attentive to questions as the other research samples, such as college students, and crowd workers might not represent the segmented populations such as cognitive science [33,34]. Although a few scholars had compared and claimed that the MTurk population is representative and no more biased than the national public in the U.S. [35,36]. More recently, Hargittai and Shaw's empirical and comparative study suggested that it would be wiser not to claim any generalizability based on the data and findings on MTurk [37].

Despite these prior works about research validity issues in crowd work, there are still lacunae in the literature. First, there lacks a comparative inquiry with scholars in different research disciplines to probe into their opinions about the validity of crowd work-based research. We wondered whether scholars in different research disciplines would have varied perceptions and expectations of validity in crowd work-based research. Second, prior scholarship largely ignored the IRBs' perspectives in crowd work-based research regarding ethical issues with data quality and validity. We posited it to be a critical gap because the IRB has an essential role in advising and supervising academic research design and ethics. Last but not least, as crowd work-based research becomes prevalent across many disciplines, associated ethical issues are pressing to be explored and discussed to hold scholars and IRBs conscious and accountable. Thus, we are motivated and feel obliged to conduct this explorative study to draw more discussion among scholars and IRBs.

3. Method

We conducted 32 semi-structured interviews with 15 scholars and 17 IRB directors and analysts in the U.S. to probe their perceptions and practices about the ethical issues in crowd work-based research. We devised our interview questions from three sources. First, since crowd work-based research involves human subject participation, we borrowed several typical questions in an IRB application template, such as what are the most common risks and benefits in crowd work-based research. Second, we designed questions deriving from the prior literature about crowd work-based research, such as how to evaluate exploitation and compensation in this context. Finally, we added a few new questions after novel themes emerged during our interview progress, such as how to perceive using a minimum wage standard as the compensation benchmark in crowd work-based research. The interviews lasted between 45-60 mins and were conducted through Skype or phone call. All the interviews were recorded and fully transcribed.

In terms of our samples, the 15 scholars we interviewed were from a pool of active and frequently cited scholars who publish crowd work-based research in top journals and conferences. The 17 IRB directors and analysts we interviewed were from a list of all the IRB directors and analysts in the top 100 universities in the U.S. [38]. Noteworthy that on both camps, our recruitment turned out to be challenging and time-consuming. Both our interviewed scholars and IRB respondents were very concerned about protecting their privacy and their institution's privacy. Hence, we cannot reveal their detailed demographic information. Table 1 portrays the categorization and distribution of our participants.

Table 1. The basic information of the participants

Role	Institution/Discipline	Number
IRB	Public University (R1)	8
IRB	Private University (R1)	6
IRB	Private University (R2)	2
IRB	College (R1)	1
Scholar	Information/Computer Science	4
Scholar	Psychology	3
Scholar	Engineering	2
Scholar	Business	2
Scholar	Communication	2
Scholar	Political Science	2

Due to the recruiting and scheduling difficulties, our interview data collection spanned several months. We kept our first round of data analysts, including note-taking and transcription, simultaneous with data collection, which is a recommended strategy for qualitative research [39]. We continued data collection until we sensed that the themes had been theoretically saturated. Then, after all our interviews had been fully transcribed, we conducted a thematic analysis on our transcriptions with Atlas.ti following Braun and Clarke's suggested steps [40]. First, we imported the interview transcriptions and familiarized ourselves with them on Atlas.ti. Second, we coded through all the transcriptions and compared them with our notes during the interviews. Third, we identified the codes and the emerging themes from the codes. Fourth, we re-examined our codes and the themes for their validity and fit with each other. Fifth, we refined and described each theme. Finally, we reviewed all the codes and themes and generated a narrative to present our findings. Our research design, sampling plan, data protection strategies had been approved by IRB.

4. Findings

We identified numerous themes about ethical issues in crowd work-based research. In the scope of this paper, we focus on reporting and reflecting on the ethical issues about data quality and validity.

4.1. Ethical deliberation on data quality

Data quality is a central ethical consideration among our interviewed scholars and IRB directors/analysts. First of all, some respondents worried about data quality decline due to crowd workers scribbling through questions or giving their answers without thought. For example, one scholar said:

> *There is a minority, actually not a minority, that can do lots of damage because these people are not paying attention to get work done but to maximize their payoff for the minimum amount of work. (P28)*

P28 immediately corrected his first thought of only a few spamming crowd workers and speculated the spammers' motive to maximize compensation with the least amount of effort. However, unlike P28, most scholars in our interviews posited that the percentage of spammers in crowd work is small, even though there has been little empirical research or estimation on the approximate percentage of spammers within the crowd workers population.

On the IRB side, many respondents shared with the scholars a critical view of spammers among crowd workers. Some of them further cast doubt on the reliability of levering crowd work for academic research. One IRB director said:

> *I am not in favor of using MTurk, and I don't know whether the data are reliable. You have people there doing it for the money…so are they answering the questions, or are they understanding the questions? (P13)*

P13 has served in public and private universities, and she explicitly told us that she would not recommend the researchers in her universities to use MTurk. Her primary concern, like P28, was that some crowd workers purported to maximize earnings as fast as they can and would care little about data quality

Besides spamming, our respondents worried about fraud in crowd work-based research, which means that some crowd workers would deceive their qualifications and eligibility to participate in tasks with specific screening criteria. One scholar told us:

> *We have done interviews with MTurk workers in India, and a lot of times, they would take on those tasks where it was completely about U.S. politics. What they would do is that they would imagine that they were like a soccer mom in the U.S. and think about what she would respond. (P29)*

P29 had interviewed Indian crowd workers' behaviors and habits previously and knew that some of them would pretend to be in the U.S. and answer questions about U.S. politics. Hence, she did not trust survey studies on MTurk since the data quality is susceptible to crowd workers' fraud. Her voice echoed Antin and Shaw's research findings on the U.S. and Indian MTurk workers' self-reporting problems, which also revealed social desirability biases [41].

Additionally, two respondents raised concerns about crowd workers' fraud using scripts or VPS to automate their answers. These respondents mentioned the "bot crisis" on MTurk, which has also been reported as a notable fraud case by Dennis et al. and Chmielewski and Kucker recently [17], [42]. One scholar, P20, having highlighted the bot crisis, proposed to screen MTurk workers with a high approval rate to avoid such data quality crisis: *"you have to set your approval rate above 95% to avoid the bots or something like this." (P30)*

An approval rate is a crowd worker's ratio of completed tasks that have been approved by previous requesters. Screening crowd workers based on their approval rate is a common approach to fend data quality. Most of our interviewed scholars favored a high rate par like 95% to filter crowd workers' eligibility. However, not every scholar agreed with setting a fixed high approval rate to screen crowd workers. Some considered the popular yet rigid 95% approval rate was arbitrary and without sufficient empirical test of its validity. For example, one scholar said he understood the rationale of applying a 95% approval rate, but he was against it ethically:

> *Well, I don't think that the psychology pool has good criteria [for screening]; it's just "Are you at the university and are you taking a class, and if so, we are thinking that you are appropriate and we can generalize to anyone based on what you have done." We see so much data from 18-20-year-old without working experience; is that useful? I don't know, and I am concerned about it. (P22)*

He implied that psychology research had no screening criterion as an approval rate for students, but he heard little criticism about it. By contrast, researchers started to set up a specific screening rate for crowd workers, so he perceived it to be a double standard.

Meanwhile, one scholar abandoned using a high approval rate and would intentionally apply a "zero" or near-zero approval rate to target novice crowd workers. He perceived it to be a better strategy for data quality:

> *This is my own little secret. I like new workers that have completed almost no tasks [be]cause they're going to try hard. They're just like "I better do a good job." And they're clicking all the stuff, and they pay attention at the end. They're grateful whereas somebody who's taken a thousand surveys there, and they're more likely to know what I'm assessing and know what I'm trying to test for. They're more used to all the attention checks, whereas a beginner is more likely to try hard. (P27)*

P27 assumed that new crowd workers would try harder and be more attentive than seasoned ones because they wanted to accrue a good reputation and were still unfamiliar with various survey questions and experimental treatments. Thus, for him, new crowd workers would produce better data quality than seasoned ones. However, several IRB respondents pointed out the pitfall in recruiting novice crowd workers. They worried that "trying too hard" in crowd work beginners is a sign of social desirability bias and may divert from a researcher's original recruitment plan and expected sample characteristics.

Apart from using an approval rate, numerous scholars proposed inserting attention check questions (ACQs) to ensure data quality. They assumed that if crowd workers neglected or failed the ACQs, their data quality in the other questions would also be problematic. Most of our IRB respondents also supported using ACQs as judgment or evidence to reject crowd workers or respond to their complaints.

However, to reject or not to reject a crowd worker is a difficult question. Scholars in our interviews struggled with it. On the one side, they were conscious that spamming and fraud would damage their research quality and waste their funding. On the other side, they did not regard it worthy of their time and stress to negotiate with crowd workers over rejection. Hence, they would approve all the crowd workers regardless of their data quality. One scholar explained:

> *There were some cases when it was obvious that some workers did not use our plugin, and we had ways of tracking whether they did use it, and they basically wanted to get paid even though they did not use it. So, we sent them an email and said, ok, you didn't use it, so we were not going to pay you. Some of these workers mailed us again, and we did end up paying them, but it was more to just remove the hassle and not wanting to fight. (P29)*

P29's choice of avoiding hassles with crowd workers was prevalent among our interviewed scholars. Such a choice could be apprehensible when an IRB director described how aggressive a crowd worker's complaint of rejection could be:

> *I would say 75% of the complaints were aggressive. One that I received last week, he missed an attention check, and he said, "I did," and the investigator said that "well, I can check it for you, but because it's de-identified, I will need a portion of your IP address to look it up." The investigator was just asking for the first few digits of the IP address, and this person got very upset and said, "I did not know that XXX University was sponsoring this kind of survey to obtain data for free." (P11)*

From P11's recount, even if a researcher had concrete evidence of a crowd worker's poor data input and wanted to cooperate to solve the issue, a crowd worker's complaint could still be antagonistic and threatening.

As an exception, one scholar, P20, disagreed with the other scholars' choice of approving every crowd work regardless of their data quality. Instead, he critiqued the practice of unselective approval and unconditional payment to crowd workers irrespective of their data quality. P20 explained:

> *Part of the reason that the HIT approval ratio no longer works so well for signaling poor quality workers is that many researchers just approve everybody, whether they are mandated to by their IRB or because they don't want to deal with this hassle at all. It's kind of a "Tragedy of the Commons" where everybody, for their self-interest, makes the decision for just paying people and leaving them alone. Then collectively, it undermines the reputational mechanism that makes MTurk work. (P24)*

P24 posited that academic researchers collectively and interdependently rely on a reputation system, i.e., the approval rate system, to filter qualified crowd workers from unqualified ones. If every researcher decides to approve all crowd workers, a crowd worker's approval rate will always be (fake) high, even if they are spamming or deceiving. Consequently, approval rates become inflated and meaningless to signal crowd workers' accountability. P24 referred to this consequence as the "Tragedy of the Commons:" a crowd worker's approval rate can no longer deter their spamming or fraud, and all the researchers would waste money on bad data and have to increase payment to solicit high-quality data.

4.2. Ethical deliberation on research validity

Besides data quality, our interviewees also deliberated the validity issues in crowd work-based research. Regarding internal validity, our respondents expressed opinions around three themes: crowd workers' non-naïvety, information diffusion, and the undue influence of payment. First, some respondents were concerned about crowd workers' non-naïvety due to their prior knowledge or predispositions to various survey questions and experimental treatments. For example, one scholar said that researchers should never run a "Prisoner's Dilemma" on MTurk again because so many MTurk workers had taken it:

> *I know from social science work on MTurk that there's almost like a running joke. Like you can't run a prisoner's dilemma experiments on mechanical Turk because so many people have done that on MTurk. The results will be skewed from what the natural population would do. (P21)*

Prisoner's Dilemma (PD) is a classic game-theory problem about cooperation and defect. Horton et al. ran one of the first PD experiments on MTurk [43]. Most recently, Capraro et al. conducted a Stag-Hunt Game (SHG) experiment on MTurk, which derives from PD [44]. In this sense, PD related experiments have been tested on MTurk for over a decade.

The majority of our respondents agreed that crowd workers' non-naïvety could spoil survey validity and experimental manipulations. For instance, one scholar explained how crowd workers' non-naïvety could damage the internal validity of political research:

> *Another thing on MTurk, which is to some degree why I stop using MTurk, is [that] people may risk paying too much attention because they know that you are evaluating their work. That doesn't serve the kind of work that I do in political science, whereas in the real world, people are not paying close attention to politics, so it's not a very accurate measure of what we are capturing. (P23)*

P23 noted that MTurk workers might pay too much attention to his questions about politics because they cared about their responses that researchers would evaluate for payment. However, such an attentive and meticulous attitude and behavior toward political questions did not resemble people's attitudes and reactions to politics in the real world, which are more indifferent and casual. Hence, such non-naïvety was not helpful for his political research and partly yielded him to stop using MTurk.

However, one scholar, P18, offered an opposing view to the other respondents. He contended that crowd workers' non-naïvety could, in effect, benefit his privacy research:

> *I and the others have been arguing for a while, that it's totally OK to keep on using MTurk even in privacy studies, so even if a participant made a lie in a privacy study, because in my experience, the results are almost always more conservative [with MTurk workers] than the results that I get from the other samples. In other words, more conservative means, because the subjects may be a little less naïve, if I try to use a certain treatment, an experimental treatment, it's harder, harder [emphasis his] to produce a statistically significant effect on MTurk samples than on non-MTurk samples. (P18)*

P18 argued that precisely because many MTurk workers were prone to be less non-naïve and more accustomed to various surveys and experimental treatments than the general public, it became harder to experiment with robust manipulations and obtain a statistically significant effect. Therefore, MTurk workers' non-naïvety is beneficial for preventing Type I error when a researcher may falsely claim a significant effect in their treatments.

Apart from non-naïvety, our respondents were also concerned about information diffusion that threatens internal validity. They noted that crowd workers might share research information on online forums outside a crowd work platform, as one IRB director critiqued:

> *You must make sure that they are not talking to one another. You must admit that limitation as well. We all know that mechanical Turk workers have their own groups that they are on all day long as they're working, and they're telling each other about the surveys that they're taking. And that is a huge threat to internal validity. (P12)*

P12 posited that MTurk workers were diffusing research information with each other, which made some MTurk workers equipped with predisposed opinions or prior knowledge before participating in a research project. Therefore, their predisposition and prior knowledge became spurious factors.

Finally, a few respondents were concerned about the undue influence of payment that could impair internal validity. For example, P31, a psychologist, explained the payment effect on internal validity:

> *I do believe it creates a power dynamic where "I better say what this requester thinks or wants in order to get my payment. I get to do what I think the requester is asking for in order to just get paid" whereas if you are just volunteering for research, you don't have that same power dynamic. (P31)*

P31 perceived that some crowd workers would intentionally cater their responses to a requester's expectations rather than out of their real thoughts to get the requester's approval and compensation. Hence, these crowd workers' social desirability bias due to the monetary incentive would impair their research's internal validity.

In terms of external validity, multiple IRB directors and scholars shared their concerns about crowd workers' representativeness of specific populations. For example, one IRB director highlighted the disparity between the crowd work population and the U.S. population regarding political research:

> *If you are a political scientist, for example, the MTurk worker population is not a representative sample of American voters. I think MTurk workers are quite bimodal in age distribution. That's like all people are young or old, and there is not too much in between. They are better educated than average Americans; they are probably a little bit wealthier than average Americans. I don't think studies done on MTurk can be presented as a representative sample of Americans. (P17)*

Meanwhile, some also expressed concerns about the generalizability of the data from crowd workers, which is not only related to crowd workers' demographics but also to the sampling challenges in crowd work-based research.

Specifically, a few respondents posited that payment had a paradoxical effect on the sampling of crowd workers. They remarked that, first, an insufficient monetary incentive could discourage some crowd workers from participating, and this exclusion could render the research sample unrepresentative. Second, they speculated that if the payment was low or devoid, yet some crowd workers still chose to participate regardless, then these participants might not constitute the right sample either. For example, one scholar explained:

> *There is a trade-off, and imagine we did not pay anybody, then the only people that would participate would be the people who already love taking political surveys, and that is not the group we want to know more about. We want to know more about people who may be not interested in politics. So, I think if we want to be more representative, we are still going to be in the position to compensate people. (P23)*

Besides, another scholar raised his concern about the statistical power in crowd work-based research. He regarded it hard to get a large sample for a specific population on MTurk, and as such, it became a more severe problem than a sample's characteristics:

> *It's hard to get enough [statistical] power on MTurk study to do something really big like among "civic Republicans" because there are fewer of them. If I have something that I want to run among Republicans and I want it to break it up, Republicans old and young, and if I only have 200 Republicans in my sample [on MTurk], then it's basically impossible to do that analysis. (P20)*

P20 noted that if he needs to divide a sample into subgroups with more refined demographics, it would be impossible to obtain sufficient statistical power because those subgroups among the crowd workers are too small.

Finally, one scholar stated that it is impossible to maintain both external validity and internal validity at the same time, and he would compromise the former for the latter in crowd work-based research:

> *[F]or me, the most important goal is internal validity, which means that as I usually don't say, "our results demonstrate that every human being will act the following way." Rather, we say, "with this kind of sample size, and the sample of the population we use, we found these results, and because we believe that we did not have any confound, we stand by our result." And then it's the broader goal and mission of science to be able to replicate, to generalize, or to invalidate the results that I obtain, or the others obtain, using different samples. (P18)*

P18 further told us that he would constrain his sample of crowd workers in the U.S. for better data quality and internal validity because it fit his research purpose and discipline better, even though it would sacrifice his research generalizability.

5. Discussion

5.1. Data quality: "Tragedy of the Commons"

First, we agreed with P20's concern that if every scholar avoided rejecting any crowd workers regardless of their data quality, then the data quality in crowd work-based research would worsen, which will impair all scholars' research and funding. The "Tragedy of the Commons" is a phenomenon described and discussed by economist Garrett Hardin. It means that if every individual relied on themselves and pursued their self-interests without concerning their relationships with the other people or the others' interests in a community, then the common resources would deplete and harm everyone in the end [45].

To some degree, Hardin's depicted phenomenon resembles the situation of crowd work-based research. Take MTurk as an example. On the one side, the "common resources" for all the scholars are the population of MTurk workers and their data input, and these common resources are limited. Prior research has estimated that the active population of MTurk workers for average laboratory sampling was only about 7,300 [46]. Even though such a number may be larger nowadays, the MTurk population for scholars to sample is arguably still quite limited. On the other hand, many MTurk workers have to do multiple tasks simultaneously to maximize their income [47]. As such, MTurk workers' attention span to an individual task and data quality is usually short and sometimes careless [47,48]. In addition, scholars not only need to compete with each other to sample from a limited population of MTurk workers and win their attention but also must compete with non-academic requesters. On the other side, the "self-interest" for all the scholars is to gather good data from qualified MTurk workers.

However, as previous research and our findings indicated, not all the MTurk workers were accountable or qualified to provide good quality data. Thus, it becomes a choice for scholars on how to pursue their "self-interest" in crowd work-based research. Our study found that most scholars would use a specific approval rate as a screening mechanism; nonetheless, they would not bother negotiating with MTurk workers and prefer to accept all the submissions regardless of data quality. Some scholars claimed that they did not want to waste time "fighting" with angry MTurk workers; some were compassionate with the meager income MTurk workers could obtain. But regardless of their rationales, we argue that if every scholar continues to pursue their self-interest as such, the limited "common resources" in crowd work-based research will deteriorate. Because as a result, all crowd workers would realize that they can be free riders in taking academic tasks on MTurk and get paid without rejection. Moreover, suppose scholars continue to reject no crowd worker; in that case, every crowd worker will likely have a fake high approval rate, which will be useless for scholars to filter out "qualified" crowd workers.

5.2. Validity: ignored disciplinary differences

As regards research validity, we found an interesting theme about disciplinary differences. P18, a computer scientist, argued that crowd workers' non-naïvety could benefit his research because it would make his experimental tests harder to be significant than with "naïve" participants. Also, unlike political scientists such as P20 and P23, P18 weighed more on the internal validity over the external validity in crowd work-based research. Hence, he would compromise crowd work-based research's generalizability, a primary concern for P20 and P23, to whether his sample of crowd workers has few confounding factors and can ensure internal validity. Albeit a minority in our respondents, P18's viewpoint implies that scholars in political science and computer science may have distinctive expectations and preferences of research validity in their respective research disciplines.

To our knowledge, prior studies lack sufficient attention and discussion about research validity issues related to disciplinary differences in crowd work-based research. Scholars seemed to presume that the validity of crowd work-based research in one discipline can hold and extrapolate to another discipline. For instance, we noticed that several computer scholars were citing a few political scientists' early works about MTurk's validity to support their research methodology (e.g., [49,50]). P18 and P23's opposing views on the validity issues of crowd work-based research reminded us that any blind reliance and extrapolation on the findings of validity in one research discipline to another could be biased or at least questionable. Hence, we propose that the future discussion about validity in crowd work-based research should not merely focus on a specific crowd work platform (e.g., whether MTurk is valid for scientific research generally). Instead, the discussion should orient at individual research disciplines (e.g., whether it is valid to conduct political research via crowd work).

Finally, we hope to help voice IRB's ethical concerns in this study. Although we knew that IRB usually categorizes crowd work-based research as exempt with minimum risks, we realized that it does not necessarily mean that they favor utilizing crowd work for academic purposes. As our findings revealed, many IRB directors and analysts expressed their deep ethical concerns about data quality and validity issues in crowd work-based research. We advocate that these concerns should be heard by more scholars rather than remain in silence or within an individual institution.

6. Conclusion

This study investigated the data quality and validity issues in crowd work-based research from the scholars' and IRB's ethical perspectives. We found common themes about data quality issues such as spamming, fraud, and validity issues such as non-naivety and information diffusion. More important, we identified and discussed two ethical issues in crowd work-based research, i.e., a potential "Tragedy of the Commons" due to scholars' data quality control and negligence of research validity preference according to different disciplines. Due to the limited sample size, our findings may not represent the diversity of the scholars' and IRBs' views. We plan to conduct a large-scale survey to gather more input from scholars in crowd work-based research and IRB directors and analysts in more institutions for our future work.

7. References

[1] J. Howe, "The Rise of Crowdsourcing," *Wired* no. 14, p. 5, 2006.

[2] A. Kittur *et al.*, "The Future of Crowd Work" presented at the CSCW '13, San Antonio, Texas, USA, 2013.

[3] K. B. Sheehan, "Crowdsourcing research: Data collection with Amazon's Mechanical Turk," *Commun. Monogr.*, vol. 85, no. 1, pp. 140–156, Jan. 2018, doi: 10.1080/03637751.2017.1342043.

[4] M. A. Graber and A. Graber, "Internet-based crowdsourcing and research ethics: the case for IRB review," *J. Med. Ethics*, vol. 39, no. 2, pp. 115–118, Feb. 2013, doi: 10.1136/medethics-2012-100798.

[5] V. Williamson, "On the Ethics of Crowdsourced Research," *PS Polit. Sci. Polit.*, vol. 49, no. 01, pp. 77–81, Jan. 2016.

[6] M. S. Silberman, B. Tomlinson, R. LaPlante, J. Ross, L. Irani, and A. Zaldivar, "Responsible research with crowds: pay crowdworkers at least minimum wage," *Commun. ACM*, vol. 61, no. 3, pp. 39–41, Feb. 2018, doi: 10.1145/3180492.

[7] Prolific, "Prolific vs MTurk," *Why Prolific?* https://www.prolific.co/prolific-vs-mturk/

[8] S. Palan and C. Schitter, "Prolific.ac—A subject pool for online experiments," *J. Behav. Exp. Finance*, vol. 17, pp. 22–27, Mar. 2018, doi: 10.1016/j.jbef.2017.12.004.

[9] "Amazon Mechanical Turk." https://www.mturk.com/ (accessed Sep. 23, 2020).

[10] C. Callison-Burch, "Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 1 - EMNLP '09*, Singapore, 2009, vol. 1, p. 286. doi: 10.3115/1699510.1699548.

[11] M. C. Haug, "Fast, Cheap, and Unethical? The Interplay of Morality and Methodology in Crowdsourced Survey Research," *Rev. Philos. Psychol.*, vol. 9, no. 2, pp. 363–379, Jun. 2018, doi: 10.1007/s13164-017-0374-z.

[12] G. Paolacci, "Running experiments on Amazon Mechanical Turk," *Judgm. Decis. Mak.*, vol. 5, no. 5, p. 9, 2010.

[13] J. Kees, C. Berry, S. Burton, and K. Sheehan, "An Analysis of Data Quality: Professional Panels, Student Subject Pools, and Amazon's Mechanical Turk," *J. Advert.*, vol. 46, no. 1, pp. 141–155, Jan. 2017, doi: 10.1080/00913367.2016.1269304.

[14] D. E. Difallah, G. Demartini, and P. Cudré-Mauroux, "Mechanical Cheat: Spamming Schemes and Adversarial Techniques on Crowdsourcing Platforms," CrowdSearch 2012 workshop at WWW 2012, Lyon, France

[15] C. Schild, L. Lilleholt, and I. Zettler, "Behavior in cheating paradigms is linked to overall approval rates of crowdworkers," *J. Behav. Decis. Mak.*, vol. 34, no. 2, pp. 157–166, Apr. 2021, doi: 10.1002/bdm.2195.

[16] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?," *Perspect. Psychol. Sci.*, vol. 6, no. 1, pp. 3–5, 2011.

[17] S. A. Dennis, B. M. Goodson, and C. Pearson, "Mturk Workers' Use of Low-Cost 'Virtual Private Servers' to Circumvent Screening Methods: A Research Note," *SSRN Electron. J.*, 2018, doi: 10.2139/ssrn.3233954.

[18] E. Dreyfuss, "A Bot Panic Hits Amazon's Mechanical Turk," *WIRED*, Aug. 17, 2018. https://www.wired.com/story/amazon-mechanical-turk-bot-panic/

[19] W. Mason and D. J. Watts, "Financial Incentives and the 'Performance of Crowds,'" in *KDD-HCOMP '09*, Paris, France, 2009, p. 9.

[20] L. Litman, J. Robinson, and C. Rosenzweig, "The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk," *Behav. Res. Methods*, vol. 47, no. 2, pp. 519–528, Jun. 2015, doi: 10.3758/s13428-014-0483-x.

[21] M. Yin, Y. Chen, and Y.-A. Sun, "Monetary Interventions in Crowdsourcing Task Switching," in *Proceedings of the Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

[22] E. Peer, J. Vosgerau, and A. Acquisti, "Reputation as a sufficient condition for data quality on Amazon Mechanical Turk," *Behav. Res. Methods*, vol. 46,

no. 4, pp. 1023–1031, Dec. 2014, doi: 10.3758/s13428-013-0434-y.

[23] C. A. Anderson, J. J. Allen, C. Plante, A. Quigley-McBride, A. Lovett, and J. N. Rokkum, "The MTurkification of Social and Personality Psychology," *Pers. Soc. Psychol. Bull.*, vol. 45, no. 6, pp. 842–850, Jun. 2019.

[24] M. E. Whiting *et al.*, "Crowd Guilds: Worker-led Reputation and Feedback on Crowdsourcing Platforms," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, Portland Oregon USA, Feb. 2017, pp. 1902–1913.

[25] L. C. Irani and M. S. Silberman, "Turkopticon: interrupting worker invisibility in amazon mechanical turk," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, Paris, France, 2013, p. 611.

[26] N. Salehi *et al.*, "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 - CHI '15*, Seoul, Republic of Korea, 2015, pp. 1621–1630. doi: 10.1145/2702123.2702508.

[27] M. Graham *et al.*, "The Fairwork Foundation: Strategies for Improving Platform Work," *Weizenbaum Conf.*, 2019, doi: 10.34669/WI.CP/2.13.

[28] L. H. Kidder and C. M. Judd, *Research methods in social science*. New York: CBS College Publishing, 1986.

[29] J. B. Ford, "Amazon's Mechanical Turk: a comment," *J. Advert.*, vol. 46, no. 1, pp. 156–158, 2017.

[30] J. Kees, C. Berry, S. Burton, and K. Sheehan, "Reply to 'Amazon's Mechanical Turk: A Comment,'" *J. Advert.*, vol. 46, no. 1, pp. 159–162, 2017.

[31] G. Paolacci and J. Chandler, "Inside the Turk: Understanding Mechanical Turk as a Participant Pool," *Curr. Dir. Psychol. Sci.*, vol. 23, no. 3, pp. 184–188, Jun. 2014, doi: 10.1177/0963721414531598.

[32] J. Oppenlaender, A. Visuri, K. Milland, P. Ipeirotis, and S. Hosio, "What do crowd workers think about creative work?," Feb. 2020, [Online]. Available: http://arxiv.org/abs/2002.10887

[33] J. K. Goodman, C. E. Cryder, and A. Cheema, "Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples," *J. Behav. Decis. Mak.*, vol. 26, no. 3, pp. 213–224, 2013.

[34] N. Stewart, J. Chandler, and G. Paolacci, "Crowdsourcing Samples in Cognitive Science," *Trends Cogn. Sci.*, vol. 21, no. 10, pp. 736–748, Oct. 2017, doi: 10.1016/j.tics.2017.06.007.

[35] A. J. Berinsky, G. A. Huber, and G. S. Lenz, "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk," *Polit. Anal.*, vol. 20, no. 3, pp. 351–368, 2012.

[36] S. Clifford, R. M. Jewell, and P. D. Waggoner, "Are samples drawn from Mechanical Turk valid for research on political ideology?," *Res. Polit.*, no. October-December, pp. 1–9, 2015.

[37] E. Hargittai and A. Shaw, "Comparing Internet Experiences and Prosociality in Amazon Mechanical Turk and Population-Based Survey Samples," *Socius Sociol. Res. Dyn. World*, vol. 6, Jan. 2020, doi: 10.1177/2378023119889834.

[38] "2021 Best National University Rankings," *U.S.News*. [Online]. Available: https://www.usnews.com/best-colleges/rankings/national-universities

[39] C. Marshall and R. Gretchen B, *Designing qualitative research*. SAGE Publications, Inc, 2014.

[40] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qual. Res. Psychol.*, vol. 3, no. 2, pp. 77–101, Jan. 2006, doi: 10.1191/1478088706qp063oa.

[41] J. Antin and A. Shaw, "Social Desirability Bias and Self-Reports of Motivation: A Study of Amazon Mechanical Turk in the US and India," Austin, TX, USA, May 2012, p. 10.

[42] M. Chmielewski and S. C. Kucker, "An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results," *Soc. Psychol. Personal. Sci.*, vol. 11, no. 4, pp. 464–473, May 2020, doi: 10.1177/1948550619875149.

[43] J. J. Horton, D. G. Rand, and R. J. Zeckhauser, "The online laboratory: conducting experiments in a real labor market," *Exp. Econ.*, vol. 14, no. 3, pp. 399–425, Sep. 2011, doi: 10.1007/s10683-011-9273-9.

[44] V. Capraro, I. Rodriguez-Lara, and M. J. Ruiz-Martos, "Preferences for efficiency, rather than preferences for morality, drive cooperation in the one-shot Stag-Hunt Game," *Journal of Behavioral and Experimental Economics*, 86, 101535.

[45] G. Hardin, "The Tragedy of the Commons," *Science*, vol. 162, no. 3859, pp. 1243–1248, 1968.

[46] N. Stewart *et al.*, "The Average Laboratory Samples a Population of 7,300 Amazon Mechanical Turk Workers," *Judgm. Decis. Mak.*, p. 15, 2015.

[47] T. Kaplan, S. Saito, K. Hara, and J. P. Bigham, "Striving to Earn More: A Survey of Work Strategies and Tool Use Among Crowd Workers," 2018, pp. 70–78.

[48] M. S. Aruguete, H. Huynh, B. L. Browne, B. Jurs, E. Flint, and L. E. McCutcheon, "How serious is the 'carelessness' problem on Mechanical Turk?," *Int. J. Soc. Res. Methodol.*, vol. 22, no. 5, pp. 441–449, Sep. 2019, doi: 10.1080/13645579.2018.1563966.

[49] B. Ur, J. Bees, S. M. Segreti, L. Bauer, N. Christin, and L. F. Cranor, "Do Users' Perceptions of Password Security Match Reality?," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, San Jose California USA, May 2016, pp. 3748–3760.

[50] N. Fitz, K. Kushlev, R. Jagannathan, T. Lewis, D. Paliwal, and D. Ariely, "Batching smartphone notifications can improve well-being," *Comput. Hum. Behav.*, vol. 101, pp. 84–94, Dec. 2019, doi: 10.1016/j.chb.2019.07.016.