Modus Operandi of Crowd Workers: The Invisible Role of Microtask Work Environments

UJWAL GADIRAJU, Leibniz Universität Hannover
ALESSANDRO CHECCO, University of Sheffield
NEHA GUPTA, University of Nottingham
GIANLUCA DEMARTINI, University of Sheffield

The ubiquity of the Internet and the widespread proliferation of electronic devices has resulted in flourishing microtask crowdsourcing marketplaces, such as Amazon MTurk. An aspect that has remained largely invisible in microtask crowdsourcing is that of *work environments*; defined as the hardware and software affordances at the disposal of crowd workers which are used to complete microtasks on crowdsourcing platforms. In this paper, we reveal the significant role of work environments in the shaping of crowd work. First, through a pilot study surveying the good and bad experiences workers had with UI elements in crowd work, we revealed the typical issues workers face. Based on these findings, we then deployed over 100 distinct microtasks on CrowdFlower, addressing workers in India and USA in two identical batches. These tasks emulate the good and bad UI element designs that characterize crowdsourcing microtasks. We recorded hardware specifics such as CPU speed and device type, apart from software specifics including the browsers used to complete tasks, operating systems on the device, and other properties that define the work environments of crowd workers. Our findings indicate that crowd workers are embedded in a variety of work environments which influence the quality of work produced. To confirm and validate our data-driven findings we then carried out semi-structured interviews with a sample of Indian and American crowd workers from this platform. Depending on the design of UI elements in microtasks, we found that some work environments support crowd workers more than others. Based on our overall findings resulting from all the three studies, we introduce *ModOp*, a tool that helps to design crowdsourcing microtasks that are suitable for diverse crowd work environments. We empirically show that the use of *ModOp* results in reducing the cognitive load of workers, thereby improving their user experience without effecting the accuracy or task completion time.

CCS Concepts: •**Information systems** → **World Wide Web; Crowdsourcing;**

Additional Key Words and Phrases: Crowdsourcing, Microtasks, User Interface, Work Environment, Design, Performance, Crowd Workers, Human Factors

ACM Reference format:
Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. 2017. Modus Operandi of Crowd Workers: The Invisible Role of Microtask Work Environments. *PACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 1, Article 1 (February 2017), 27 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

Author's addresses: U. Gadiraju, L3S Research Center, Leibniz Universität Hannover, Germany; A. Checco and G. Demartini, Information School, University of Sheffield, United Kingdom; N. Gupta, University of Nottingham, Nottingham, United Kingdom.

DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

"Sometimes the Internet fee is greater than the rewards I earn (due) to images, audios or videos in tasks."

– CrowdFlower Worker from India

We are currently in an age of pervasive computing where various kinds of sensors facilitate *smart* environments at home or work, improving our lives in numerous ways [76]; ranging from optimizing energy consumption [52] to facilitating structural health monitoring [46]. Recent work has showcased how visual sensors (in the form of CCTV cameras) and social sensors (such as Twitter feeds) can be combined to improve event detection and aid in understanding the evolution of situations [70]. The rapid growth and ubiquity of mobile devices has resulted in making participatory sensing feasible on a large scale [7]. Such opportunities of using people as sources of sensory information allows us to build useful applications that have implications on urban mobility [73], environment [38], personal health monitoring [58], and so forth. An effective way to use people as sources of sensory information is to collect data from them directly [43]. Online marketplaces like Amazon's Mechanical Turk[1] (AMT) or CrowdFlower[2] provide a large and diverse workforce of people accessible around the clock and on demand, for participation in return for monetary rewards.

Microtask crowdsourcing is being used widely these days to solve a multitude of problems that go beyond the capability of machine intelligence. Over time, crowdsourcing platforms like AMT have been used for diverse purposes including content creation and running surveys [12]. The potential to reach thousands of people around the globe at will and on demand [33] has led to innovative and unprecedented solutions to problems in the space of ubiquitous and pervasive computing [32, 51, 61]. On the other hand, the ubiquity of the Internet and rise in prevalence of electronic devices have led to the rise of applications centered on mobile crowdsourcing [15, 18, 24, 54, 74]. Recent work by Laput et al. that introduced *Zensors*, leverages real-time human intelligence from online crowd workers to create robust, adaptive and intelligent sensors for any visually observable property [44]. This is a great example of how human input and intelligence can play a pivotal role in pervasive computing.

An important aspect at the core of crowd-powered solutions (especially those based on the microtask crowdsourcing paradigm) is controlling the quality of work that is produced by workers. Several research works have focused on improving the quality of crowdsourcing results by using a variety of techniques ranging from worker pre-screening methods and effective crowdsourcing task design [42], to using gamification and incentive mechanisms [19, 59], and answer aggregation methods [67]. Due to the low entry barrier, crowdsourcing has become truly ubiquitous [68]. Prior ethnographic works have also shown that workers who participate in microtask crowdsourcing are embedded in diverse environmental contexts that impact their work routines [26, 48]. A considerable amount of research effort has focused on the *motivations* behind worker participation in crowdsourced microtasks [8, 39]. Yet, little is understood about the *tools* that support and drive such worker participation. While some previous works have focused on the question of *why* crowd workers spend time and complete tasks on crowdsourcing platforms, in this work we focus on the question of *how* crowd workers complete tasks, and investigate the affect of different work environments on worker performance.

In this paper we draw attention to the less addressed realm of the *modus operandi* of workers in crowdsourcing microtasks. How exactly do workers participate and contribute to crowd work? We are particularly interested in the work environments that influence the observable behavior of crowd workers. We aim to understand how the different *work environments*, defined and characterized by the hardware and software that enable worker participation in crowd work, affect the quality of work that is produced. With the rampant increase in mobile device use around the world[3], do crowd workers use mobile devices more commonly than laptops and PCs? Are some devices more suitable to specific task types than others? How up-to-date are crowd workers with the available software and

[1] http://www.mturk.com
[2] http://www.crowdflower.com
[3] Forecasted to reach 11.6 billion mobile devices by 2020, exceeding the world's projected population at that time (7.8 billion) [1].

hardware? How do UI elements and work environments interact while facilitating crowd work? By '*suitability*', we refer to the quality of being appropriate in supporting workers complete a given task at hand. For example, a work environment that is characterized by low Internet connection bandwidth may impede workers in completing tasks that contain high resolution images or videos, as opposed to a work environment that is characterized by a high Internet connection bandwidth. Similarly, some devices may be more appropriate in supporting workers complete tasks. For example, desktops or laptops are intuitively more appropriate for tasks that involve searching the web and retrieving information than mobile phones. Even in a novel system such as *WearWrite* [55], that enables users to write documents from their smartwatches, crowd workers who tested the system found issues with the interface such as its complexity and language. To elucidate, one user said "I don't know what a bullet point is", with one of the recommendations being better support in providing a cross-device functionality so workers can work on smartwatches, mobile phones or desktop computers. Understanding how crowd workers operate during their participation in microtasks can help us in the following ways:

- Improve crowdsourcing task design by facilitating better interaction of crowd workers with the tasks.
- Develop quality enabling and control mechanisms which leverage known affordances that workers rely on.

To this end, we carried out three studies. In the first study, we gathered responses from 100 distinct workers on CrowdFlower, in a survey regarding common issues with UI elements that workers faced while completing tasks in this microtask marketplace. Based on our findings, we ran a large-scale study to investigate the influence of work environments on worker performance across different task types, and the interplay with UI elements in tasks. We compared the impact of work environments on the quality of work produced by workers from India and USA, leading to several novel observations. In the third study, we carried out semi-structured interviews with Indian and American workers to validate the data-driven observations from the second study and to further understand the role of work environments in shaping the quality of crowd work. Based on the presented research findings, we developed a software tool for requesters to checks microtask for UI design issues and suggests ways to improve them before they are run on a crowdsourcing platform.

2 RELATED LITERATURE

"Ubiquitous computing is a method of enhancing computer use by making many computers available throughout the physical environment, but making them effectively invisible to the user... They weave themselves into the fabric of everyday life until they are indistinguishable from it."

– The most basic fundamentals of ubiquitous computing as described by Mark Weiser (1991, 1993)[4]

2.1 Crowdsourcing for Ubiquitous and Pervasive Computing

Over the last decade, many parallels have been drawn between crowdworkers and artificial intelligence systems; crowdsourcing even being called artificial artificial intelligence [34], making the worker and their work invisible, an issue that the field of 'Social Computing' has been trying to tackle over the years.

Prior work in participatory sensing [14, 63] have identified issues with usability in the design of crowd applications that could lead to significant data quality issues. For instance, Ding et al. have studied the issues with using crowd-based personal spectrum sensors (such as through smartphones and in-vehicle sensors) where the sensing data may have been unreliable or untrustworthy due to unexpected equipment failures or malicious behaviors, amounting to some abnormal data, and, making crowd-sensing schemes ineffective [14]. They discussed reactive measures to robustly cleanse out abnormal data components from the original corrupted sensing data. Hara et al. identified that heterogeneous devices come to be involved in crowdsourcing environments [27]. While finding generic infrastructures can be very difficult, they think having a generic reusable platform to support development

[4]http://www.ubiq.com/hypertext/weiser/UbiHome.html

of crowdsourcing applications would help in supporting heterogeneous mobile devices as well as manage large numbers of users. Chatzimilioudis et al. discuss multiple crowdsourcing applications by the use of smartphones [9]. They deliberate on the issues of running crowdsourcing applications with smartphones such as different Internet connection modalities (2G, 3G, 4G) each with different energy and data transfer rates.

In contrast to these previous works, we investigate the affect of (i) UI element design choices, and (ii) role of work environments (characterized by the software and hardware affordances at the disposal of workers) on the quality of work that is produced by workers. We also explore how both *good* and *bad* designs interact with the work environments. Thus, we shed light on the 'invisible' aspects of crowd work environment - a key component for participation in microtasking.

2.2 Task Types in Microtask Crowdsourcing

In our work we focus on the impact of crowd work environments and investigate their effects on different types of tasks. A taxonomy of task types in microtask crowdsourcing platforms has been developed in [21] where a two-level structure with 6 categories at the top level has been proposed. In our work we leverage such top level categorization to compare the effects of work environments on different types of tasks.

In our previous work we ran a large scale supervised classification job to analyze 130 million HITs published on Amazon's Mechanical Turk (AMT) over 5 years with the goal of understanding patterns in task type changes over time [12]. We observed, for example, that content creation tasks (i.e., where workers are asked to generate some content like an audio transcription or a document summarization task) are the most popular on AMT, and in our experiments, are the ones in which workers performed poorly (See Table 6). Another popular task type on AMT are surveys [12], as a crowdsourcing platform allows easy and immediate access to large populations of participants.

2.3 Worker Differences & Participation Bias

Crowd workers are not all the same. For example, different workers have different skills and interests. In [13] we previously showed that it is possible to profile workers based on their social network activities and assigned tasks based on such profiles which model their interests, to increase the accuracy of crowd work.

Other types of differences in the crowd that participates and completes specific tasks are caused by incentive schemes. For example, different reward levels may attract different types of workers [16] thus creating a participation bias in a study run on platforms like AMT. Jiang et al. analyzed the perceived benefits of participation in crowd work, and found that American and Indian workers differed in their perceptions of non-monetary benefits of participation. Indian workers valued self-improvement benefits, whereas American workers valued emotional benefits [36]. Hsieh and Kocielnik showed how different reward strategies (e.g., lottery-based reward models) result in different types of crowd workers deciding to participate and complete the available tasks. They highlighted the consequent difference in the crowdsourced task results [31]. Along the same line, in [28] authors showed that rewarding workers when they quit their participation in a batch of HITs allows to filter out low-quality workers early, thus retaining only highly accurate workers. Recently, Findlater et al. showed that results of online HCI experiments are similar to those achieved in the lab for desktop interactions, but this was less so in the case of mobile devices [20].

Prior work has studied the reasons that drive senior adults to participate in crowd work and show both a low participation of such population as well as an interest for incentive types that differ from monetary ones [6]. In contrast to this, in our work we analyze potential barriers to crowd work from a technological perspective showing important geographical differences in the type of devices and tools used by crowd workers which can also create participation bias in crowdsourcing studies. We also focus on how the technical infrastructure used by workers has an impact on participation and work quality in paid microtask crowdsourcing platforms.

2.4 Crowd Work Context and Barriers

Recently, ethnography-based research has been carried out to understand the contexts in which crowd workers are embedded. Authors of [49] focused on the Indian and US worker communities and highlighted the effects of current crowdsourcing platforms and marketplaces on crowd worker experience. In [48], authors studied how crowd workers on AMT use web forums to create communities where they share experiences and voice crowd work-related problems. McInnis et al. report on the qualitative analysis of 437 comments where AMT workers were asked to comment on parts of the AMT participation agreement through an online discussion website [50]. They indicate 'unfair rejection' as a major issue for workers, and identify risk factors that lead to this issue. Workers discuss 'flaws in task or interface design' as a key factor, while the authors suggest sounding an alarm through various tools to intimate requesters about a broken task.

Narula et al. noted that microtask marketplaces were often inaccessible to workers in developing countries, and introduced a mobile-based crowdsourcing platform called Mobileworks for OCR tasks, thereby lowering a barrier for participation [54]. Khanna et al. studied usability barriers that were prevalent on AMT, which prevented workers with little digital literacy skills from participating and completing work on AMT [40]. Authors showed that the task instructions, user interface, and the workers' cultural context corresponded to key usability barriers. To overcome such usability obstacles on AMT and better enable access and participation of low-income workers in India, the authors proposed the use of simplified user interfaces, simplified task instructions, and language localization. Vasantha et al. report an initial study of the demographic of 22 rural homeworkers in Scotland including computer skills, views on rural infrastructure and their skills in solving spatial visualization tests [66]. The authors present results equivalent to survey-based studies conducted in the past, and suggest that the homeworkers can solve knowledge-intensive industrial spatial reasoning problems with minimum training. They asked participants to report on their computer and Internet skills, to which most participants reported 'good', while some reported 'fair'. In their work, the authors also call for more research on rural infrastructure (such as Internet connection and road connectivity) that support crowdsourcing work, as most participants expressed satisfaction with their infrastructure yet a few did not find them adequate for crowdwork. Jones et al. explored what it means to be a mobile phone user situated at the lower end of the socio-economic ladder in developing economies like India, South Africa and Kenya to own and operate digital and mobile devices with almost no access to computers [37]. The authors suggest to engage with such users to help sketch out a technology road-map that will lead to devices and services which will be of value in the near future.

Several prior works have stressed the positive impact of good task design, clear instructions and descriptions on the quality of work produced [41, 47, 62]. However, as pointed out by Kittur et al. task interfaces are often poorly designed or even have bugs that make it impossible to complete tasks [42]. Poor quality work often arises from poorly designed crowdsourcing tasks. Morris et al. discuss the value of subcontracting microtask work and present value propositions for doing so [53]. In that they hypothesize a contracting model specifically based upon the need for task improvement such that workers can fix issues with user interface components and task structure amongst other things, which currently takes place by way of informal back-channels [23, 35].

In contrast to previous works, we aim to investigate the unexplored interaction between task design (through UI elements) and work environments (characterized by the technical hardware and software infrastructures that crowd workers use). We study the impact of these aspects on worker performance and advance the current understanding of the contexts in which crowd work takes place.

3 STUDY I : UI ELEMENTS

The aim of this first study was to identify typical problems that crowd workers face during interactions with different UI elements embedded in tasks. During the course of task completion, crowd workers are exposed to the various UI elements that may or may not be carefully designed by requesters publishing tasks on a crowdsourcing platform. Recent work that analyzed 5 years of crowd work on AMT [12], found that there is an organic growth in the number of new requesters (over 1,000 new requesters each month in 2013, 2014). Such new requesters are typically unfamiliar with the process of task design and may put less effort to ensure adequate UI design before deployment. Even experienced requesters do not necessarily consider the work environments in which crowd workers contribute to piecework. Prior studies have highlighted the importance of appropriate task presentation, reflecting on the impact it has on a worker's perception of task complexity, cognitive load, and eventually on worker performance within the task [2]. Recently, Yang et al. investigated the role of *task complexity* in worker performance, with an aim to better the understanding of task-related elements that aid or deter crowd work [75]. In this paper, we study the interplay between task design (in terms of the UI elements) and work environments (i.e., the context and differing conditions that crowd workers are embedded in). To this end, Study-I plays an important role to understand the typical issues that workers confront on a regular basis in crowd work.

3.1 Methodology and Survey Design

We designed a survey[5] asking workers about the issues that they typically faced with during their contributions in previous crowdsourcing tasks. The survey consisted of a few background questions, followed by questions corresponding to worker experiences while dealing with various UI input elements (*input boxes*, *text areas* spanning multiple lines, *checkboxes*, *dropdown menus*, *radio buttons* and *submit buttons*). Questions also covered other UI aspects such as *external navigation*, use of *colors*, experiences with *audio / video* content. To avoid misinterpretation, we presented workers with pictorial examples of each UI element. Finally, we provided workers with an opportunity to raise UI issues that were not addressed by the preceding questions in an open text field. We deployed the survey on CrowdFlower[6] and gathered responses from 100 distinct crowd workers. On average, each worker took just over 5 minutes to complete our survey and was compensated according to a fixed hourly rate of 7.5 USD on task completion. To detect untrustworthy workers and ensure reliability of the responses received, we followed the recommended guidelines for ensuring high quality results in surveys [22]. To this end, we interspersed two attention check questions within the survey. We also used the filter provided by CrowdFlower to ensure the participation of high quality workers only (i.e., *level 3* crowd workers as prescribed on the CrowdFlower platform). We flagged 7 (out of 100) workers who failed to pass at least one of the two attention check questions and do not consider them in our analysis.

3.2 Survey Results

We found that 43% of the workers who participated in the survey identified themselves as females (and 57% were males). Crowdsourcing microtasks served as a primary source of income for 42% of the workers. Table 1 presents the distribution of workers according to their age groups. We note a fairly even distribution of workers with respect to their age. As shown in Table 2, the workers who participated in the survey were also highly experienced in crowdsourcing work.

Based on the responses from workers, we observe that the issues raised can be distinguished between those that are a result of work environment constraints, and those that are a result of task design choices. By manually analyzing and aggregating the open-ended responses from workers and the responses to questions regarding different aspects of UI elements, we make the following key observations.

(1) **Input Boxes & Text Areas** – We found that 36% of workers raised issues that they faced with input boxes and text areas. 64% of the workers suggested that they did not experience problems in this regard. A recurring issue cited by workers with respect to input boxes and text areas was that the size of the input box

[5] https://sites.google.com/site/crowdworkenvironments/
[6] http://crowdflower.com

Table 1. Distribution of workers accoding to their age

Age	No. of Workers
18 - 25 Years	17.20%
26 - 35 Years	29.03%
36 - 45 Years	29.03%
46 - 55 Years	20.43%
Older than 55 Years	4.30%

Table 2. Experience of workers

Crowd Work Experience	No. of Workers
3 to 6 months	16.13%
1-3 Years	54.84%
3-5 Years	16.13%
Over 5 Years	12.90%

and character limit were disproportionate, often leading to only a part of the entered text being visible upon entry (mentioned by 5% of the workers). The following is an excerpt from a worker who raised this issue:

> *'Usually, text fields just work well as I can put things that I want in the list. The only big issue is when people want a long text answer like this one expect me to fit it into that small text field. It's so much harder to type what I want in and to proofread and edit my typed out text in a small text field. I could probably do my typing in an external text editor but why should I go through all that trouble when I can do all my typing in the browser. In general these things just work, though wrong usages like wanting large text in that small field is just terrible design.'*

Over a decade ago, researchers suggested that matching the size of input boxes to the expected length is an important guideline to follow when designing forms on the Web [10, 11, 72]. Another issue cited by workers was that of *input format validation* and *auto-correction*. Workers described occasions where text that was input was not accepted due to flawed format validation (especially, in case of URLs). These issues were cited by 15% of the workers. In other cases, input text was unfavorably auto-corrected, thereby hindering workers (mentioned by 6% of workers). Yet again, we found that a guideline that was suggested years ago with respect to accurate format validation [4] is sometimes violated by requesters during task design. Workers also reported that sometimes default text in the input field gets appended to the text that is entered, instead of disappearing on input. Finally, workers brought to light that input fields are not enabled sometimes, leading to a situation where it is not possible for workers to enter any response (mentioned by 6% of workers).

(2) **Checkboxes, Radio Buttons & Dropdown Menus** – We found that nearly 70% of the workers on average claimed to never have faced issues with checkboxes, radio buttons and dropdown menus. A recurring issue with checkboxes, radio buttons and dropdown menus was cases with too many options (mentioned by 10% of workers on average). This is a well-studied issue in UX design on the Web [3, 5]. Another common problem was found to be the small size of checkboxes, radio buttons or the dropdown menu icon (mentioned by 6% of workers on average). Several workers referred to issues with selecting checkboxes and radio buttons due to the active region not including the corresponding text, or multiple clicks being required for selection (mentioned by 10% of workers on average). The following is an excerpt from a worker who raised this issue corresponding to checkboxes:

> *'It's easier to mark a checkbox when not only the checkbox itself is [an active region], but a region around it too, because sometimes it's difficult to put the cursor in a region a little bit greater than a point. For instance, sometimes we can mark a checkbox just clicking on the line where the checkbox is in. Globally speaking, we can choose the alternatives we want in a more fast way, and we can complete the jobs more easily. But not everyone remember of build a job*

thinking of us this way, but I think it would be a good practice if did always, better for you and us too.'

Finally, some workers (approx. 5% on average) reflected on the difficulty to scroll within dropdown menus due to the small active region that disappears if the cursor moves outside the menu.

(3) **Audio & Visual Content –** 40% of the workers raised issues that they faced with tasks involving audio or visual content. Workers primarily raised issues related to the poor resolution and quality of audio and visual content within some tasks (mentioned by 25% of workers on average). The position and size of the visual content, loading and buffering times of the audio/visual content were other commonly cited problematic aspects (mentioned by 15% of workers on average). Some excerpts from the responses of workers that reflect these issues are presented below.

> *'Slow loading, waiting very long for the audio/video to load completely.'*
>
> *'I've faced multiple. Most issues are with the sound quality. Sometimes things are too quiet or have too much static which makes it hard to hear. With these issues I can't hear what I'm supposed to hear and can't really work with it. A lack of volume control is also a bit of a problem when sounds are of varying volumes and I want to not blow my ears out... Waiting to hear multiple sound files and do work on them is not so fun at all as it wastes time when I want to swiftly do work at my pace.'*

(4) **External Links & Navigation –** Over 50% of the workers reported issues they typically faced with external links and navigation. A recurring issue with external links was found to be the resolution of links to `http` URLs instead of `https` URLs on clicking. This results in warnings on some browsers and workers are often unsure about proceeding thereafter (nearly 25% of workers mentioned issues related to this). Other issues include opening links in the same tab instead of a new window or tab on clicking, broken links, and the loading time of the new page (nearly 20% of workers mentioned at least one such issue). Some excerpts from the responses of workers that reflect these issues are presented below.

> *'many dead links, many old test questions linking to sites that have been changes so the test question is no longer valid'*
>
> *'Some links don't work, pages never open, by the way, i lost a level 3 badge with a problem like this.'*

(5) **Colors Used & Submit Button –** We found that 80% of the workers did not typically face issues with the colors used in tasks. Some workers however, pointed out that poor color contrasts used in interfaces sometimes makes it hard to read the content of given tasks (nearly 15% of workers mentioned related problems). Around 50% of the workers claimed to have faced problems with the submit button. A common issue raised regarding the submit button was the poor positioning of the button (mentioned by 5% of the workers). Workers also complained that in some cases, the submit button was not enabled or had to be clicked multiple times (mentioned by 20% of the workers); another design violation [45]. Other issues pointed out include missing or unresponsive submit buttons and errors on clicking. Some excerpts from the responses of workers that reflect these issues are presented below.

> *'For me, some colors applied to data make it difficult to read the data - e.g., light grey.'*
>
> *'when you hit the keyboard 'enter' the task is automatically submit your work, even though you're not yet done.'*
>
> *'Sometimes the submit button didn't work at all. Unable to be pressed at all for a few minutes.'*

At first glance from a crowd worker's perspective, some of the issues raised might appear to be trivial to resolve or overcome through trial and error. However, prior ethnographic work has rightly stressed on the importance of considering the environmental context of crowd workers [26, 48]. It is also well understood that a fair portion of workers are not entirely familiar with using computers and other devices which play a crucial role in their participation on crowdsourcing platforms [40]. We thereby believe that these issues play an important role in shaping the quality of crowd work and the fluidity with which it is completed.

Next, we present the results of a study aimed at understanding what are the affects of the UI design issues identified so far in this first study on crowd work effectiveness.

4 STUDY II : WORK ENVIRONMENTS

The aim of this study is to understand how workers deal with UI design choices made by crowdsourcing requesters during the course of their work. We also investigate how the crowd work environments interact with the design choices and influence quality of the work produced. Thus, we address the following research questions.

RQ#1 : How is the performance of crowd workers influenced by design choices made by requesters with respect to UI elements in crowdsourcing microtasks?

RQ#2 : How do microtask crowdsourcing work environments influence the quality of work produced by crowd workers?

4.1 Methodology and Task Design

Based on responses from Study-I, we identified important aspects and design choices related to UI elements that crowd workers often encounter in microtasks. The companion webpage[7] presents these variations which either aid or hinder workers during the course of task completion. Depending on whether particular variations help workers or obstruct their work, we classify them as either *Good* or *Bad*, implying a good or bad experience as cited by workers in Study-I. In some cases, for the sake of completeness we additionally consider other extremities not mentioned by workers. For example, workers pointed out that disproportionately small text-areas (`ta_smallSize`) are troublesome to type into; we also consider the other extremity, that of disproportionately large text-areas (`ta_largeSize`). In other cases, catering to an extremity was deemed to be unnecessary owing to an unrealistic scenario. For example, UI elements with disproportionately large active regions such that options get selected by clicking anywhere on the window is an unrealistic extremity.

To analyze the influence of different design considerations with respect to UI elements on worker performance, and their interplay with varying worker environments, we manually created a batch of 129 microtasks accounting for each of the 43 variations (each variation $\times$ 3 tasks), shown in the table on the companion webpage.These tasks consist of different types; *information finding*, *verification and validation*, *interpretation and analysis*, *content creation*, *surveys* and *content access* [21]. The table in the companion page also presents sample tasks that we created corresponding to each of the UI element variations; these tasks are noticeably designed to reflect real-world microtasks that have previously been deployed on crowdsourcing platforms.

Since understanding work environments would be a crucial part of this study, we deployed two identical batches of the 129 tasks on CrowdFlower, one addressing workers based in USA and the other addressing workers based in India. We considered USA and India since they represent two of the largest populations of crowd workers [33], and due to the potentially different work environments they entail. In both cases, we used the inbuilt CrowdFlower feature to restrict participation to only the highest quality level of the crowd. We gathered 50 distinct judgments from workers in response to each of the 129 tasks in the batch, resulting in 6,450 $\times$ 2 responses (from USA and

[7] Companion Webpage - https://sites.google.com/site/crowdworkenvironments/

India workers). Workers were randomly assigned to tasks in the batch and the order of tasks was also randomized. Workers were compensated in proportion to the amount of work completed at a fixed hourly piece-rate of 7.5 USD.

While workers completed tasks, we recorded work environment related aspects in the background: the screen resolution of devices used by the workers, the CPU `speed` of machines used by workers, and the `user-agent string` using JavaScript embedded in the tasks. Note that CPU *speed* is computed by means of a Javascript benchmark using loops, hashing and random number generator functions, obtaining a score where the speed can be directly compared to that of a reference fast machine. To investigate further affordances at the disposal of workers such as input devices (keyboards or touchpads), mouses or mousepads and so forth, we implemented mousetracking using Javascript and the JQuery library, and logged user activity data ranging from mouse movements to keypresses. Our analysis regarding the hardware aspects of work environments are limited to these. We plan to extend the hardware features considered in the future.

4.2 Results and Analysis

Note that we report our results including the test statistic, degrees of freedom and effect sizes (Cohen's *d*, Hedge's *g*) for statistically significant findings. We acquired responses from 90 Indian and 95 American workers. The overall comparison between the two groups of workers is presented in Table 3. We did not find significant differences between Indian and American workers in terms of their overall accuracy or retention rate (i.e., number of HITs completed by each worker). However, we found that the American workers were significantly faster in completing the tasks, with an average task completion time (TCT) of 0.89 mins compared to the Indian workers (1.39 mins); *t(183)=3.06, p<.05, Cohen's d=.45*.

Table 3. Overall comparison between workers from India and USA. The asterix ('*') indicates a statistically significant difference between the two groups.

	INDIA	USA
No. of Workers	90	95
Avg. Accuracy of Workers (in%)	81.56	78.41
Avg. TCT (in mins)	1.39	0.89*
Avg. Retention Rate (in %)	55.56	52.63

Table 4. Overall comparison between workers from India and USA with respect to *good* and *bad* variations of UI elements. The asterix ('*') indicates statistically significant differences between the *good* and *bad* variations for a group.

	INDIA	USA
Avg. Accuracy (in%) *Good*	82.93*	79.51
Avg. Accuracy (in%) *Bad*	73.82	75.07
Avg. TCT (in mins) *Good*	1.49	0.76*
Avg. TCT (in mins) *Bad*	1.23	1.01

We also investigated the overall differences in the average accuracy and task completion times of Indian and American workers on *good* and *bad* variations. Table 4 presents our findings. By using multiple t-tests with Bonferroni correction, we note that there are statistically significant differences between the accuracy of Indian workers on tasks with *good* versus *bad* variations; *t(178)=2.15, p<.05, Cohen's d=.36*. On the other hand, we found that American workers exhibited a significant difference in the task completion time across *good* and *bad* variations; *t(188)=3.66, p<.001, Cohen's d=.29*, requiring significantly less time to complete tasks with *good* design variations as intuitively expected. We also observe that workers from USA take less time to complete tasks (see Table 3) than workers from India (independently from the *good* or *bad* design). In summary, this suggests that American workers deal better with tasks that are poorly designed.

4.2.1 Performance Across UI Element Variations. To understand how workers coped with tasks involving the different UI element variations, we analyzed the performance of workers corresponding to each UI element grouped into *good* and *bad* variations. The results of our analyses are presented in Table 5.

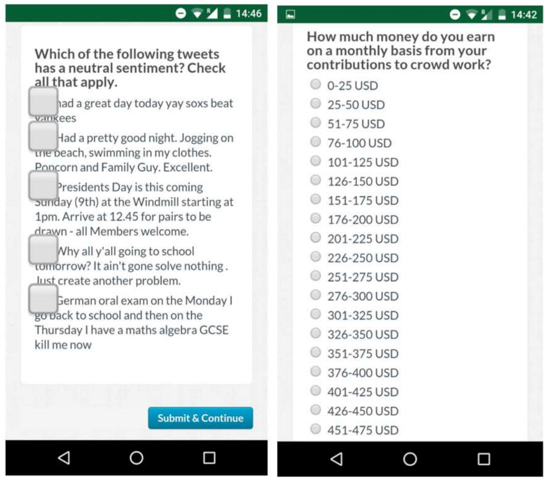


Fig. 1. Left– An *interpretation and analysis* task with large checkboxes (`cb_largeSize`); Right– A *survey* task with many radio buttons (`rb_manyOptions`), as rendered on an Android mobile phone and viewed on a Chrome browser.

Statistical Significance – We computed the statistical significance of our observations in Table 5, using multiple t-tests. To control for Type-I error inflation in our multiple comparisons, we use the Holm-Bonferroni correction for family-wise error rate (FWER) [30], at the significance level of $\alpha < .05$. Statistically significant differences between tasks with *good* and *bad* UI element variations are marked with an asterix (*).

- **Input Boxes** – Both Indian and American workers performed considerably better in tasks having *good* variations of input boxes, in comparison to *bad* variations. Further investigation revealed that disproportionately large input boxes (`ib_largeSize`) corresponded to the least accuracy: 30.67% in the case of Indian workers and 34.67% in the case of American workers. We did not find significant differences in the average task completion time (TCT) for both Indian and American workers between the two variations. This indicates that workers spent roughly the same amount of time on tasks with *bad* design variations despite the potential hindrance, reflecting their genuine attempt to provide high quality responses.
- **Text Areas** – In the case of tasks corresponding to text areas, Indian workers depict a better performance in the *good* variations when compared to the *bad* variations. No significant differences were observed for American workers. We found that both groups of workers performed with the least accuracy in the (`ta_smallSize`) variation where the visible size of the text area is small, making it inconvenient for workers to see all the text that is entered. Similar to our findings with respect to tasks with variations of input boxes, we did not find significant differences in the average TCT of workers between the *good* and *bad* variations.
- **Checkboxes** – We found that both Indian and American workers performed much better in tasks with *good* variations of checkboxes than in *bad* variations. An example of a bad variation is shown in Figure 1. In contrast to Indian workers where we found no significant difference in their average TCT, the American workers were faster in tasks with *good* variations. Further investigation revealed that the (`cb_manyOptions`) variation corresponded to the least accuracy among both Indian and American workers in tasks with checkbox variations.
- **Radio Buttons** – Tasks with variations of radio buttons correspond to better performance from both Indian and American workers in the *good* as opposed to in the *bad* variations. However, the performance of workers in the *bad* radio button variations when compared to that in case of the checkboxes is significantly

higher (Indian workers: *t(125)=6.452, p<.001*, American workers: *t(163)=4.805, p<.001*). This is explained by the simpler nature of radio buttons, where questions are modeled to accept a single response (even in case there are many options; as in `rb_manyOptions`) (an example is shown in Figure 1).

- **Audios –** In tasks with audio variations, we note that compared to other tasks, the overall level of accuracy drops in the case of Indian and American workers, both in the *good* and *bad* variations. We attribute this drop in accuracy to the inherently more complex nature of audio transcription tasks [57], where work environment specifics (such as device volume, headsets or other equipment) may play a role. This is exacerbated in the case of audios with poor quality (`audio_poorQuality`). We found that Indian workers perform considerably better in audio transcription tasks with good quality variations when compared to the poor quality variations. They also take more time on tasks with the `audio_goodQuality` variations. On further scrutiny, we found that in tasks with poor audio quality, several Indian workers gave up after trying to transcribe the audio, condemning the poor quality of the audio in their responses. In contrast, we found that American workers performed similarly in both the *good* and *bad* variations, without a significant difference in the average TCTs.
- **Images –** We found that Indian and American workers performed similarly in the tasks with either *good* or *bad* image variations, without significant differences in TCTs between the two variations. Across both groups of workers we found that the `img_smallSize` variation corresponded to the lowest accuracy of workers in image variation tasks.
- **Videos –** In case of tasks with videos, we note that both Indian and American workers do not exhibit a significant difference in their performance between the *good* and *bad* design variations. We found that American workers took less time to complete the tasks with *good* variations as opposed to *bad* variations (i.e., tasks with the `video_poorQuality` variation).
- **Dropdown Menus –** In the case of tasks with dropdown menu variations (related to active region and icon size), we found that both Indian and American workers perform with similar accuracy and take similar amounts of time in both the *good* and *bad* variations.

  External Links – We found no effect on the accuracy of Indian and American workers or their task completion times based on the type of external links (`elink_HTTP` or `elink_HTTPS`), or whether the links opened in the same or new tab (`elink_sameTab` or `elink_newTab`).
- **Rating Scales –** Due to the subjective nature of rating scales, we only consider the TCT of workers as a measure of performance. We did not find significant differences across different rating scales.

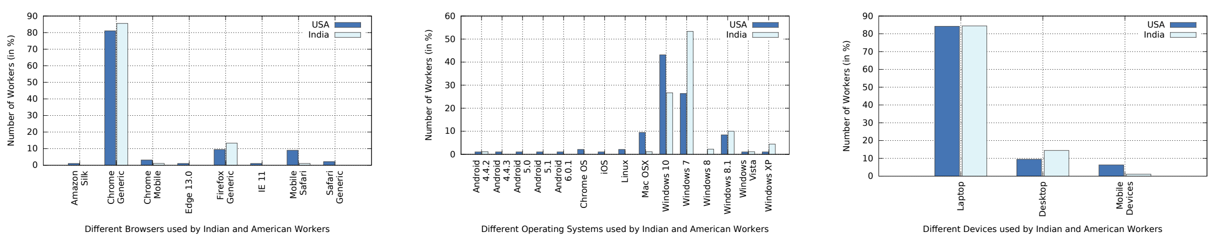
4.3 Role of Work Environments

In the earlier section, we presented our findings with respect to the performance of workers in tasks with different UI element variations and found several differences when comparing American and Indian workers. With an aim to understand whether work environments influence the performance of workers, we investigated the aspects that characterize the work environments of the Indian and American workers in our studies.

4.3.1 Browsers, Operating Systems and Devices. By resolving the `user-agent` strings of workers who participated in our tasks, we identified their browsers, operating systems and devices. The distribution of workers according to these specifics are presented in Figure 2. We found that there is a far greater variety in browsers, operating systems and devices used by crowd workers from the USA than those from India. However, the most popularly used browsers (*Chrome Generic*, *Firefox Generic*), operating systems (*Windows 7*, *Windows 10*) and devices (*Laptops*, *Desktops*) are similar across both groups of workers. It is noteworthy that the American workers appear to use more of the latest available technology such as the Windows 10 operating system and Macbooks, in comparison to Indian workers in our tasks.

Table 5. Performance of Indian and American workers corresponding to each of the UI elements, grouped into *good* and *bad* variations. Statistically significant differences between tasks with *good* and *bad* UI element variations are marked with an asterix (*).

	India		USA	
UI Element Variation	Avg. Accuracy [%]	Avg. TCT [min]	Avg. Accuracy [%]	Avg. TCT [min]
Input Boxes - Good	**85.95***	1.21	**85.08***	0.81
Input Boxes - Bad	71.64	1.25	70.62	0.99
Text Areas - Good	**80.79***	1.57	76.62	0.98
Text Areas - Bad	65.69	1.55	72.02	1.23
Check Boxes - Good	**88.29***	1.19	**93.55***	0.51
Check Boxes - Bad	63.11	0.86	69.86	**0.83***
Radio Buttons - Good	87.19	0.81	86.89	0.46
Radio Buttons - Bad	85.65	0.55	81.08	0.51
Audios - Good	**38.28***	**5.33***	43.99	3.01
Audios - Bad	22.75	3.80	44.59	3.12
Images - Good	66.52	1.61	65.85	1.13
Images - Bad	67.24	1.43	68.14	0.96
Videos - Good	90.8	1.16	88.73	0.89
Videos - Bad	84.08	1.45	86.61	**1.12***
Dropdown Menus - Good	98.67	0.59	92.51	0.48
Dropdown Menus - Bad	98.88	0.60	95.4	0.52
External Links - HTTP	98.04	0.70	96.92	0.58
External Links - HTTPS	98.77	0.74	96.23	0.60
External Links - Same Tab	95.54	0.60	98.41	0.68
External Links - New Tab	96.79	0.79	97.62	0.76
Rating Scales - Horizontal	–	0.72	–	0.49
Rating Scales - Vertical	–	0.83	–	0.45
Rating Scales - Slider	–	0.56	–	0.52

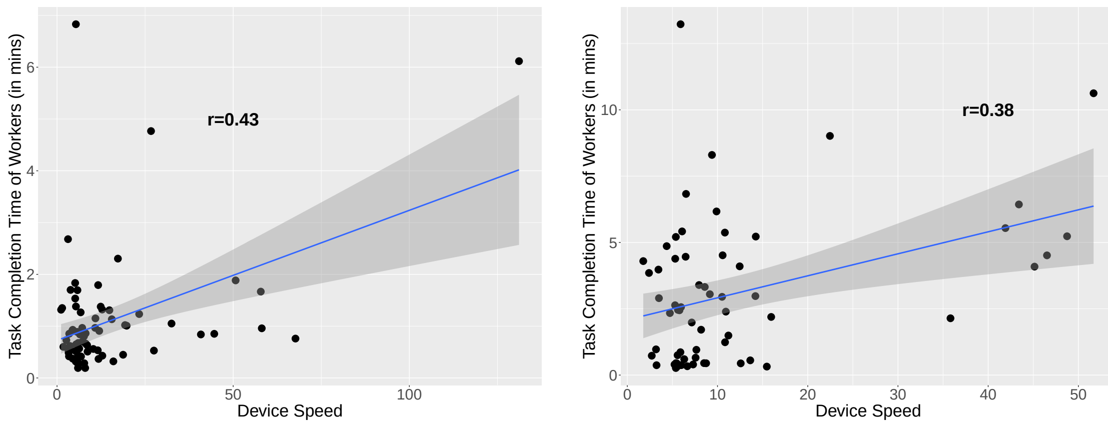


(a) Browsers (b) Operating Systems (c) Devices

Fig. 2. Distribution of Indian and American workers in the 129 tasks according to their browser, operating system and devices used.

4.3.2 Impact of Device Speed. We investigated the relationship between the speed of devices that Indian and American workers used, and their accuracy and TCTs across different tasks. Figure 3 presents the statistically significant relationships that we found.

We found a moderate positive correlation between the speed of the device used and the task completion time of American workers in tasks with text areas (see Figure 3a); $r(72)=.43$, $p<.001$. This suggests that American workers use more time to complete tasks with text areas when the device used is relatively faster. Thus in tasks with text areas, we found that the speed of the device used by American workers accounts for nearly 18.5% of the variance



(a) *TCT* and *device speed* of American workers who completed tasks with *text area* variations.

(b) *TCT* and *device speed* of American workers who completed tasks with *audio* variations.

Fig. 3. Impact of device speed on different types of tasks (only statistically significant linear relationships are presented).

in their task completion times (the coefficient of determination, R^2=0.184). Further scrutiny revealed that faster devices led to American workers providing more tags as well as more unique tags in the tagging task (see Table 1 in the companion webpage) corresponding to text areas. We investigate and discuss these findings further in a follow-up qualitative study described in Section 5.

Similarly, in tasks involving audio media, we found a moderate positive correlation between the task completion time of American workers and the speed of the devices used (see Figure 3b); *r(62)=.38*, *p*<.001. Accordingly, the speed of the devices used by American workers in tasks with audio variations accounted for nearly 14.5% of the variances in their task completion times (the coefficient of determination, R^2=0.144). We did not find significant correlations with respect to the devices used by Indian or American workers across other task variations.

4.3.3 Impact of Devices on Worker Performance.

Next, we investigated the impact of devices used on the performance of Indian and American workers (i.e., their accuracy and task completion time) across the different UI element variations.

In the case of Indian workers, we found that workers who used desktop computers needed significantly more time to complete tasks (*M=1.55, SD=.99*) when compared to those who used laptops (*M=1.18, SD=.76*) in tasks with *input boxes*; *t(78)=1.528*, $p < .05$, Hedge's *g=.46*. On investigating why this was the case, we found that the speeds of the desktop computers used by the Indian workers was significantly lower (*M=5.96, SD=19.36*) than the laptop computers, probably indicating that desktops are older machines (*M=20.05, SD=27.34*); *t(78)=1.768, p<.05*, Hedge's *g=.54*. Indian workers who used laptops exhibited a significantly higher accuracy *(M=68.45, SD=20.00)* than those workers who used desktops to complete tasks with *check boxes (M=79.48, SD=16.73)*; *t(78)=2.043*, $p < .05$, Hedge's *g=.64*. We did not observe a significant impact of devices used by Indian workers on their performance in tasks with other UI element variations.

American workers who used laptops completed tasks with *text areas* in a significantly faster time *(M=1.09, SD=1.13)* than those who used desktops *(M=1.92, SD=1.40)*; *t(80)=1.829*, $p < .05$, Hedge's *g=.72*. On investigating why this was the case, once again we found that the speeds of the desktop computers used by the workers who completed tasks with *text areas* was lower (*M=6.61, SD=1.71*) than those who used laptops (*M=22.4, SD=35.18*). American workers who used laptops exhibited a higher accuracy *(M=63.99, SD=26.88)* than those who used desktops *(M=80.11, SD=21.99)* in tasks with *check boxes*; *t(80)=1.932*, $p < .05$, Hedge's *g=.7*. Similarly, American workers who used laptops performed more accurately *(M=89.95, SD=20.37)* than those who used desktops in tasks with *videos (M=72.86, SD=34.82)*; *t(80)=1.947*, $p < .05$, Hedge's *g=.78*.

Our findings regarding the impact of devices on worker performance are further explored and discussed in Section 5, through follow-up individual interviews with Indian and American workers.

4.3.4 Impact of Screen Resolution. We analyzed the screen resolution of devices used by Indian and American workers to investigate the potential influence of screen resolution on worker performance in tasks with different UI element variations.

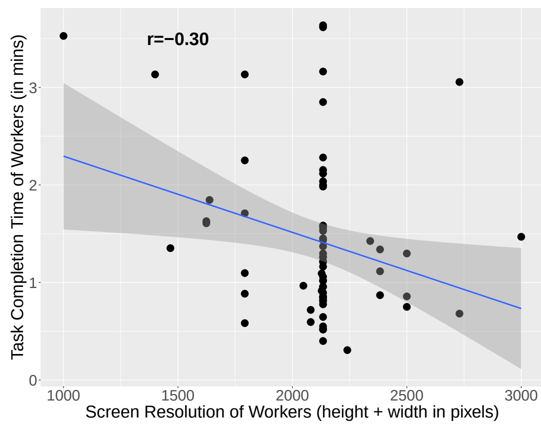


Fig. 4. Relationship between *screen resolution* and *TCT* of Indian workers who completed tasks with *image* variations.

We found that most Indian and American workers used devices with a high screen resolution; many reporting a *HD* screen resolution of 720 × 1280 or higher. 84 out of the 90 Indian workers were found to be using devices with HD or higher screen resolutions, while 85 of the 95 American workers did the same.

We found a weak negative correlation between the screen resolution and the task completion time of Indian workers in tasks with UI element variations corresponding to images (see Figure 4); *r(88)=-.30*, *p*<.001. This indicates that lower screen resolutions can hinder workers in tasks that involve images, resulting in longer task completion times. Accordingly, the screen resolution of Indian workers accounted for 9% of the variance in their task completion times (the coefficient of determination, R^2=.09). We did not find significant correlations with screen resolution of devices across other UI element variations.

4.3.5 Impact on Different Task Types. Based on the taxonomy of microtasks proposed in [21], we analyzed the impact of work environments on the different task types; *information finding* (IF), *interpretation & analysis* (IA), *verification & validation* (VV), *content creation* (CC), *surveys* (SU) and *content access* (CA). Table 6 presents the distribution of the 129 tasks according to the taxonomy, and the overall average work accuracy of Indian (IND-Acc) and American (USA-Acc) workers corresponding to each task type. While we found differences in accuracy within each group across the different task types, we did not find statistically significant differences in worker performance between Indian and American workers across the task types.

Although we did not find significant differences in worker accuracy in each of the task types across the devices (desktops, laptops or mobile devices), we found that in information finding (IF) and content creation (CC) tasks, both Indian and American workers using mobile devices required significantly more time for task completion. This indicates that laptops and desktops are more suitable than mobile devices for certain task types. We reason that content creation tasks typically involve typing content, which is inherently easier to accomplish using a keyboard

Table 6. Distribution of the tasks deployed according to their type.

Task Type	#Tasks	% Tasks	IND-Acc (in%)	USA-Acc (in%)
CC	24	18.60	64.50	68.67
IA	39	30.23	83.85	85.64
IF	12	9.30	69.00	69.83
SU	24	18.60	97.50	84.58
VV	30	23.26	79.00	79.53

on a laptop or desktop computer, as opposed to a mobile device (considering that it is easier to deal with active regions corresponding to input boxes and text areas on laptops and desktops). In the case of information finding tasks, workers are typically required to search the Web, return to the task, and provide their responses. We reason that such toggling between tabs or windows is easier to accomplish on desktops and laptops in comparison to mobile devices.

In our final study, presented next, we aim at validating such hypotheses we draw based on the analysis of collected data on the affect that work environments have on the performances obtained by workers.

5 STUDY III : FOLLOW-UP PERSONAL INTERVIEWS

In Study I, we investigated the typical problems that crowd workers faced owing to UI element design in tasks. In Study II, we revealed how work environments of crowd workers interacted with various UI element design considerations. We analyzed the consequent impact on the quality of work produced by American and Indian workers in different types of tasks. With an aim to understand our findings in Study II better, and whether or not the observed correlations between work environmental aspects and the quality of work produced can be supported by further evidence, we conducted a follow-up study (Study III) involving personal interviews of American and Indian workers from CrowdFlower.

5.1 Methodology

To better understand the influence of work environments on crowdsourced microtasks, we conducted 7 semi-structured interviews with CrowdFlower workers [65] who completed all tasks in the batches described earlier in Study II.

We randomly selected 20 American and 20 Indian workers who completed all the tasks, and contacted them via e-mail requesting their participation in a follow-up interview over Skype or Google Hangout. In the email[8], workers were notified about the purpose and nature of the interview, the participation reward and mode of compensation (i.e., a bonus payment on CrowdFlower according to the hourly-rate of 7.5 USD), and an assurance of anonymity. We sent out the recruitment emails to workers nearly 3 months after their participation in Study II, to avoid any bias in their responses and perception of tasks stemming from the recency of participating in our batch of tasks with *good* and *bad* variations. 5 American and 10 Indian workers showed interest to participate in the interviews. Of these, 2 American and 5 Indian workers scheduled and completed the interviews. Table 7 presents some characteristics of the interview participants in Study III.

Two of the authors conducted semi-structured interviews with the interested workers. Participants were first briefed about the identity and work of the authors, and the structure of the interview. They were informed that, with their consent, the interviews would be audio recorded and transcribed. After receiving a verbal consent, workers were asked a series of questions ranging from their general background, experience and motivation for participation

[8]Full text of the recruitment email is available at the companion webpage – https://sites.google.com/site/crowdworkenvironments/

Table 7. Characteristics of American (AW) and Indian (IW) Interview Participants

ID	Gender	Age	Education	Experience	#Tasks Completed	Income
AW1	F	50	Some college, no degree	4.5 years	> 200,000	Secondary
AW2	F	63	Bachelor's degree	5 years	> 4,000	Secondary
IW1	M	41	Post-graduate diploma	7 months	> 1,000	Secondary
IW2	M	24	Bachelor's degree	1.5 years	> 50,000	Primary
IW3	M	32	Master's degree	2 years	> 10,500	Secondary
IW4	M	37	Bachelor's degree	5 months	> 5,000	Secondary
IW5	M	32	Bachelor's degree	9 months	> 4,000	Primary

in crowd work, to details about their work environment. The wordings of questions were made intentionally neutral to elicit responses from participants without being influenced by the interviewer. Some of the questions asked during the interviews are presented below.

A sample of questions asked during the interviews with participants.

- Since when have you been using computers? How often do you make software/hardware upgrades?
- Which platforms or websites do you use to participate and contribute to crowd work?
- What type(s) of device(s) do you use for your participation in crowd work?
- Do you switch between devices? If yes, when and why?
- Based on your experience, what are your thoughts on the suitability of devices to different types of tasks that you encounter?
- What type of Internet connection do you have? How much do you pay for it on a monthly basis? How would you describe your Internet connection in the context of the tasks you typically complete on crowdsourcing platforms?
- What are your most memorable experiences with different UI elements and media types (input/text boxes, checkboxes, radio buttons, images, audios, videos) that you encounter in tasks?
- Based on your experience, how would you describe the importance of being proficient in the English language for participating in crowd work?

Interestingly during the participant recruitment phase, far more Indian workers were willing to participate in the Study III than American workers, given the quick and robust response to the survey study. These interviews were used to elicit details about the worker that were beyond the scope of Study II such as determining language proficiency of the workers, and their own perspectives regarding the effectiveness of their work environments. Study III also provided an opportunity to gather further personal and subjective data, such as to questions: what did workers do when their broadband was slow, or when they had issues with a task which they were unable to solve from their end. In the next section we share the results from this qualitative study with supporting anecdotes from the participants, as well as highlight the themes that were interesting.

5.2 Results and Findings

We transcribed the audio recordings after the interviews and qualitatively assessed the responses of participants. On average, each interview lasted for approximately 20 minutes. Our focus was to look for instances of hardware, software and task design related issues and any demographic and socio-economic factors that might influence 'the doing of the work' in the workers' work environment. By closely examining their responses, we identify and unearth the following prevalent details regarding work environments and their influence on the quality of work produced. We summarize the results concerning work environments under three main themes below: (1) Device usage by individuals (2) Internet, device speed and screen resolution (3) Language proficiency. We then discuss what the participants' perspectives were regarding dealing with poor design.

5.2.1 Work Environments – Device Usage. Through our interviews we found an interesting mix of rationale behind the choice of devices being used by workers for participation in microtasks. 4 of the Indian workers (**IW1**, **IW2**, **IW3**, **IW5**) claimed to use multiple devices for participating in crowd work. **IW1** said he switches between a laptop and a desktop, depending on where he is in his 2-storey house. **IW2** claimed to switch between using his laptop and mobile phone depending on the type of the task; he said that some tasks are convenient to complete on his mobile phone, such as image validation or sentiment analysis. In those cases, he lies down and uses his mobile phone. Otherwise, **IW2** sits at a desk and uses his laptop. **IW3** uses his laptop for participating in crowd work when he is at home. When he is at work or traveling, or completing microtasks during his free time, he uses his tablet. Similarly, **IW5** uses his desktop while participating from home, and uses his laptop when outside.

In contrast to these workers, **IW4**, **AW1** and **AW2** reported the usage of a single device to complete crowdsourced microtasks. **IW4** uses an assembled desktop computer, **AW1** uses a laptop, and **AW2** uses a desktop. **AW1** indicated her preference to use her laptop by saying, "*I only use my laptop. I think mobile devices are too small, (my) laptop is big. I like to be able to see what I'm doing. So there we go!*". **AW2** indicated her preference to use her desktop since most of the tasks she completes involve typing, and she is better at typing on a desktop with a keyboard. She also said, "*I can see the screen better on my desktop*".

Workers differed in their views on the potential influence of device type on their performance. **IW1** feels that there is no difference in the way he performs using either his laptop or desktop. He added, "*I don't think there is much impact of type of tasks on the device I choose to use*". **IW2** believes that some tasks are more suitable to complete on laptops or desktops. He said, "*(In) 90% of the tasks I have to use my laptop. Because sometimes it is easy to copy-paste when I have to find business profiles or (such) related tasks. I have to constantly go to Google and find some information related to the tasks*". **IW5** expressed a preference to use a desktop, saying that "*Desktop is more convenient because of the keyboard and mouse*".

5.2.2 Work Environments – Internet Connection, Device Speed & Screen Resolution. We asked workers about the quality of their Internet connections as well as their devices. **IW1** reported having a 2 MBPS unlimited broadband connection. His desktop is 10 years old, and his laptop is 2 years old. **IW1** said he experienced issues due to bandwidth limitations in tasks that contained images. **IW2** mentioned that his laptop is 2 months old, and that he replaced the old one due to the slow processor, and small screen size and keypad. He reported that he bought a new laptop to mainly do CrowdFlower tasks and described his Internet connection as follows; "*In the past I had a very slow Internet connection. But now I have high-(speed) and unlimited Internet connection and everything is fine. New laptop is making a huge difference in speed and especially helping my accuracy*". However, he cited having issues with loading tasks that contain media such as images, audios/videos. **IW3** also reported having an unlimited broadband connection. He explained that despite his high speed connection, he faced problems with tasks containing videos, "*Sometimes I feel a bit odd, some videos might not be like buffering fast and I have to wait and wait for sometime. And yesterday I did some tasks and the videos were buffering, so too slow. I got a bit frustrated. I think it was a problem with that task actually. It's not happening with all the tasks, but some tasks, videos are not opening, its taking time to buffer, even though my Wi-Fi has speed but its taking time, buffering buffering buffering. I just pressed the play button and went to the kitchen and did my cooking and then yeah, came back and started doing the task. Luckily had enough time to do that. It's a bit more time-consuming. I have to be looking into it, let it buffer, it takes time. So I am losing time, time is money. If I can do it fast I can go for the next task. It's money*". **IW4** reported having a dial-up connection that he uses via a modem. He said, "*Sometimes the Internet is slow, so I convert my Internet to 3G to download images. Time is sufficient, it downloads ... to finish it* *before the task ends*". **IW5** uses a data card. Referring to the slow speeds he deals with, he said "*The service is pathetic, very slow Internet. The network drops often, sometimes in the middle of task and this means a lot of time and effort wasted*". In contrast, **AW1** affirmed that she has no issues with her Internet connection. She said, "*I have high bandwidth, the highest you can get*". **AW2** also claimed to have a high-speed Internet connection and said she did not face any problems due to her Internet connection.

The workers were divided in their opinions on the importance of screen resolution in enhancing their performance. Most workers believed that they did not face problems, and that their performance was not hindered due to their screen resolutions (**IW1, IW4, IW5, AW1**). A few workers indicated that high screen resolution can help improve performance (**AW2, IW2, IW3**). **IW2** said, "*I totally think screen resolution also makes a difference. With a better resolution you can do the task more easily*". **IW3** said, "*I always set the background to proper color. I always set brightness for each task. Some (tasks) have pictures or some (tasks) have videos. So, I always set the background to adjust my view and get that clarity*". The data collected during Study-II showed that lower screen resolutions can hinder workers in tasks that involve images.

5.2.3 Language Proficiency. Previous works [25, 40] have studied how work is completed within different crowdsourcing tasks on AMT and establish that there are language proficiency differences amongst workers. These differences where shown to range from bare-minimal functional knowledge of English to highly sophisticated, contextual use of the language. Language proficiency can have a large impact in tasks with poorly designed UI elements, creating amplified difficulties for workers. For example in tasks with poor audio or large interface elements that hide text, workers who are highly proficient in the language have a better chance of guessing unclear words and providing accurate transcriptions and responses. The range in proficiency became apparent in Study III as we asked participants questions about their use of digital devices for personal use in the day to day, and gauged their effective use of English during the interviews. For example, worker **IW4** who has a degree in Homeopathic Science and ran a Homeopathic clinic in a small town operating in a regional language in India, acquired a basic education of English but hardly used the language on a daily basis except for when he worked on CrowdFlower.

Most workers suggested that language proficiency can aid in quick and accurate completion of tasks (**IW1**, **IW2**, **IW5**, **AW1**, **AW2**). **IW1** believed that he is proficient in English, and that this gives him an edge over other workers. He said language proficiency can improve speed and accuracy in tasks; he can grasp things faster while working on different types of tasks. **IW2** said, "*Instructions are very simple usually. But you need to be very good at English, otherwise you cannot do the tasks properly. For me the instructions are very simple. But you need to have a grasp over English. If your English is not that good, it will take longer to complete the task. Even if you are a new contributor it takes time. If you are experienced you will see the same tasks repeatedly and you won't have to read the instructions again*". Similarly, **IW5** believed that experienced workers can overcome language related impediments. He said, "*Language fluency does affect your performance. But it is a one time thing, and not much of an issue. I can take the blame for it sometimes, where maybe I don't understand the task correctly because English is not my native (language)*". **AW2** said, "*Language fluency helps speed. It definitely is an advantage over non-English speakers*". **AW1** said, "*Sometimes there are issues with how the directions are explained. This doesn't happen very often. But this may also be due to how experienced I am. If I feel something is not right, I don't do those tasks. I've done tasks for 4 years now. So I know which tasks I should avoid. In some tasks you definitely need a better grasp over English than others. Maybe in about 50% of tasks if you have a good grasp over English you can be faster in completing the tasks. Half the times you can benefit if you're fluent*".

5.2.4 Dealing with Issues Emerging from Poor Task Design. Participants gave us examples of practices they followed to deal with poor task design. **IW2** said, "*Some tasks can be time consuming. Sometimes the instructions are very difficult to understand, other times they are very lengthy...takes 30 mins just to read and understand. But I still do them when I really need the money. But these days I try to find tasks which pay well*". **IW5** followed a similar strategy. He said, "*If the pay is good I put in extra effort. Otherwise, I still work on such tasks only if no* *other tasks are available*". **IW4** said, "*I try to understand what the task is and then do it, and do it in good time. I set no targets...try to finish the task before the completion time that's all*". In contrast, **AW1** and **AW2** said that they tend to skip tasks that are poorly designed or unclear.

5.2.5 Summary of Insights from Study III. Summing up the key points from Study III, workers use multiple digital devices, often switching between them, to carry out microtasks based on their social contexts and personal needs, for example, a larger screen for tasks that require a fair amount of reading. There are important implications of this on microtask design; requesters need to be mindful of device usage and switching to enable better interactions and improve the quality of work.

The Internet connections that workers used varied from very slow 56K dial up modems to 2 MBPS broadband, high-speed unlimited download. We also found a variety of devices and screen resolutions: traditional mobile phones, second-hand desktop computers to smart phones and laptops via different browsers, running on varying versions of operating systems and document editing suites. These insights calls upon requesters to be more mindful about the resources and work environment of their workers when designing and allocating tasks, since this directly affects the quality of work.

Poor language proficiency aggravates the difficulty in successful task completion for workers, especially for novice workers who do not have a deep, contextual knowledge of the English language. Instructions given in a task work hand in hand with the UI design and flow of the task, and hence need to be consistent and complementary to each other. Poorly designed UI elements can exacerbate language proficiency related constraints that workers may have, adversely affecting their quality of work.

Based on the studies we described above, the next section discusses the implications of our findings. We introduce a new tool, *ModOp*, to help requesters design tasks that consider how UI elements interact with the varied work environments we found during our research.

6 MAIN FINDINGS AND IMPLICATIONS

Based on the three studies presented so far, we list some of the key and novel findings presented in this paper.

General Insights –

(1) Prime examples of poorly designed UI elements that negatively impact crowd worker performance are large input boxes, disproportionately small text areas, and multiple-choice questions having many radio buttons/check boxes.
(2) In information finding and content creation tasks, workers using mobile devices required significantly more time for task completion in comparison to those using laptops or desktops.

American Workers versus Indian Workers –

(1) American workers on average were faster and performed better in tasks with poorly designed UI elements compared to Indian workers across all task types and considering all work environments.
(2) American workers outperformed Indian workers in audio transcription tasks (performing well in tasks with poor quality audio as well).
(3) More variety was observed in the work environments of American workers than Indian workers. This variety was also concomitant with more recent technology (latest operating systems, browsers) in the case of American workers.

American and Indian Workers : Finer Details –

(1) American workers with faster devices (laptops were found to be faster than desktops) provided higher quality responses (more tags, more unique tags) to questions with text area variations and audio media. We found a positive correlation between speed and accuracy of American workers using laptops. The workers using laptops also performed more accurately than those using desktops in tasks with video media.

(2) Indian workers using laptops were found to be faster than those using desktops in tasks with input box variations. Personal interviews with workers in Study III, revealed that this could potentially be due to old and outdated desktop computers.
(3) Low screen resolutions induce longer task completion times for Indian workers in tasks containing images.

The main implications for researchers and practitioners planning to use microtask crowdsourcing with the aim of conducting pervasive and ubiquitous computing experiments are the following. When the data to be labeled is large (e.g., video streams) it may be more appropriate to target American workers as their environments appear to be better on average to sustain the load, with higher bandwidths and hardware specifications in comparison to Indian workers on CrowdFlower. Requesters should always follow good UI design guidelines during crowdsourcing task design: we have observed that there is a strong negative effect of badly designed tasks on worker performance, and this is exacerbated in cases where workers have less suitable work environments. Along this line, planning for reactive UIs that can nicely adapt to different work environments would empower many workers with the capability of being even more effective. Our main findings have important implications on task allocation in crowd work. Tasks that require fast execution or those that may benefit from certain work environments can be allocated to specific workers with suitable work environments. To implement this, workers in crowdsourcing marketplaces consisting of a heterogeneous batch of tasks may be assigned the 'most suitable' tasks based on their current work environment (e.g., do not allocate information finding tasks to workers currently on a mobile device). A final implication of our results is that tasks should be adapted based on work environments. A first step in this direction is a tool that we describe next to support better task designs, taking into account work environments.

6.1 ModOp – A Tool for Environment-Aware Task Design

Based on our findings we developed a software tool, *ModOp*[9] to help requesters crowdsource better work and make environment-aware HIT designs. *ModOp* parses HITs as HTML forms and guides a requester during the task design phase, by providing appropriate warnings and feedback according to our key findings. The elements that our tool monitors are:

- *Input Box / Text Area size* – Warning triggered if the size is disproportionately small or large.
- *Image size and resolution* – Warning triggered if the image size or resolution is disproportionately small.
- *Checkboxes* – Warning triggered if number of checkboxes is not optimal; requesters are advised to split checkbox questions where there are more than 10 options.
- *Radio Buttons* – Warning triggered if the number of radio buttons corresponding to a question is > 4, based on [5].

Apart from the design feedback that is provided by *ModOp* on-the-fly, the tool can also supports requesters in making work environment-aware decisions during task design.

- *Device Type – ModOp* automatically detects the device type of workers by leveraging the user agent string.
- *Device Speed – ModOp* automatically provides an estimate of the worker's device speed based on a target relative speed.
- *Screen Resolution – ModOp* automatically detects the screen resolution of the worker's device.

With minimal effort, requesters can integrate *ModOp* into their task desgin workflow and make informed work environment-aware decisions; this can facilitate more inclusive pay schemes (for example, awarding bonuses to workers for good performance despite poor work environments), shape task assignment and routing (for example, routing tasks that contain high resolution media to workers with large Internet connection bandwidths and fast devices), and have broad implications on fairness and transparency in crowd work.

[9]The *ModOp* tool is available for public use as Chrome browser extension and bookmarklet at http://github.com/AlessandroChecco/ModOp.

We believe that this tool can help crowdsourcing requesters in designing better tasks by directing requesters' attention to otherwise neglected attributes of UI element design and work environments. We envision that this would improve the overall crowd worker experience and help in reducing their cognitive load [71].

6.2 Evaluation of *ModOp*

We performed an evaluation of the ModOp tool using real-world microtasks by considering the impact of *ModOp* on the cognitive load of workers. Cognitive load refers to the total amount of mental effort being used in the working memory of a person with respect to the task at hand. Early work showed that instructional design can be used to reduce cognitive load in learners [64]. Cognitive load theory has been used widely to inform and evaluate several web-based tools and systems; Feinberg et al. proposed to leverage cognitive load to inform web-based design [17], Oviatt used the theme of cognitive load to design human-centered interfaces [56], Wang et al. studied website complexity from the cognitive load perspective [69], Schnabel et al. proposed the use of shortlists to improve the performance of recommender systems and reduce the cognitive load of users [60]. Thus, our intuition behind using cognitive load as a metric to evaluate *ModOp* was to observe whether the input propelled by *ModOp* related to the design of UI elements affects the task perception of crowd workers.

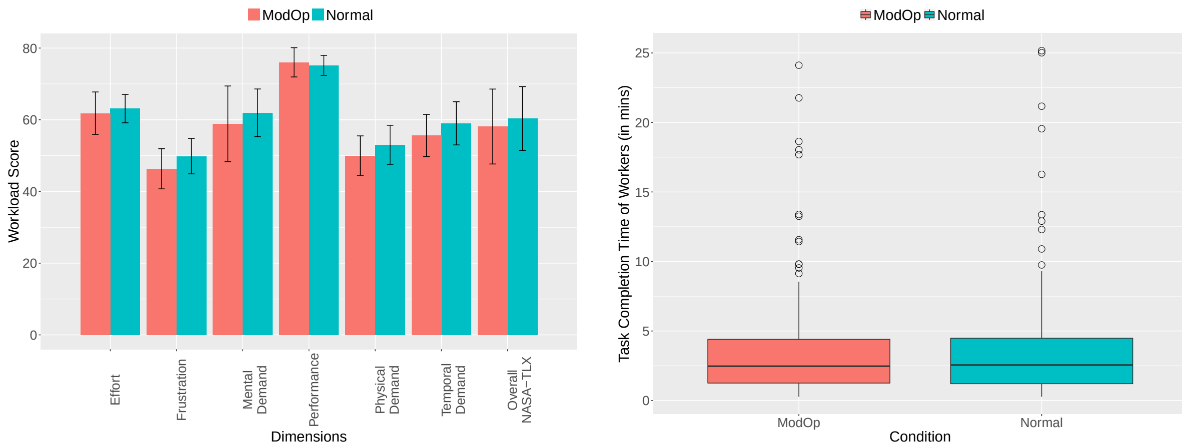
We considered a dataset of 61 tasks that were previously deployed on Amazon's Mechanical Turk, and consisting of different task types [75]. By running the *ModOp* tool on the original task designs, we identified over 20 tasks for which *ModOp* suggested UI design improvements. This accounts for nearly one-third of tasks in the dataset. However, not all modifications suggested by *ModOp* were feasible to implement given the dataset constraints: for example, increasing image resolution is infeasible without possession of high resolution sources. Thus, we consider 9 different tasks where improvements suggested by *ModOp* were feasible to implement. The goals of these tasks included guided learning, interpretation and analysis, image classification and content moderation. In the *Normal* condition, we deployed these tasks on CrowdFlower in their original form and collected 20 judgments from distinct crowd workers. In an identical setup barring the modifications in task design suggested by ModOp, i.e., the *ModOp* condition, we deployed the improved task designs on CrowdFlower. In both conditions, workers were compensated with the same reward level computed based on a fixed hourly rate of 7.5 USD. On completion of each task, workers were asked to complete an online version of the NASA-TLX questionnaire [29] to measure the associated cognitive load of completing the crowdsourcing task.

Figure 5a presents results comparing the cognitive load of tasks measured using the 6 sub-scales of NASA-TLX as well as the overall workload between the two conditions. Through a two-tailed T-test we found that workers in the *ModOp* condition (M=58.11, SD=11.45) perceived a statistically significant lower workload than in the *Normal* condition (M=60.35, SD=9.52); $t(41)$=2.524, $p < .05$. Figure 5b draws a comparison in the task completion time of workers in the two conditions. This suggests that the modifications in task design suggested by *ModOp* can reduce the cognitive load experienced by workers.

We did not find a statistically significant differences in the task completion time for workers across the *Normal* (M=3.86, SD=4.37) and *ModOp* (M=3.77, SD=4.15) conditions at the $p < .05$ level.

Finally, we analyzed the performance of workers in both the *Normal* (M=57.20, SD=33.23) and *ModOp* (M=52.29, SD=36.59) conditions. Figure 6 presents our findings. We did not find a statistically significant difference in the accuracy of workers across the two conditions using a two-tailed T-Test; $t(105)$=0.525, p=.47.

ModOp suggests modifications for all *bad* designs of UI elements. Thus, our findings from Study II with respect to the impact of *bad* and *good* design of UI elements on the accuracy and task completion time (TCT) of workers can be directly interpreted to hold. Contrary to our expectations, in the evaluation of *ModOp* we did not observe significant differences in TCT and the accuracy of workers when compared to the *Normal* condition. On closer inspection, we found that this can be attributed to the relatively few modifications in these tasks (M=3.89, SD=2.57). This suggests that even a few modifications recommended by the tool can reduce the cognitive load of workers, but may not result in a significant improvement in accuracy or TCT in such cases.



(a) Cognitive Load

(b) Task Completion Time

Fig. 5. Cognitive load of workers who completed tasks in the *Normal* condition compared to the *ModOp* condition, measured using the NASA Task Load Index (NASA-TLX), and the corresponding task completion time.

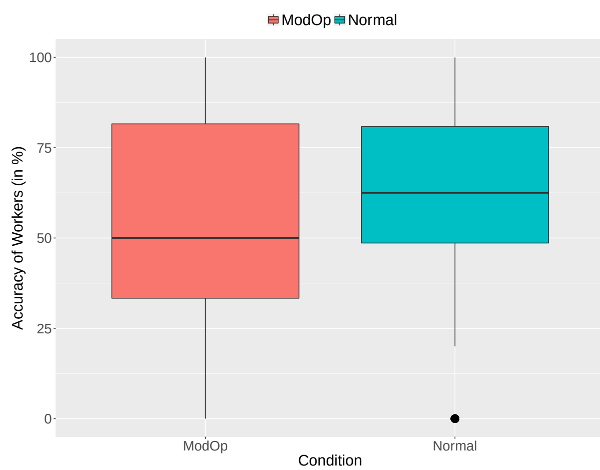


Fig. 6. Performance of Workers in the two Conditions

Thus, based on our experimental results, we can conclude that the *ModOp* tool can be useful in reducing the perceived cognitive load of crowd workers without adversely effecting their task completion times or accuracy.

7 DISCUSSION, CAVEATS AND LIMITATIONS

From the results of our first study, we found that UI elements in crowdsourced microtasks pose several issues to crowd workers. We note that some of the widely cited issues from workers, emerge from violations of classic UI design guidelines highlighted in previous work related to web form design [5, 72]. This indicates that requesters do not take special care to ensure optimal design of UI elements. By accounting for *good* and *bad* design of UI elements in crowdsourcing tasks, we explored the role of crowd work environments in determining the quality of work that is produced.

On analyzing the results from Study-II, we found that on average across the different types of tasks, American workers exhibited lower task completion times than Indian workers. Further scrutiny revealed that American workers were significantly faster at completing tasks with *good* design when compared to those with *bad* design. However, American workers do not exhibit a difference in their overall accuracy between tasks with *good* and *bad* design, as opposed to Indian workers who performed with a significantly higher overall accuracy on tasks with *good* design. This indicates that American workers can better cope with poorly designed tasks.

Our rationale behind restricting the participation of workers in Study II to the highest level of quality on CrowdFlower, was to observe the impact of work environments on the performance of workers who were experienced and genuinely high quality workers. Thus, we analyzed the interplay between UI elements and work environments, and how the interaction shaped quality of crowd work in a real-world setting. Another participation bias may have occurred in Study III towards more experienced workers being willing to participate in individual interviews.

Based on our findings through Study II and Study III, *language proficiency* can potentially influence the task completion time of workers, especially in tasks that they are unfamiliar with. Through the personal interviews carried out in Study III, we revealed a number of aspects (such as device type, device speed) that highlight the role of work environments in shaping the quality of work that is produced.

Finally, with regard to the *ModOp* tool and its evaluation, we believe a preliminary study from the workers' lens in controlled settings was required to establish an affect of *ModOp* on the perception of workers. We propose that selecting random samples from real world microtasks of different types is a good surrogate for an experiment meant to estimate the affect *ModOp* would have on workers in real-world crowdsourcing. The main limitation of this approach is the absence of feedback on the perceived value of *ModOp* from the requesters side. However, it is not easy to recruit a representative or reasonable sample of real-world requesters; the academic background of potential requesters that the authors could exploit in such an evaluation and the resulting selection bias, would have a strong effect on the results. We plan to extend the evaluation to include real-world task requesters in the future.

We believe that our findings will have broad implications on the design and workflow of crowdsourcing microtasks. Considering the hidden role that work environments play in shaping the quality of crowd work, can lead to a fairer treatment of workers, rebalancing the existing power asymmetry between requesters and workers.

8 CONCLUSIONS AND FUTURE WORK

In this paper we studied the effect of work environments (characterized by the hardware and software affordances of crowd workers) on the efficiency and effectiveness of microtask crowdsourcing across two different worker populations. By carrying out three studies, we revealed (i) the most common HIT design challenges crowd workers need to face when completing jobs on microtask crowdsourcing platforms, and (ii) the effect that HIT design choices and work environments have on task execution time, work accuracy, and worker cognitive load. Our findings indicate a significant impact of *good* and *bad* HIT designs for certain task types and across American and Indian crowd workers, (**RQ#1**). We found substantial evidence that confirms the invisible role of work environments in shaping crowd work (**RQ#2**), through experimental findings in Study II and supported by individual interviews in Study III. The findings in this paper reveal the importance of work environments and HIT design on the participation of crowd workers and on the quality of the data collected on microtask crowdsourcing platforms. We encapsulated the important lessons learned through our work into a tool called *ModOp*. The tool helps in validating HIT designs by triggering warnings and providing feedback to requesters who wish to deploy HITs that are congenial (with respect to UI elements), and at the same time are work environment-aware.

This paper touches on studies of user experiences and societal impact, which elaborate on the mission and broader definition of ubiquitous computing. The work offers perspectives into how we can design crowdsourcing tasks to enable broader participation, concerning in particular mobile and device-agnostic design.

As a next step, we plan to further develop *ModOp* to adapt tasks dynamically after detecting the work environment used to complete the task; serving different workers with adapted versions of the same task and personalized it according to their work environments. The tool will adapt multimedia files to support low-bandwidth workers, and help in routing time/quality sensitive jobs to the appropriate workers as determined by their work environments.

REFERENCES

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015-2020. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html. (????). Accessed: 2016-10-01.

[2] Harini Alagarai Sampath, Rajeev Rajeshuni, and Bipin Indurkhya. 2014. Cognitively inspired task design to improve user performance on crowdsourcing platforms. In *CHI*. ACM, 3665–3674.

[3] Javier A Bargas-Avila, Olivia Brenzikofer, Alexandre N Tuch, Sandra P Roth, and Klaus Opwis. 2011. Working towards usable forms on the worldwide web: optimizing multiple selection interface elements. *Advances in Human-Computer Interaction* 2011 (2011), 4.

[4] Javier A Bargas-Avila, Sébastien Orsini, Hannah Piosczyk, Dominic Urwyler, and Klaus Opwis. 2011. Enhancing online forms: Use format specifications for fields with format restrictions to help respondents. *Interacting with Computers* 23, 1 (2011), 33–39.

[5] Javier A Bargas-Avila, AN Tuch, K Opwis, O Brenzikofer, S Orsini, and SP Roth. 2010. *Simple but crucial user interfaces in the world wide web: introducing 20 guidelines for usable web form design.* INTECH Open Access Publisher.

[6] Robin Brewer, Meredith Ringel Morris, and Anne Marie Piper. 2016. "Why Would Anybody Do This?": Understanding Older Adults' Motivations and Challenges in Crowd Work. In *CHI (CHI '16)*. New York, NY, USA, 2246–2257. DOI:http://dx.doi.org/10.1145/2858036.2858198

[7] Jeffrey A Burke, Deborah Estrin, Mark Hansen, Andrew Parker, Nithya Ramanathan, Sasank Reddy, and Mani B Srivastava. 2006. Participatory sensing. *Center for Embedded Network Sensing* (2006).

[8] Dana Chandler and Adam Kapelner. 2013. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization* 90 (2013), 123–133.

[9] Georgios Chatzimilioudis, Andreas Konstantinidis, Christos Laoudias, and Demetrios Zeinalipour-Yazti. 2012. Crowdsourcing with smartphones. *IEEE Internet Computing* 16, 5 (2012), 36–44.

[10] Leah Melani Christian, Don A Dillman, and Jolene D Smyth. 2007. Helping respondents get it right the first time: the influence of words, symbols, and graphics in web surveys. *Public Opinion Quarterly* 71, 1 (2007), 113–125.

[11] Mick P Couper, Michael W Traugott, and Mark J Lamias. 2001. Web survey design and administration. *Public opinion quarterly* 65, 2 (2001), 230–253.

[12] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G Ipeirotis, and Philippe Cudré-Mauroux. 2015. The dynamics of micro-task crowdsourcing: The case of amazon mturk. In *WWW*. ACM, 238–247.

[13] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2013. Pick-a-crowd: Tell Me What You Like, and I'Ll Tell You What to Do. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13)*. ACM, New York, NY, USA, 367–374. DOI:http://dx.doi.org/10.1145/2488388.2488421

[14] Guoru Ding, Jinlong Wang, Qihui Wu, Linyuan Zhang, Yulong Zou, Yu-Dong Yao, and Yingying Chen. 2014. Robust spectrum sensing with crowd sensors. *IEEE Transactions on Communications* 62, 9 (2014), 3129–3143.

[15] Nathan Eagle. 2009. txteagle: Mobile crowdsourcing. In *International Conference on Internationalization, Design and Global Development*. Springer, 447–456.

[16] Carsten Eickhoff and Arjen P de Vries. 2013. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval* 16, 2 (2013), 121–137.

[17] Susan Feinberg and Margaret Murphy. 2000. Applying cognitive load theory to the design of web-based instruction. In *Proceedings of IEEE professional communication society international professional communication conference and Proceedings of the 18th annual ACM international conference on Computer documentation: technology & teamwork*. IEEE Educational Activities Department, 353–360.

[18] Zhenni Feng, Yanmin Zhu, Qian Zhang, Lionel M Ni, and Athanasios V Vasilakos. 2014. TRAC: Truthful auction for location-aware collaborative sensing in mobile crowdsourcing. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*. IEEE, 1231–1239.

[19] Oluwaseyi Feyisetan, Elena Simperl, Max Van Kleek, and Nigel Shadbolt. 2015. Improving paid microtasks through gamification and adaptive furtherance incentives. In *WWW*. ACM, 333–343.

[20] Leah Findlater, Joan Zhang, Jon E Froehlich, and Karyn Moffatt. 2017. Differences in Crowdsourced vs. Lab-based Mobile and Desktop Input Performance Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 6813–6824.

[21] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. 2014. A taxonomy of microtasks on the web. In *Proceedings of the 25th ACM conference on Hypertext and social media*. ACM, 218–223.

[22] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In *CHI*. 1631–1640.

[23] Mary L Gray, Siddharth Suri, Syed Shoaib Ali, and Deepti Kulkarni. 2016. The crowd is a collaborative network. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 134–147.
[24] Aakar Gupta, William Thies, Edward Cutrell, and Ravin Balakrishnan. 2012. mClerk: enabling mobile crowdsourcing in developing regions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1843–1852.
[25] Neha Gupta. 2017. An Ethnographic study of Crowdwork via Amazon Mechanical Turk in India. (2017).
[26] Neha Gupta, David Martin, Benjamin V Hanrahan, and Jacki O'Neill. 2014. Turk-life in India. In *Proceedings of the 18th International Conference on Supporting Group Work*. 1–11.
[27] Tenshi Hara, Thomas Springer, Gerd Bombach, and Alexander Schill. 2013. Decentralised approach for a reusable crowdsourcing platform utilising standard web servers. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. ACM, 1063–1074.
[28] Christopher G. Harris. 2015. The Effects of Pay-to-Quit Incentives on Crowdworker Task Quality. In *CSCW (CSCW '15)*. ACM, New York, NY, USA, 1801–1812. DOI:http://dx.doi.org/10.1145/2675133.2675185
[29] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage Publications Sage CA: Los Angeles, CA, 904–908.
[30] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* (1979), 65–70.
[31] Gary Hsieh and RafałKocielnik. 2016. You Get Who You Pay for: The Impact of Incentives on Participation Bias. In *CSCW (CSCW '16)*. ACM, New York, NY, USA, 823–835. DOI:http://dx.doi.org/10.1145/2818048.2819936
[32] Ting-Hao Kenneth Huang, Amos Azaria, and Jeffrey P Bigham. 2016. InstructableCrowd: Creating IF-THEN Rules via Conversations with the Crowd. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1555–1562.
[33] Panagiotis G Ipeirotis. 2010. Demographics of mechanical turk. (2010).
[34] Lilly Irani. 2015. The cultural work of microwork. *New Media & Society* 17, 5 (2015), 720–739.
[35] Lilly C Irani and M Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 611–620.
[36] Ling Jiang, Christian Wagner, and Bonnie Nardi. 2015. Not Just in it for the Money: A Qualitative Investigation of Workers' Perceived Benefits of Micro-task Crowdsourcing. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on*. IEEE, 773–782.
[37] Matt Jones, Simon Robinson, Jennifer Pearson, Manjiri Joshi, Dani Raju, Charity Chao Mbogo, Sharon Wangari, Anirudha Joshi, Edward Cutrell, and Richard Harper. 2016. Beyond âĂIJyesterdayâĂŹs tomorrowâĂİ: future-focused mobile interaction design by and for emergent users. *Personal and Ubiquitous Computing* (2016), 1–15.
[38] Salil S Kanhere. 2011. Participatory sensing: Crowdsourcing data from mobile smartphones in urban spaces. In *Mobile Data Management (MDM), 2011 12th IEEE International Conference on*, Vol. 2. IEEE, 3–6.
[39] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. 2011. More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk.. In *AMCIS*, Vol. 11. 1–11.
[40] Shashank Khanna, Aishwarya Ratan, James Davis, and William Thies. 2010. Evaluating and improving the usability of Mechanical Turk for low-income workers in India. In *Proceedings of the first ACM symposium on computing for development*. ACM, 12.
[41] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *CHI*. ACM, 453–456.
[42] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *CSCW*. ACM, 1301–1318.
[43] Vassilis Kostakos, Jakob Rogstadius, Denzil Ferreira, Simo Hosio, and Jorge Goncalves. 2017. Human Sensors. In *Participatory Sensing, Opinions and Collective Awareness*. Springer, 69–92.
[44] Gierad Laput, Walter S Lasecki, Jason Wiese, Robert Xiao, Jeffrey P Bigham, and Chris Harrison. 2015. Zensors: Adaptive, rapidly deployable, human-intelligent sensor feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1935–1944.
[45] Matthew Linderman and Jason Fried. 2004. *Defensive Design for the Web: How to improve error messages, help, forms, and other crisis points*. New Riders Publishing.
[46] Jerome P Lynch and Kenneth J Loh. 2006. A summary review of wireless sensors and sensor networks for structural health monitoring. *Shock and Vibration Digest* 38, 2 (2006), 91–130.
[47] Catherine C Marshall and Frank M Shipman. 2013. Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *Web Science*. ACM, 234–243.
[48] David Martin, Benjamin V Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *CSCW*. ACM, 224–235.
[49] David Martin, Jacki O'Neill, Neha Gupta, and Benjamin V. Hanrahan. 2016. Turking in a Global Labour Market. *Computer Supported Cooperative Work (CSCW)* 25, 1 (2016), 39–77. DOI:http://dx.doi.org/10.1007/s10606-015-9241-6
[50] Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. 2016. Taking a HIT: Designing around rejection, mistrust, risk, and workers' experiences in Amazon Mechanical Turk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2271–2282.

[51] Róisín McNaney, Mohammad Othman, Dan Richardson, Paul Dunphy, Telmo Amaral, Nick Miller, Helen Stringer, Patrick Olivier, and John Vines. 2016. Speeching: Mobile Crowdsourced Speech Assessment to Support Self-Monitoring and Management for People with ParkinsonâĂŹs. In *Proceedings of the 2016 ACM SIGCHI Conference on Human Factors in Computing Systems*. 7–12.

[52] Marija Milenkovic and Oliver Amft. 2013. An opportunistic activity-sensing approach to save energy in office buildings. In *Proceedings of the fourth international conference on Future energy systems*. ACM, 247–258.

[53] Meredith Ringel Morris, Jeffrey P Bigham, Robin Brewer, Jonathan Bragg, Anand Kulkarni, Jessie Li, and Saiph Savage. 2017. Subcontracting Microwork. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, To Appear.

[54] Prayag Narula, Philipp Gutheim, David Rolnitzky, Anand Kulkarni, and Bjoern Hartmann. 2011. MobileWorks: A Mobile Crowdsourcing Platform for Workers at the Bottom of the Pyramid. *Human Computation* 11 (2011), 11.

[55] Michael Nebeling, Alexandra To, Anhong Guo, Adrian A de Freitas, Jaime Teevan, Steven P Dow, and Jeffrey P Bigham. 2016. WearWrite: Crowd-assisted writing from smartwatches. In *Proceedings of CHI*.

[56] Sharon Oviatt. 2006. Human-centered design meets cognitive load theory: designing interfaces that help people think. In *Proceedings of the 14th ACM international conference on Multimedia*. ACM, 871–880.

[57] Gabriel Parent and Maxine Eskenazi. 2010. Toward better crowdsourced transcription: Transcription of a year of the let's go bus information system data. In *Spoken Language Technology Workshop (SLT), 2010 IEEE*. IEEE, 312–317.

[58] Sasank Reddy, Andrew Parker, Josh Hyman, Jeff Burke, Deborah Estrin, and Mark Hansen. 2007. Image browsing, processing, and clustering for participatory sensing: lessons from a DietSense prototype. In *Proceedings of the 4th workshop on Embedded networked sensors*. ACM, 13–17.

[59] Markus Rokicki, Sergej Zerr, and Stefan Siersdorfer. 2016. Just in Time: Controlling Temporal Performance in Crowdsourcing Competitions. In *WWW*. 817–827.

[60] Tobias Schnabel, Paul N Bennett, Susan T Dumais, and Thorsten Joachims. 2016. Using shortlists to support decision making and improve recommender system performance. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 987–997.

[61] Oliver S Schneider, Hasti Seifi, Salma Kashani, Matthew Chun, and Karon E MacLean. 2016. HapTurk: Crowdsourcing Affective Ratings of Vibrotactile Icons. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 3248–3260.

[62] Aaron D Shaw, John J Horton, and Daniel L Chen. 2011. Designing incentives for inexpert human raters. In *CSCW*. ACM, 275–284.

[63] Matthias Stevens and Ellie DâĂŹHondt. 2010. Crowdsourcing of pollution data using smartphones. In *Workshop on Ubiquitous Crowdsourcing*.

[64] John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science* 12, 2 (1988), 257–285.

[65] Rannie Teodoro, Pinar Ozturk, Mor Naaman, Winter Mason, and Janne Lindqvist. 2014. The motivations and experiences of the on-demand mobile workforce. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 236–247.

[66] Annamalai Vasantha, Gokula Vijayumar, Jonathan Corney, Nuran Acur Bakir, Andrew Lynn, Ananda Prasanna Jagadeesan, Marisa Smith, and Anupam Agarwal. 2014. Social implications of crowdsourcing in rural scotland. *International Journal of Social Science & Human Behavior Study* 1, 3 (2014), 47–52.

[67] Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. 2014. Community-based Bayesian Aggregation Models for Crowdsourcing. In *WWW*. 155–164. DOI:http://dx.doi.org/10.1145/2566486.2567989

[68] Maja Vukovic, Soundar Kumara, and Ohad Greenshpan. 2010. Ubiquitous crowdsourcing. In *Proceedings of the 12th ACM international conference adjunct papers on Ubiquitous computing-Adjunct*. ACM, 523–526.

[69] Qiuzhen Wang, Sa Yang, Manlu Liu, Zike Cao, and Qingguo Ma. 2014. An eye-tracking study of website complexity from cognitive load perspective. *Decision support systems* 62 (2014), 1–10.

[70] Yuhui Wang and Mohan S Kankanhalli. 2015. Tweeting cameras for event detection. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 1231–1241.

[71] Kathryn Whitenton. 2013. Minimize cognitive load to maximize usability. *Pozyskano* 4 (2013), 2014.

[72] Luke Wroblewski. 2008. *Web form design: filling in the blanks*. Rosenfeld Media.

[73] Xiao-Feng Xie and Zun-Jing Wang. 2015. An empirical study of combining participatory and physical sensing to better understand and improve urban mobility networks. In *Transportation Research Board 94th Annual Meeting*.

[74] Tingxin Yan, Matt Marzilli, Ryan Holmes, Deepak Ganesan, and Mark Corner. 2009. mCrowd: a platform for mobile crowdsourcing. In *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*. ACM, 347–348.

[75] Jie Yang, Judith Redi, Gianluca Demartini, and Alessandro Bozzon. 2016. Modeling Task Complexity in Crowdsourcing. In *HCOMP*.

[76] Andrea Zanella, Nicola Bui, Angelo Castellani, Lorenzo Vangelista, and Michele Zorzi. 2014. Internet of things for smart cities. *IEEE Internet of Things journal* 1, 1 (2014), 22–32.