Mobilizing Crowdwork: A Systematic Assessment of the Mobile Usability of HITs

Senjuti Dutta
sdutta@vols.utk.edu
University of Tennessee, Knoxville
Knoxville, Tennessee, USA

Rhema P. Linder
rlinder@utk.edu
University of Tennessee, Knoxville
Knoxville, Tennessee, USA

Doug Lowe
dlowe13@vols.utk.edu
University of Tennessee, Knoxville
Knoxville, Tennessee, USA

Richard Rosenbalm
rrosenb4@vols.utk.edu
University of Tennessee, Knoxville
Knoxville, Tennessee, USA

Anastasia Kuzminykh
anastasia-kuzminykh@utoronto.ca
University of Toronto
Toronto, Ontario, Canada

Alex C. Williams
acw@utk.edu
University of Tennessee, Knoxville
Knoxville, Tennessee, USA

ABSTRACT

There is a growing interest in extending crowdwork beyond traditional desktop-centric design to include mobile devices (e.g., smartphones). However, mobilizing crowdwork remains significantly tedious due to a lack of understanding about the mobile usability requirements of human intelligence tasks (HITs). We present a taxonomy of characteristics that defines the mobile usability of HITs for smartphone devices. The taxonomy is developed based on findings from a study of three consecutive steps. In Step 1, we establish an initial design of our taxonomy through a targeted literature analysis. In Step 2, we verify and extend the taxonomy through an online survey with Amazon Mechanical Turk crowdworkers. Finally, in Step 3 we demonstrate the taxonomy's utility by applying it to analyze the mobile usability of a dataset of scraped HITs. In this paper, we present the iterative development of the taxonomy, highlighting the observed practices and preferences around mobile crowdwork. We conclude with the implications of our taxonomy for accessibly and ethically mobilizing crowdwork not only within the context of smartphone devices, but beyond them.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in ubiquitous and mobile computing**.

KEYWORDS

Crowdwork, Mobile Usability, Taxonomy, Human Intelligence Tasks.

ACM Reference Format:
Senjuti Dutta, Rhema P. Linder, Doug Lowe, Richard Rosenbalm, Anastasia Kuzminykh, and Alex C. Williams. 2022. Mobilizing Crowdwork: A Systematic Assessment of the Mobile Usability of HITs. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3491102.3501876

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3501876

1 INTRODUCTION

Crowdwork is a contemporary form of on-demand information work that involves the completion of independent tasks of varying complexity, difficulty, knowledge demands, and time constraints. These tasks, often referred to as Human Intelligence Tasks (HITs), are created by *requesters* on crowdsourcing marketplaces and platforms such as Amazon Mechanical Turk[1] (MTurk) in order to find *crowdworkers* who will complete the task in exchange for monetary reward. Research repeatedly suggests that these platforms have become integral as digital professions in the 21st century with "tens of thousands of new workers" arriving on crowdwork platforms each year [18, 51]. Online surveys with residents in the United Kingdom and the European Union suggest crowdwork is widespread in nature, reporting that millions of citizens engage in crowdwork with a substantial percentage mentioning it as their full-time job [40, 41]. Practical examinations of crowdwork suggest that it has not only established itself as a digital profession of the 21st century, but also become an important component in the pipeline of many academic research areas. This makes crowdwork a compelling subject for further research both as a computational tool and a work profession [50, 72, 73].

Studies of crowdwork generally describe the digital profession as one centered around the workstation computer. Quantitative and qualitative studies similarly report that the vast majority of crowdworkers, as in many other information work professions, recognize workstation and laptop computers as their primary device for work-related activities [36, 96]. In crowdwork specifically, much of the utility afforded by workstation or desktop computers stems from their ability to support the nature of on-demand work in which tasks must be captured and completed efficiently. Specific motivations for desktop-centric work practices in crowdwork include screen-size demands [36], limitations of productivity (e.g., HIT finding [96]), and general ease in completing administrative tasks related to crowdwork (e.g., reviewing requesters [77, 87]). A variety of efforts ranging from individuals apps (e.g., Respeak [90]) to full-blown platforms (e.g., Google's *Crowdsource* [12]) independently have facilitated crowdsourced work experiences that are designed to be completed on smartphones. In contrast to these prior contexts, modern crowdsourcing platforms allow requesters to build task interfaces themselves, providing few to no formatting

[1] Amazon Mechanical Turk. https://www.mturk.com/

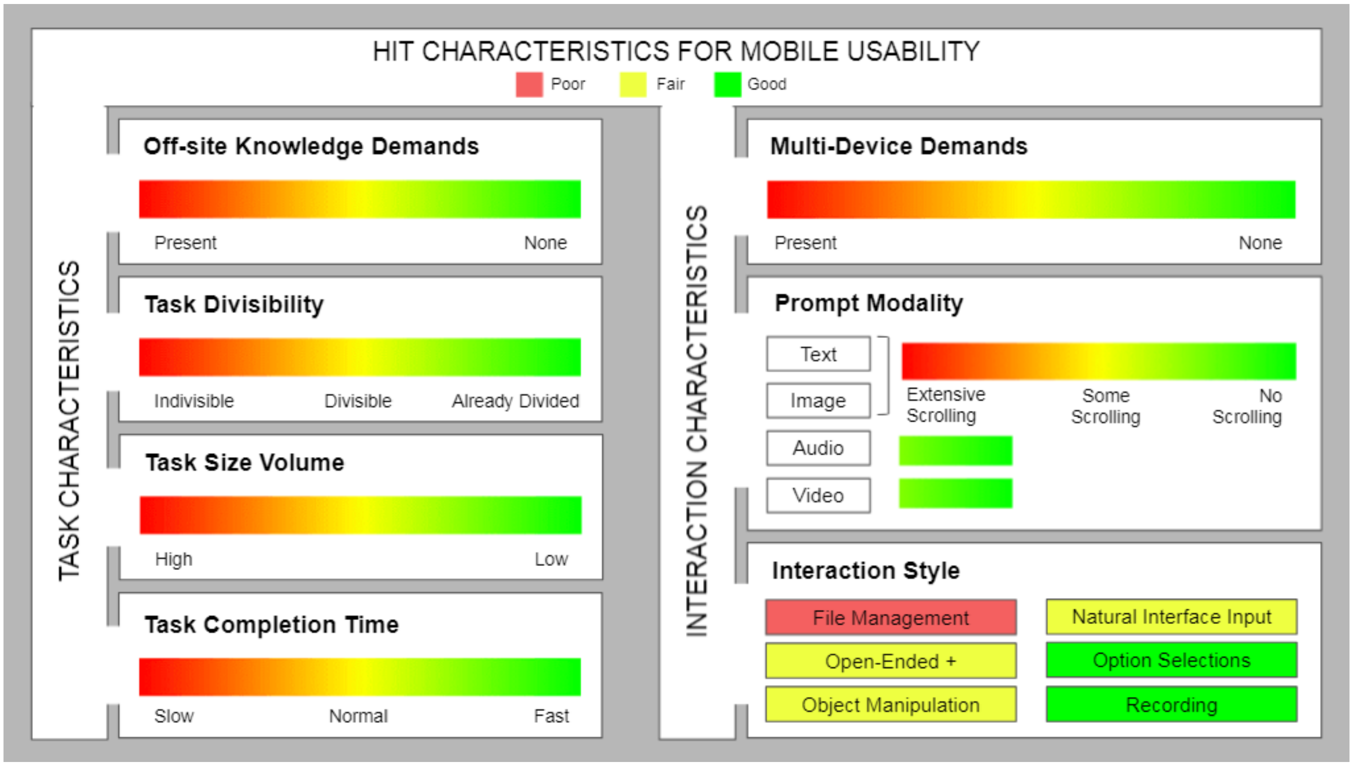


Figure 1: We present our taxonomy on HIT characteristics for mobile usability. We have organized the taxonomy according to Task and Interaction characteristics. The Poor, Fair, and Good legend map to the Mobile Usability Rating (MUR) more defined throughout this paper, with definitions of underlying values in Table 1 and results from an in-the-wild HITs demonstration in Section 5 (e.g. Figure 8).

constraints on way in which a task's question its posed or or the way in which a task's response is collected. Each HIT is structurally unique from the next, and for this reason, it remains unclear how crowdwork, more generally, can be translated to a more mobile context.

The notion of mobilizing crowdwork hinges on understanding the characteristics that shape the usability of a HIT's interface on a mobile device. *Usability* is an aspect of software design that impacts the satisfaction and effectiveness of its users in performing tasks and activities [15]. Usability can be evaluated with design processes from criteria, empirical tests in laboratories, and is contextually valid based on the goals of stakeholders [71]. Prior studies have shown that the usability of mobile devices and the usability of traditional workstations are distinguished by having different constraints, such as a necessity for small screen size and reduced computational power [99]. Therefore, HITs that have high usability – e.g. are usable – for workstations may lack mobile usability. By better characterizing the distinctions that separate mobile and workstation usability in crowdwork, researchers, requesters, and crowdwork platforms will be able to understand pathways for making work more accessible on mobile devices.

Today, both the extent to which work opportunities on crowdwork platforms are effective for workers on their mobile devices (e.g., smartphones) and how to characterize that usability remain unclear. More people on crowdwork platforms own smartphones than have desktops or laptops. A survey by Jacques et al. indicates that around 7% of Mechanical Turk workers use smartphones on HIT because it is the only suitable device they own [45]. In the United States, 85% of adults own a smartphone and 77% own a laptop or desktop computer [10]. Even fewer Americans with an income of less than $30,000 annually have broadband internet at home (57%) and laptop or desktop computers (59%) [92]. Despite a lack of broadband and workstation, 76% of adults in the United States still own smartphone that could potentially participate in crowdwork, if the tasks were usable on mobile devices.

In this paper, we aim to address three research questions toward the goal of mobilizing crowdwork:

(1) RQ1. What characteristics relate to a human intelligence task (HIT) interface being usable on a mobile smartphone?
(2) RQ2. What is the practical and contextual prominence of each characteristic?
(3) RQ3. How do common types of HITs vary in their usability on mobile smartphones?

We first provide background on the role of mobility in work and crowdwork contexts. The next three sections outline a three-step study that details our methods and results for generating a new taxonomy (Figure 1 and Table 1) of characteristics that define mobile

HIT usability for smartphone devices. In Step 1, we first contextualize our work by using a workshop to generate and discuss characteristics in relation to relevant literature. Then, we apply this derived knowledge towards accumulated HIT suitability research which is focused towards the taxonomy of characteristics that relates to mobile suitability. In Step 2, we issue and analyze an online survey that provides empirical support to the developed taxonomy with added nuance from the perspective of Mechanical Turk workers. Finally, in Step 3, we use the taxonomy to evaluate the mobile usability of 519 HITs on Amazon Mechanical Turk. The results from our study with three steps open a discussion on the current state of mobilization of crowdwork and highlights the characteristics relevant to usability in mobile crowdwork. We conclude by discussing the implications of our presented taxonomy as it relates to making crowdwork more ethical, more accessible, and more mobile-friendly.

2 BACKGROUND WORK IN THE ROLE OF MOBILITY

The purpose of our study is to assess and understand the characteristics that relate to the mobile usability of HITs. Here, we describe the relevant background work related to the goal of understanding the role of mobility in crowdwork.

2.1 Professional Crowdwork: A Desktop-Centric Practice

Crowdwork is an emergent work practice that centers around the completion of tasks for payment. In the terminology of Amazon Mechanical Turk, *requesters* issue HITs and pay *workers* a reward in currency. HITs are often relatively small, routinely involving only seconds of work [19]. Amazon Mechanical Turk is a popular platform for a large range of industry and academic researchers because it provides a convenient way to access a large and diverse set of participants and workers [79]. HITs are often regarded as tasks that are arbitrary in complexity, difficulty, knowledge demands, and time constraints. Today, crowdwork marketplaces, such as Amazon Mechanical Turk, Prolific, and Clickworkers, allow requesters to create HITs and outsource their completion to crowdworkers in exchange for monetary compensation [50]. Alongside the completion of HITs, crowdworkers engage in a range of supplemental tasks that support their work, including finding HITs, communicating with other crowdworkers, and reviewing requesters [27, 58].

Unlike traditional information work contexts, much of crowdwork is fueled by voluntary or community-based efforts in which crowdworkers self-manage infrastructure to improve their work, such as hosting platforms for reviewing HITs or requesters (e.g., Turkopticon [44]), creating forums to connect with other crowdworkers (e.g., TurkerView [77]), or building or sharing new productivity tools [96]. Generally speaking, the workstation computer is the intended use-case for these tools, though there exists some evidence that suggests a broader desire for their use on mobile phones [96]. It is generally assumed that crowdworkers utilize, or have access to, a workstation computer for their work due to the multitude of capabilities that these machines maintain over there mobile. Recent studies, however, report that challenges and barriers to the work practice are nonuniform, particularly in rural areas where assumptions around the availability of specific devices or personal expertise may be biased [23, 30]. In general, studies suggest that the desktop computer is the predominant gateway to many of the opportunities that exist in crowdwork today.

2.2 Mobile Usability and the Mobilization of Crowdwork

Understanding, measuring, and characterizing the usability of task interfaces on mobile devices is a long-standing research problem that spans decades of research [8, 78]. Through these studies, researchers often target a specific domain and subsequently define and validate a set of "variables" that impact usability with the domain [86]. For example, Agarwal and Venkatesh provided five categories of variables for assessing website usability (i.e., content, ease of use, promotion, made-for-the-medium, and emotion) with the intent of using these variables as usability heuristics [1]. Serving as a demonstration to the development of the categories, Venkatesh et al. used Agarwal et al's heuristics as a tool to assess the usability and develop automated predictions about its use [91]. Meta-analyses of empirical studies show that mobile usability can be assessed through the lens of tasks, technology, environments, and even individual user characteristics [14]. Importantly, these approaches often involve a range of measures that range from traditional HCI measures for usability (e.g., cognitive effort, usefulness) to more practical measures of usability for design research (e.g., responsiveness, aesthetics). Generally speaking, the ideal mechanism for defining, measuring, and assessing the mobile usability of an interface is one that captures the interplay of relevant factors within a specific context [46].

Across decades of research, studies have demonstrated that mobile crowdsourcing can be achieved in a variety of ways with mixed usability outcomes. TxtEagle [21] and mClerk [29] were respectively deployed in Kenya and India and allowed interested individuals to complete several types of tasks via SMS text messages, including language translation, market research, audio transcription, and low-cost image classification. Narula et al. developed and studied *MobileWorks*, a smartphone-based crowdsourcing platform that administers optical character recognition tasks [62]. Yan et al. introduced an iOS application to post and submit sensor-related crowdsourcing tasks using smartphones [98]. Vashistha et al. demonstrated how voice-based interactions on mobile could be facilitated with Respeak, a voice-driven, crowd-powered system for assisted transcription on mobile devices [90].

More recent research in mobile usability has emphasized the increasingly multi-device landscape of crowdwork. Hettiachchi et al. found that crowdworkers are receptive toward vocally engaging with crowdsourcing tasks administered through smart speakers [35]. In conjunction with new possibilities for device-specific work, researchers have sought to understand the challenges that stem from completing a particular type of task on a particular device [17]. In a recent study of task acceptance, preferences for six different types of common HITs (e.g, sentiment analysis, information finding, audio tagging, speech transcription, image classification, bounding box) across different devices, Hettiachchi et al. not only observed that the task acceptance rate varies between devices, but also found that the acceptance rate between desktop and smartphone devices was substantially comparable [36]. Beyond task-level considerations, studies suggest that crowdworkers have additional uses for their mobile devices that extend beyond simply completing HITs on the smartphone (e.g., being notified about new HIT opportunities) [96].

Though *"mobile usability"* remains relatively undefined within the niche domain of crowdwork, researchers have used mobile platforms (e.g., smartphones, tablets) to create crowdsourcing experiences that embody aspects of "good" mobile usability. For example, Pei-Yu et al. examined the motivations for participating in mobile crowdsourcing tasks provided by Google's "CrowdSource", a crowdsourcing platform that engages people with a limited set of mobile task types and rewards them with badges [12]. Similar approaches have been leveraged in citizen science in which the general public can use a freely-available mobile app to complete a myriad of question-oriented tasks (e.g., "Is there a galaxy in this image?") [34]. As their intended audience is often individuals who are intrinsically motivated (e.g., an interest in furthering science), many of these tasks are designed to have the combined goal of minimizing cognitive load and making the act of contributing enjoyable [75]. Though several platforms (e.g, Zooniverse [81]) facilitate effective mobile experiences, they recognize the design of usable interactions for mobile devices as an on-going challenge [69].

2.3 Beyond Crowdwork: Mobilizing Information Work with Microproductivity

Microwork, like crowdwork, utilizes a *microtasks* paradigm that breaks larger tasks down into manageable chunks. Human-computer interaction research utilized a variety of HIT characteristics to motivate the design and functionality of novel systems (e.g., Soylent [6]), but were later inspired by the idea of microtasks to apply its principles to information work more broadly. Teevan et al. distinguishes microtasks as distinct from HITs because they *always* include the necessary context to complete and *always* require only several seconds of attention [84]. Further, microtasks are often automatically generated by an algorithm rather than by hand. Such algorithms operate by algorithmically decomposing a larger task (i.e., a macrotask) into a series of smaller microtasks that can be quickly and easily answered (e.g., a binary Yes-No question) [11]. For this reason, many HITs exceed the academic definition of what may be considered a microtask. In an analysis of 130 million HITs collected in 2014, Difallah et al. found that that batches of microtasks are becoming increasingly more common on Amazon Mechanical Turk [19]. Despite these technical differences, "crowdwork" and "microwork" are often used interchangeably both by researchers and practitioners.

Research in HCI and ubiquitous computing has given ample attention to designing new interactive systems to better facilitate mobile work experiences through the lens of microwork. Apparition helps designers in rapidly creating functional interface prototypes from sketches and verbal descriptions using any kind of device with a web browser [53]. MicroWriter supports mobile phones along with laptops and desktops and subdivides larger writings into smaller manageable microtasks to improve productivity [85]. PlayWrite leverages a similar model of microtasking on the smartphone to facilitate productive writing during limited attention scenarios [43]. Mercury utilizes a custom model of function-based microtasking to allow programmers to continue their work while on-the-go [95]. Finally, "Slide to X" employs a unique microtask design to facilitate a new, low-effort mechanism for unlocking smartphones [88]. Beyond the smartphone, research has also explored microtask completion on even further constrained devices (e.g., smartwatches). These studies generally conclude that completing these tasks is feasible, yet many tasks lack the appropriate design or structure for such completion [65]. Many of these systems are systematically similar in that as each provide interactive experiences that map to the constraints of the device and assume attention is either divided or significantly limited.

2.4 Summary of Contribution

There is a growing interest in extending crowdwork beyond the traditional desktop-centric practice of professional crowdwork. However, researchers, requesters, and platforms, today, fail to understand the mobile usability requirements of human intelligence tasks. Inspired by prior studies that develop characterizations of usability [14], we employ a three-step approach toward the development, empirical extension, and demonstration of a new taxonomy of mobile usability characteristics contextualized to human intelligence tasks. Our contributions include the taxonomy of characteristics, a brief examination of current and desired mobile HIT preferences, and a mobile usability assessment of HITs collected from Amazon Mechanical Turk. Collectively, we contribute an understanding of the task and interaction characteristics of HITs that make them suitable for completion on the smartphone devices.

3 TAXONOMY GENERATION - STEP 1: TARGETED LITERATURE ANALYSIS

The goal of our research is to understand the characteristics of HITs that contribute to their usability on mobile smartphone devices. To better understand these characteristics, we designed and employed a process heavily inspired by the Nominal Group Technique [16] that allow us to generate a taxonomy through brainstorming, ideation, discussion, and deliberation. In this section, we describe this process and conclude by presenting the final taxonomy of characteristics.

3.1 Procedure

We designed a procedure that centers around the Nominal Group Technique [16], a group-based decision making workshop method that aims to facilitate the generation of ideas from a small group of qualified experts or professionals. Our research team includes experts in crowdsourcing that have acted as requesters and workers and have a background in HCI. This justifies our process for proposing areas of interest for mobile usability. The technique's four-step process involves 1) generating ideas, 2) recording ideas, 3) discussing ideas, and 4) deliberating ideas. Methodological analyses of the Nominal Group Technique (NGT) specifically reinforce its use as a tool for computer-mediated ideation and brainstorming [22]. The procedure has been used in several contexts within human-computer interaction settings, including summaries of workshop activities [32], understanding system requirements [20], and improving computer-mediated decision-making [52]. Here, we leverage this technique to facilitate group-based decisions about the development of our taxonomy while drawing inspiration from related work.

3.1.1 Overview of Procedural Execution. Three members of the research team participated in the procedure. Two researchers currently hold a PhD in Computer Science while the third researcher is a PhD student in Computer Science. All three researchers have significant expertise and published research papers in crowdsourcing and crowdwork.

Before beginning the NGT procedure, each member of the research team independently conducted an review of literature toward the goal of searching for issues or concerns related not only to completing HITs on mobile devices, but also general usability related to mobile devices. Specific guidelines, suggestions, or taxonomies for designing HITs for mobile are not discovered during this preparatory phase of research. With the familiar literature in mind, our research goal shifted to summarizing research as characteristics, rather than to discovering or creating novel conceptualizations of what contributed to mobile usability on smartphone devices. In support of this goal, we engaged in a four-step, NGT-inspired process:

- **Stage 1. Idea Generation.** Our procedure began by assigning one researcher as the "moderator" who was responsible for facilitating the experience. The moderator asked each researcher, including themselves, to work silently and independently toward the goal of answering the following question:

  "What characteristics relate to a human intelligence task being completed on a mobile smartphone?"

  The overarching goal of beginning independently stems directly from the Nominal Group Technique which suggests that the quality and quantity of ideas that are produced [33, 60].
- **Stage 2. Idea Recording.** The moderator provided each researcher with a document (i.e., a Google Doc) in which they were required to record characteristics related to HIT mobility. Researchers were encouraged to provide a list of characteristics they perceived to be comprehensive and motivated by at least one prior work. Upon completion, researchers submitted the completed document to the moderator.
- **Stage 3. Idea Discussion.** After receiving each of their completed documents, the moderator organized a videoconferencing meeting in which they instructed each researcher to share and discuss the contents of their document. The primary goal of this stage was to ensure that members of the research team were given an opportunity to convey the importance of their documented characteristics. Each researcher took turns providing and receiving feedback, asking for clarifications, and producing new inquiries based on the discussion.
- **Stage 4. Idea Deliberation.** Following the presentation of ideas, the research team engaged in a deliberative process toward the goal of arriving at a consensus of ideas, which was used as the basis for developing the taxonomy of characteristics. As the research team developed more focused research questions, heuristics for including or excluding characteristics from the taxonomy emerged naturally. The inclusion of a specific characteristic was primarily motivated by referencing prior literature that supported its relevance toward answering the question defined in Stage 1. In contrast, a characteristic was excluded from consideration if one of the following conditions held: (1) related to a device, rather than an aspect of a HIT that could be configured by a requester, (2) related to aspects of "meta-work", such as finding or managing HITs, or (3) a characteristic that could not be easily assessed through visual or manual inspection (e.g., in a screenshot of the HIT). The research team concluded Stage 4 by arriving at a consensus set of ideas that underscore the mobile usability requirements of HITs. This included assigning a Mobile Usability Rating (MUR) of Good, Fair, or Poor for each characteristic's value per related work.

An important consideration for the design of this study is that our research team was required to operate in a remote and distributed fashion due to the on-going COVID-19 pandemic. While prior studies have utilized NGT to facilitate real-time brainstorming sessions that often take place in-person (e.g., [32]), our procedure took place entirely in a computer-mediated fashion using email communication and videoconferencing software (e.g., Zoom).

3.2 Results: A Literature-Fortified Taxonomy of Mobile Characteristics

The first three stages of the NGT-inspired procedure resulted in a set of initial "ideas" for summarizing characteristics from literature. For clarity, we henceforth refer to the generated "ideas" simply as our "taxonomy of characteristics". Through the Idea Generation and Recording phases (*Stage 1 and 2*), a total of 20 non-unique characteristics were developed independently and suggested to the moderating researcher. The number of characteristics contributed by each researcher ranged from three to 11. Each characteristic was presented during the Idea Discussion phase (*Stage 3*). During this phase, a total of 8 suggested characteristics were identified as duplicates (i.e., reported by more than one researcher), leaving a total of 12 unique candidate characteristics for consideration in the taxonomy. During the Idea Deliberation phase (*Stage 4*), we developed heuristics for focusing characteristics as practical for assessments. This refined and modified how our summaries of characteristics in literature as attributes of these characteristics were more deeply and narrowly defined. The process of narrowing was guided by literature references that we subsequently used to fortify our characteristics with practical justification.

The NGT procedure concluded with a set of 7 characteristics of HITs that relate to their suitability for completion on mobile devices. Our taxonomy of characteristics is divided between two types: (1) *Task Characteristics* and (2) *Interaction Characteristics.* All possible values for each characteristic are driven by examples that arose from prior literature discussed throughout the later stages of the NGT procedure. Each possible value is mapped to a MUR value of "Good", "Fair", and "Poor" usability based on hypothetical contexts that arise through discussion as well as observed contexts reported

	Characteristic Name	MUR	Values	Description	Refs
Task Characteristics	Off-Site Knowledge Demands	🟩	None	HITs that do not require information beyond the task interface.	[7, 26, 100]
		🟥	Present	HITs that require information beyond the task interface.	[7, 26, 100]
	Task Divisibility	🟩	Already Divided	HITs that are already divided into small tasks.	[93]
		🟨	Divisible	HITs that have clear boundaries for division into smaller tasks.	[37]
		🟥	Indivisible	HITs that lack clear boundaries for division into smaller tasks.	[97]
	Task Size Volume	🟩	Low	HITs with content that includes two pieces of media or less.	[53]
		🟥	High	HITs with content that includes more than two pieces of media.	[25]
	Task Completion Time	🟩	Fast	HITs that require 15 seconds or less to complete.	[5]
		🟨	Normal	HITs that require 15 to 60 seconds to complete.	[54]
		🟥	Slow	HITs that require more than 60 seconds to complete.	[42]
Interaction Characteristics	Multi-Device Demands	🟩	None	HITs that do not require the use of multiple devices to complete.	[36, 74]
		🟥	Present	HITs that require the use of multiple devices to complete.	[36, 74]
	Prompt Modality	🟩	Video	HITs that involve annotating and/or understanding video.	[2]
		🟩	Audio	HITs that involve annotating and/or understanding audio.	[2]
		🟨	Image*	HITs that involve annotating and/or understanding images.	[2]
		🟨	Text*	HITs that require annotation and/or understanding text.	[2]
	Interaction Style	🟩	Option Selections	HITs that require selecting a set of options (e.g. radio button).	[82]
		🟩	Recording	HITs involving audio and video authoring.	[59]
		🟨	Open-ended+	HITs that require using fill-in-the-blanks or free-form text-entry.	[82]
		🟨	Object Manipulation	HITs that involve direct manipulation (e.g., bounding box).	[82]
		🟨	Natural Interface Input	HITs that involve unconventional input (e.g., gestures)	[82]
		🟥	File Management	HITs that require manipulating or uploading files.	[49]

Table 1: The taxonomy of characteristics and their associated values. Each characteristic was rated as contributing to a particular Mobile Usability Rating (MUR): 🟩 Good, 🟨 Fair, or 🟥 Poor. A * indicates that MUR is heavily context-dependent.

in the literature. Importantly, each characteristic in the taxonomy is intended to assess a particular aspect of a HIT's suitability for use on smartphone devices. To assess the overall mobile HIT usability, each of these characteristics can be considered together. The set of characteristics and their associated values are shown in Table 1.

3.2.1 *Task Characteristics.*

We identified a total of four characteristics that relate to a HIT's task design (i.e., the structural representation of task-related information). Task design is well-studied aspect of crowdsourcing that is concerned with the *efficiency* of crowdsourced tasks [55]. We draw on the following characteristics for our taxonomy:

(1) *Task Completion Time* describes the amount of time required to complete a HIT. Prior work suggests that crowdworkers and information workers alike have an interest in using smartphones only briefly for activities that can be completed quickly [2, 42, 95, 96]. The assumption is that faster tasks are preferred for mobile contexts.
(2) *Task Divisibility* refers to the notion that a HIT can be broken down into smaller subtasks. Prior studies suggest that translating macrotasks to microtask counterparts increases the usability of these tasks on mobile devices [43, 95]. Similar characteristics had been suggested and discussed during Stages 2 and 3 (e.g., *Task Size Steps*), but were eliminated due to the difficulty associated from assessing its presence (e.g., from a screenshot).
(3) *Task Size Volume* describes the information content within a HIT's task interface. Previous research indicates that larger task size volume negatively impacts crowdworkers' productivity [57].
(4) *Off-Site Knowledge Demands* characterizes the underlying need to navigate away from a HIT's primary task interface (e.g., to find information on another webpage or make use of another web resource) in order to successfully complete it [7, 26, 100], an activity that is generally recognized as inefficient on mobile devices [76].

3.2.2 *Interaction Characteristics.*

Alongside our four *Task Characteristics*, we identified a total of three characteristics that relate to a HIT's *interaction design* (i.e., the ways in which a task is posed and a crowdworker must complete it). The specific characteristics include:

(1) *Multi-Device Demands* describes a HIT's underlying need for the use of multiple devices in order to complete it. Several recent studies on multi-device experiences in crowdwork suggest that the use of multiple devices is becoming increasingly more common [36].
(2) *Prompt Modality* refers to the type of media that crowdworkers are prompted with and required to interface with (e.g., annotate, classify, etc). We draw directly from Peng et al. [70] to motive this characteristic's inclusion as it highlights four types of prompts.
(3) *Interaction Style* details how a HIT requires crowdworkers to engage with it, whether it be through free-form text to complex natural language. We, again, chose to simplify this characteristic toward the goal of observing interaction styles that can be assessed visually (e.g., with a screenshot) [82].

Thus, we have addressed RQ1 by developing characteristics in our taxonomy that are relevant to mobile HIT usability. We group the relevant usability characteristics in two sets *Task Characteristics* and *Interaction Characteristics* in Table 1.

4 TAXONOMY SUPPORT - STEP 2: MECHANICAL TURK SUPPORT SURVEY

Observations from our NGT-inspired study provided us with a taxonomy that describes the characteristics that contribute to a HIT's usability for completion on mobile devices. To better understand whether these characteristics capture the mobile usability requirements of HITs in reality, we draw from data collected from an online survey aimed at assessing mobile crowdwork through the lens of various mobile devices (e.g., smartphones, smartwatches, smart speakers).

4.1 Method: Online Survey

The original survey design was motivated by a broader research project aimed to understand the challenges and opportunities of engaging with crowdwork on mobile devices. The IRB approved survey includes 43 questions across five sections and is available as supplemental material[2]. Though the research questions for this study were not specifically related to understanding the mobile usability requirements of HITs, several survey questions focused specifically on understanding the types of HITs that are suitable for mobile devices. We therefore conducted a targeted analysis of relevant questions to support the outlined taxonomy.

4.1.1 Survey Design. The survey began by inquiring about participants' personal demographics (i.e., age, gender, education) and their work experience on Amazon Mechanical Turk (e.g., completed HITs, current work hours, HIT approval rate). Thereafter, the survey was split into four sections collectively aimed at understanding current and desired mobile work practices in crowdwork. Our analysis specifically draws on three questions from two sections of the survey:

1. *Section 2. Understanding HIT Completion.* This section includes multiple-choice and open-ended questions about the types of HITs that crowdworkers both (1) currently try to complete on mobile devices and (2) would like to see better supported on mobile devices. It also includes multiple-choice and open-ended questions about the frequency and scenarios that crowdworkers utilize devices to complete HITs on MTurk. From this section, we specifically analyze responses to the following questions:
   - Q15.1. *Please briefly describe what types of HITs you currently try to complete on your smartphone.*
   - Q17.2. *For your MTurk work, what types of HITs would you like to see better supported on the smartphone?*
2. *Section 5. The Magic Wand.* This section asks participants to consider a scenario in which they have a magic wand that allows them "to change whatever you'd like to change about work on Amazon Mechanical Turk to work on the platform how you want to work". From this section, we analyze responses to the following question:
   - Q24.2. *How would you use the magic wand to make managing and performing HITs better for a smartphone? What would you change? Why?*

4.1.2 Recruitment and Remuneration. We recruited a total of 111 participants for the study by deploying the survey to MTurk. To ensure participants' data was both reliable and motivated by experience with crowdwork, we employed a HIT qualification that required participants to have completed at least 10,000 HITs and have an a minimum acceptance rate of 98.0%. Prior research has found that crowdworkers on Amazon Mechanical Turk may multitask when payment is too low. We therefore chose to reward participants with $5.00 USD as it both ensures they are paid fairly and feel more comfortable devoting their complete attention to our HIT [96]. One participant reported that they do not own a smartphone, and six participants demonstrated spamming behavior in their survey responses. We chose to remove these seven participants, thus limiting our analysis to 104 participants.

4.1.3 Analysis Methods. All three survey questions relevant to the study at hand collected open-ended responses from participants. We conducted top-down coding analysis of two responses (Q15.1 and 17.1), in which responses are categorized into pre-existing codes. We coded the third question (Q24.2) with a bottom-up approach, creating codes to find themes [9].

To analyze the responses responses to questions Q15.1 and 17.1, we mapped participant responses to task types identified in our prior studies related to cross-device crowdwork [3, 17, 24, 31, 36, 89]. Specific HIT type labels include *Content Generation* [3, 17, 24, 31], *Image Classification* [3, 17, 36], *Image Transcription* [17, 36], *Information Finding* [24, 36], *Qualification* [3, 24], *Survey* [24, 31] and *Text Classification* [24, 36]. An "Other" label was added to accommodate task scenarios that fail to fit within this label paradigm. The labeling process was conducted by two annotators, and inter-rater reliability was determined to be substantially high for Q15.1 (κ = 0.8) and for Q17.1 (κ = 0.7) [94].

In contrast to Q15.1 and Q17.1, Q24.2 is more open-ended such that it allows participants to provide responses that are not necessarily limited to, but may include HIT characteristics. We therefore chose to conduct bottom-up coding process in which themes were developed through standard open-coding. Our underlying intent is to provide an unbiased mechanism for capturing a wealth of characteristics to naturally observe how participants gravitate toward characteristics described in our taxonomy instead of other characteristics that may be relevant (e.g., device constraints). Two annotators engaged in thematic labeling, again, with substantial reliability (κ = 0.8) [94].

4.2 Findings

Demographic information about our participants suggest that they have substantial experience in working on MTurk. Participants identified as male or female near-equally (M=55;F=47;NB=2). 47 participants (45.2%) held at least a Bachelor's degree. In terms of work experience on MTurk, 78 (73.5%) of participants identified as having worked on the platform for 2 or more years. 35 participants (33.7%) stated that they work 10 to 20 hours per week on the platform with a slightly smaller report for the 23 participants (22.1%)

[2] Please see the Supplemental Material section of Precision Conference.

who work 30 or more hours per week. The median of total HITs was 26,500 (σ=135,519), and the median approval rating was 99.58% (σ=43.73).

Through our analysis, we observe that that more than half of the survey respondents currently use their phone to complete HITs. Further, we also find that an even larger percentage of participants have explicit ideas for improving mobile HIT usability. 59 participants' responses (57%) made explicit reference to currently completing either at least one of the task types that we described in Section 4.1.3 or HITs that have a particular characteristic that makes them suitable. In contrast to their current practice, 89 participants (86%) provided a response to Q17.1 or Q24.2 that outlined a particular way in which the mobile HIT usability could be improved on smartphones. Across the remaining responses, we observe three high-level themes for using the "magic wand" change how they manage or complete HITs: (1) *Design and Compatibility*, (2) *Task Interaction*, and (3) *Tools, Scripts, and Apps.* Despite being viewed distinctly, each of these themes' responses collectively work toward the goal of improving crowdworkers' efficiency and productivity. The remaining 15 participants (14%) stated explicitly that they do not have any desire to manage or complete HITs on their smartphone. The distribution of responses across these themes is shown in Figure 2.

We now present our observations made through the lens of these themes. We draw specific attention to understanding how the characteristics of our taxonomy surface through participant responses. We conclude the presentation of our findings by connecting observations to the types of tasks that are currently practiced by participants alongside those that they believe should be better supported.

4.2.1 Theme 1: Design and Compatibility.

As reported by 34 participants (33%), one of the most prominent theme that emerged from our analysis of responses was centered around the resolution of *Design and Compatibility* issues that exist when accessing HITs on smartphones. The general sentiment of these responses was that HITs are designed under the assumption that they be completed on a desktop computer and often fail to render correctly on mobile devices:

> *"Maybe make HITS that just work on that small of a screen. I do other surveys on other platforms on my phone, and they always look better than any of the MTurk one that require a phone." (P38)*

By rendering incorrectly on the phone, the design of the HIT introduces new barriers that require additional effort to complete on the smartphone devices. For example, "Penny HITs" are a common type of HIT that are already divided, are limited in size, and can be completed in less than a few seconds on a desktop computer. Penny HITs often involve snap-judgements about a specific type of *Prompt Modality* with a binary question (e.g., "Is there a cat in this image?"). Even in the case where a task may be divided and design for efficiency, its associated media (e.g., a large image) may appear differently on smartphones, which lead to a hindrance in usability:

> *"I would make it so tasks are easier to see and do on the smartphone. Some of them are not made for smartphone use. So, they end up looking weird and not sized correctly. I would make it easier to do quick penny hits, so you can move through them at a quicker speed." (P79)*

A particular aspect of unresponsive HIT interfaces is that the occlusion of other relevant information on the task interface page is common. Such issues were also mentioned in the case of survey HITs in which P8 described their frustration with the navigational demands that arise through HITs that are not well-designed for use on the smartphone:

> *"[Questions in surveys should have] correct sizing so you don't need to scroll all over with Qualtrics surveys. It's annoying and inefficient."* (P8)

Alongside navigational demands that occur within the limits of a task interface, we find that navigational demands beyond it were voiced as well. Two participants explicitly mentioned challenges related to *Off-Site Knowledge Demands* highlighting the need to "make it easier to switch between browser tabs." (P45) and more generally switch between application windows on the phone:

> *"I worry about going from the MTurk page to the survey page and losing my work. So, I would want to change the process and be able to stay on the same page instead of a separate link to go to." (P45)*

A small number of participants suggested resolving *Design and Compatibility* issues, such "an auto-reformatting feature that formats HIT pages to better display on tiny phone screens" (P70) or "requiring requesters to design their projects for both [desktop and mobile] platforms" (P6). From the perspective of crowdworkers' the ideal composition of these characteristics for mobile HIT usability is one that makes "it easier to go through questions without stopping" (P25). The vast majority of *Design and Compatibility* issues suggest that mobile efficiency is stunted by tasks that are not well-divided, have context that exceeds the smartphone screen, have prompts that are not supported across devices, and are generally slow to complete in comparison to their desktop counterparts. Each of these confirm the representation of the *Off-Site Knowledge Demands*, *Task Divisibility*, *Task Size Volume*, *Task Completion Time*, and *Prompt Modality* characteristics within our taxonomy.

4.2.2 Theme 2: Task Interaction.

Elements of *Task Interaction* were discussed by 20 participants (19%). Responses within this theme centered around interactive challenges that occur within task interfaces on mobile devices. Prior studies have identified a plethora of efficiency challenges for touch-based smartphones [4, 56]. Three participants voiced explicit remarks around touch-based interaction, citing that they would use the magic wand to "just improve interfaces to take better advantage of touch screens" (P63). As P29 states:

> *"I'd make it more realistic to do certain tasks on smartphone, such as a bounding-box tasks, i.e. make it support touch-screen devices."* (P29)

At a high-level, we observe that the limitations in interactivity that stem from touch-based input contribute significantly to our participants' lack of interest in using their smartphone to complete HITs. Despite not explicitly mentioning the smartphone's touch-based input, the remaining 17 participants provided remarks that highlight how mobile activities are stifled by the speed of mobile

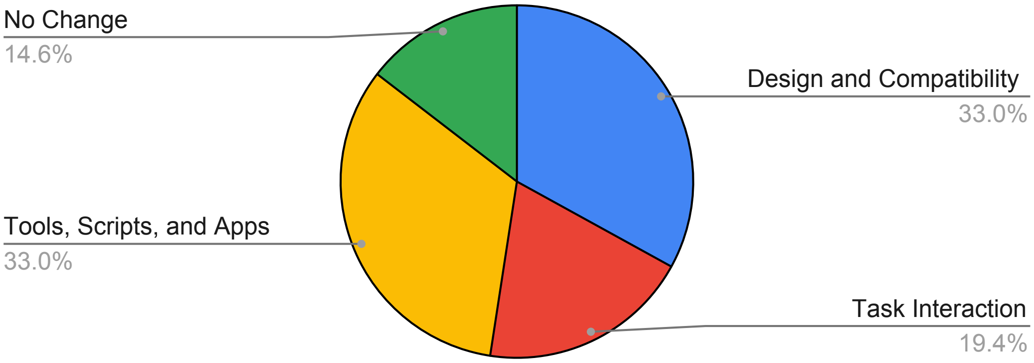


Figure 2: Representation of themes in responses to Q24.2.

interactions. For example, 11 of the 20 participants experienced difficulty when typing with a virtual keyboard on a smartphone:

> *"There is no mouse and keyboard. Everything is slow. It needs to be faster."* (P43)

Seven participants made explicit remarks about the limitations that stem from a lack of traditional input devices. Five of the seven participants further elaborated by explicitly suggesting that the ability to attach a mouse and keyboard to their smartphone would substantially make HITs more usable on the smartphone. As a crowdworker who does not currently complete HITs on their phone, P12 stated:

> *"Perhaps mouse and keyboard support, but this would feel weird. At that point, I guess I may be able to work on a smartphone if forced."* (P12)

Alongside concerns with general efficiency within task interactions, one participant voiced a desire to improve how information is transferred across devices, citing that it was *"hard to copy-and-paste things such as the survey code on a smartphone."* (P5). Across this theme, we specifically observe the presence of concerns and recommendations that span elements of *Interaction Style*, *Prompt Modality*, and *Multi-Device Demands*. Through the lens of *Task Interaction*, our specific observations are that "good" mobile usability is best achieved when task interaction is limited to multiple choice, text entry is not required, and multi-device demands are not present.

4.2.3 Theme 3: Tools, Apps, and Scripts. As reported by 34 participants (33%), the final theme of responses collectively referenced aspects of *Tools, Apps, and Scripts* that relate to the mobile usability of phones. Prior research has found that the vast majority of tools that exist are functionally limited to desktop computers [96]. These tools are often specifically designed to aid crowdworkers in finding and auto-accepting HITs for crowdworkers in order to accelerate their productivity. The entirety of the responses within this theme centered around tools that not only finds and auto-accepts HIT opportunities, but does so in a fashion that targets HITs designed for completion on the smartphone:

> *"I would have an app like I mentioned before that lists only available hits that are smartphone-friendly."* (P64)

Through these responses, we find that crowdworkers believe that there exists a set of HIT characteristics or specific HIT types that make a HIT conducive to complete on smartphones. 21 participants (20%) responded to Q24.2 by explicitly mentioning the need to "mirror the scripts' functionality from my desktop" (P14) in order to find and manage HITs as efficiency on the smartphone. Several participants' responses made reference to the use of the phone was situational and that the smartphone may be used in certain circumstances (e.g., "when I'm not at home" [P63]). In general, these responses not only reinforce the nature of our taxonomy, but also highlight barriers within work practices that exceed beyond the scope of discussion of our taxonomy's characteristics. We elaborate on the frontier for tooling research further in Section 6.3.

4.2.4 Reinforcing Mobile Usability by Task Example. Through our analysis, we observe that crowdworkers hold strong preferences for engaging in specific HITs on their smartphones. Figure 3 shows the representation of labels for the HIT types that are currently completed by crowdworkers alongside the HIT types they believe need further support. Reported by 33 participants (31%), we find that Survey HITs are the most prominent type of HITs currently completed on their smartphone, accounting for 47% of the responses to the question. Surveys, in particular, were often accompanied by anecdotal evidence that described their underlying interaction as a motivator for their mobile suitability:

> *"[I'll complete] some surveys that allow it. Anything that does not involve a lot of writing. Some batches that require picking radio buttons."* (P28)

Alongside survey HITs, "Other" was the second-most prominent label, reported by 35 participants (33%). Specific responses

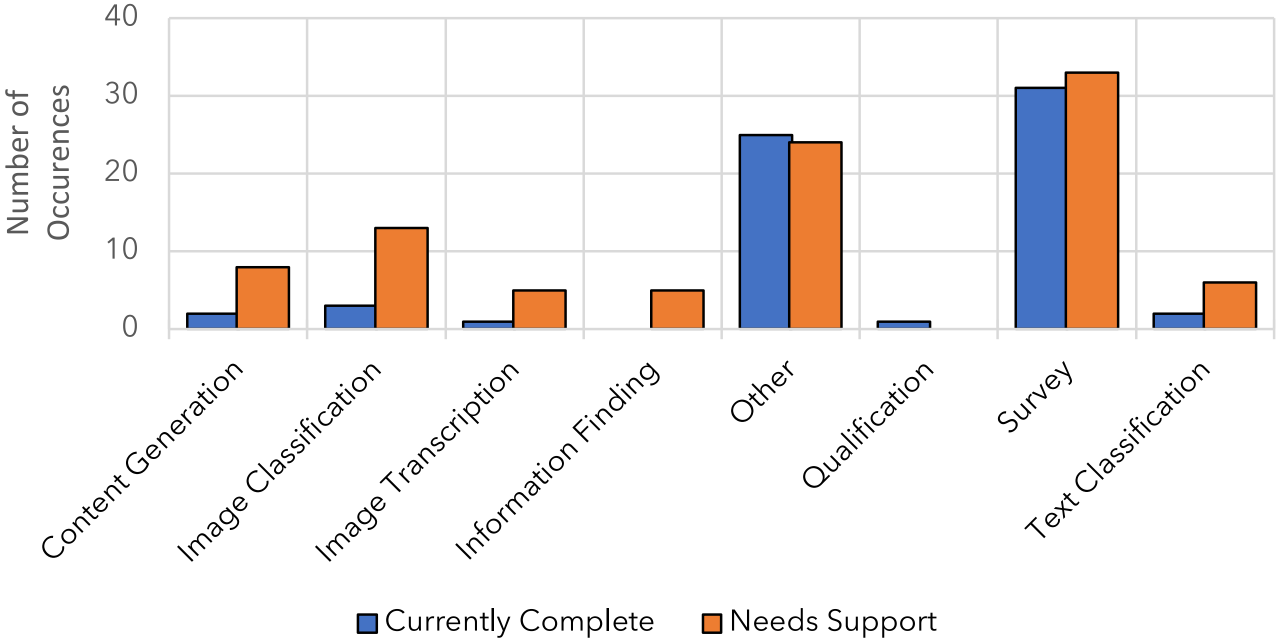


Figure 3: Histogram of task types labels that are currently completed or require additional support on smartphones.

emphasized the description of characteristics that facilitate mobile usability rather than a specific type of tasks itself:

> *"HITs that look relatively quick. So, [there isn't] much risk if I have to return if they are taking too long. I used to do the dating profile pic HITs as well, but have not recently. I always try to avoid HITs with writing on my smartphone."* (P51)

Other descriptions included "tasks that are simple and don't involve writing" (P87), "quiz-type of HITs" (P60), and "anything with bubble questions to fill out easily" (P25). Several workers referred to these task types as "batch HITs" (e.g., where workers can auto-select from a pool similarly structured tasks from the same requester) and "bubble HITs" (e.g., where workers select from multiple choice answers) Only three participant responses mentioned an explicit type of task, all of which referred to "a lot of app and website testing" (P30).

In contrast to survey HITs and miscellaneous "Other" HITs, the remaining HIT type labels were far less reported. Taken as a collective, our participants engage in tasks that require little navigation beyond the task interface, facilitate quick responses, and require a small amount of time to complete. Mirroring the themes that naturally arose through our coding of Q24.2, we find that much of the reluctance to engage with specific HITs on the smartphone as they are generally not "mobile-optimized":

> *"I will typically only do surveys on a phone if they are mobile-optimized because batch work is impractical on a phone. Most batches use iFrames and want the work done there, which just doesn't translate well to iPhone."* (P29)

Alongside observations related to HITs that crowdworkers currently tend to engage on their smartphone, we observe that crowdworkers not only have an increased desire for currently unsupported HITs, but also have a desire to improve the HITs they're already engaging with while mobile. 34 responses were categorized as voicing a need for further supporting survey HITs while 25 responses were categorized for "Other" types (e.g., HITs that involve downloading files on the smartphone). The specific motivation for these HITs related to the themes that emerged in our analysis of responses to Q24.2, namely *Design and Compatibility*:

> *"I would like surveys to be better supported on the smartphone. I want them to fit the screen and not involve a whole bunch of scrolling."* (P89)

Unlike reports for current counts, participants voiced a need for desiring additional support for all HIT types on smartphones with the exception of qualification HITs, which was not observed in this data in any capacity. Within these specific HIT types, responses often mentioned elements that would improve usability by reducing the effort required to complete a particular task. For example, Image Classification and Bounding Box HITs, which were reported by 13 and 6 participants, could benefit from "easier image bounding" (P33) on the smartphone. Specific recommendations for other HIT types (e.g., Sentiment Analysis) were often not included in responses due to the breadth of the question.

Overall, the findings in this section address RQ1 and RQ2 by illustrating that Mechanical Turk workers do indeed see the same characteristics we have organized from literature.

5 TAXONOMY DEMONSTRATION - STEP 3: AN ANALYSIS OF HITS IN-THE-WILD

The results from Step 2's online survey demonstrate that crowdworkers have clear and strong preferences and practices for completing HITs that have characteristics that make them "mobile-optimized". Our qualitative analysis of participant responses specifically suggests that these preferences and practices are molded around the characteristics embodied by our proposed taxonomy. We now aim to conclude our research with a study aimed at using our taxonomy toward assessing the mobile usability of tasks that exist on crowdwork platforms.

5.1 Method: Web Scraping and Taxonomy Application

The goal of this study is to demonstrate the utility of our taxonomy. In pursuit of this goal, we build a dataset of HITs using data collected from Amazon Mechanical Turk (MTurk) and leverage our proposed taxonomy as a tool for assessing the mobile usability of the HITs within this dataset. Our approach is specifically inspired by prior studies that manually sample a dataset of HITs from MTurk for manual investigation and analysis [83].

5.1.1 Dataset Generation. We designed web scraping software to scrape HITs and their associated metadata directly from Amazon Mechanical Turk (Figure 9). Alongside its collection of basic metadata, the tool was designed to capture two screenshots of each scraped HIT's interface as it renders both in the desktop browser viewport and in the mobile browser viewport. In order to ensure all aspects of the task interface were captured, automated screenshot behavior for both viewports was configured to iteratively screenshot, scroll down, and repeat until the entirety of the vertical visual space had been captured. All information was temporarily stored on two researchers' machines. We developed the tool in NodeJS with Puppeteer[3], a Node library which provides a high-level API to programmatically control a Chrome browser. We collected a total of 519 HITs during June 2021 using a functional worker account on Mechanical Turk.

5.1.2 Qualitative Coding for Taxonomy Characteristics. Using the characteristics in the developed taxonomy as a qualitative codebook, we sought to apply a top-down coding process to the dataset of scraped HITs. The goal of this coding process was to evaluate the usability characteristics of each HIT in our dataset. For each HIT, two coders selected the most appropriate value within each characteristic. When labeling a HIT, for every characteristic in Table 1, coders selected one of it's values. For example, a HIT's Mobile-Device Demands could be assigned a value of "None" or "Present", based on the definitions in Table 1.

During this coding process, a subset of 100 HITs (20%) was randomly sampled from the generated dataset of 519 HITs, and two researchers were tasked with labeling each characteristic in the taxonomy. Instructions for labeling involved the use of both the collected HIT metadata and the automatically captured screenshots. Researchers were instructed to make judgements based on what was shown and captured in a HIT's associated screenshots rather than make subjective judgements.

Over the course of labeling the subset of 100 HITs, the two labeling researchers encountered HITs that displayed content that was partially visible or entirely invisible. To account for these scenarios, we developed an eighth label for each HIT in the dataset named "Content Visibility". Thus, Content Visibility was coded with a bottom-up coding scheme to further our goal of evaluating the mobile usability of HITs in the dataset. This label is relevant to our scraping and coding process and is not included as a characteristic in our taxonomy. Below, we describe the possible values for this label that emerged in reviewing and discussing the sampled HIT instances in the subset of 100 HITs:

- *Visible Content*: The HIT interface is mostly or entirely visible and can be assessed for mobile usability.
- *External Survey Link*: The HIT interface cannot be assessed due to including task instructions alongside an external link to a survey to be completed on a different platform (e.g., Qualtrics, SurveyMonkey, etc).
- *Requester Configuration*: The HIT interface cannot be assessed due to Mechanical Turk algorithmically prohibiting crowdworkers from accessing it via a mobile browser or viewport. An explicit error message is shown in the browser.
- *Acceptance Required*: The HIT interface cannot be assessed due to Mechanical Turk requiring that crowdworkers accept the HIT in order to view the task. Instructions are often shown, but the task itself is not.
- *Data Collection Gap*: The HIT interface cannot be assessed due to missing relevant information stemming from a failure caused by our web scraping software or by the Mechanical Turk platform.
- *Language Mismatch*: The HIT interface cannot be assessed due to being written in a language that is not English.

Following the inclusion of this label, the two researchers revisited the subset of 100 HITs and assigned labels accordingly. For each HIT, researchers first assessed "Content Visibility" as it is a prerequisite for the presence of other labels. If content was visible for a HIT, a label was assigned for each of the seven characteristics. Label agreement was observed to be high across both "Content Visibility" (κ=0.95) and each of the seven labels (κ=0.80; κ=0.92; κ=0.88; κ=0.86; κ= 0.9; κ=0.97; κ=0.71). Any disagreements or uncertainties were resolved through follow-up discussion. The remaining 419 HITs in the dataset were divided equally among the two researchers to label independently.

After all HITs with Visible Content were labeled with values for each characteristic, we assigned each value label with Good, Fair, or Poor usability. Each value in our taxonomy is associated with Good, Fair, or Poor usability (see Table 1). For example, the "None" value in the Multi-Device Demands Characteristic is mapped to Good (green). We use this mapping in our findings in charts and to compare usability among HIT types. The taxonomy provides a direct mapping for all characteristic values in our taxonomy except for "Image" and "Text" in the "Prompt Modality". For Prompt Modality, we conditioned Image and Text to have Poor usability if

[3] https://github.com/puppeteer/puppeteer

the HIT had excessive scrolling. Otherwise, Image and Text was set to Good usability. This method of mapping usability values (MUR) enabled this research to generate charts that show usability per characteristic and HIT type e.g., in Figures 5, 6 and 8 as well as χ^2 tests.

5.2 Findings

Content visibility issues were prominent in the dataset of scraped HITs. Among the dataset of 519 HITs, a total of 261 HITs (50.3%) were labeled as having issues related to content visibility and were therefore impossible to evaluate with respect to usability. Among the HITs which had visibility constraints, HITs labeled as having an *External Survey Link* were among the most prominent, accounting for 136 of the HITs in the dataset (26.2%). The second most common label for problematic instances were HITs labeled as *Data Collection Gap* where the visibility issues occurred related to errors in our web scraping software. The remaining 75 (14.4%) of HITs experienced visibility issues related to configuration issues where requesters disallowed HITs on mobile devices by detecting iOS or Android user-agent strings, requiring the HIT to have been accepted before viewing, or having been written in a non-English language. 20 HITs (3.9%) out of 75 HITs were labeled as a *Language Mismatch* were all written in Spanish. We now present observations from an analysis of the remaining and fully-labeled 261 HITs (50.3%) examining trends in characteristics and HIT types.

5.2.1 Assessing Mobile Usability with Task Characteristics. The taxonomy's *Task Characteristics* are *Off-site Knowledge Demands*, *Task Divisibility*, *Task Size Volume* and *Task Completion Time.* Overall, the mobile usability is mixed depending on the particular characteristic. Our analysis suggest that in terms of *Off-site Knowledge Demands* 193 HITs (73.9%) belong to Good Usability as they do not require to navigate away from the main HIT interface. HITs are well suited on smartphones when they have been split into small subtasks. Task Divisibility seems to have the worst usability, as only 3.4% of the HITs are *Already Divided.* In contrast, 21.8% of HITs are *Indivisible* and 74.7% are *Divisible*, but the requesters have not divided them. *Task Size Volume* exhibits even less usability. Only 49 HITS (18.8%) from our dataset have *Low Volume*, leaving 81.2% as High Volume which are less usable on smartphones. Only 8.8% of HITs can be accomplished *Fast*, 64.8% *Normal*, and 26.4% are *Slow.*

5.2.2 Assessing Mobile Usability with Interaction Characteristics. The *Interaction Characteristics* of our taxonomy consist of *Multi-device Demands*, *Prompt Modality* and *Interaction Style.* From scraped HITs, we observe that on average majority of them show *Good Usability* characteristics. It indicate that they would be substantially supported on smartphone. In terms of *Multi-device Demands*, our data show that 252 HITs (96.6%) have *Good Usability* on smartphones as they belong to *None*, while 9 HITs (3.4%) belong to *Poor Usability.* Our data shows that most of the HITs can be highly mobile usable. 205 HITs (78.5%) can be fully supported on smartphone in terms of modality. The remaining 56 HITs (21.5%) that consist of *Image or Text* have *Poor Usability* because of excessive scrolling or occlusion. From our dataset we can see that 247 HITs (94.6%) can be supported using mobile phones with respect to the characteristic *Interaction Style.* These HITs were interacted with using the methods which include *Good-Fair Usability.* They include *Option Selections, Recording, Open-ended+, Object Manipulation* and *Natural Interface Input.* The remaining 12 HITs (4.6%) have *Poor Usability* as their interactions were through *File Management.*

HIT Type	Good	Fair	Poor	Usability Percentage %
Image Classification	247	227	93	83.6
Survey	32	13	11	80.4
Text Classification	57	17	24	75.5
Other	63	31	32	74.6
Image Transcription	91	75	58	74.1
Information Finding	200	173	166	69.2
Qualification	6	8	7	66.7
Content Generation	72	32	92	53.1

Table 2: Usability values (MUR) for Good, Fair and Poor and usability percentage aggregated across seven characteristics across each HIT type. We order each HIT type by the Usability Percentage.

5.2.3 Assessing Mobile Usability Across HIT Types. Four HIT types – Image Classification, Information Finding, Content Generation, and Image Transcription – accounted for 214 (81.9%) of the 261 HITs. Image Classification is the most the most commonly observed HIT type 81 of the 261 HITs (31%), followed by 77 (29%) *Information Finding* HITs and 32 (12%) *Image Transcription* HITs. The remaining observed HIT labels include 28 *Content Generation* HITs (10%), 14 *Text Classification* HITs (5.4%), 8 *Survey* HITs (10%), and three *Qualification* HITs (10%). A total of 18 HITs, such as text moderation, image quality assessment, and copy editing, are identified as the "*Other*" category

Figure 8 shows each HIT Type across the 261 fully labeled HITs that represent their distribution of our taxonomy's seven mobile usability characteristics. Figure 7 and Table 2 show the percent of Good, Fair, and Bad usability aggregated across characteristics. It also includes a Usability Percentage, which we calculate per Task Type by adding the total Good and Fair characteristic ratings by the total number of ratings. This creates a ranking of HIT Types ordered by Good+Fair usability: Image Classification, Survey, Text Classification, Other, Image Transcription, Information Finding, Qualification, and Content Generation. However, if we were to only consider the total Good usability per HIT Type, Text Classification and Survey would have the most usability. This agrees with the HIT types that our survey participants mentioned. Both Text Classification and Survey have *Good usability* relative to other HIT Types, especially their Offsite-Knowledge Demands, Multi-Device Demands, and Interaction Style.

We observe that some characteristics are more commonly identified as "Good" or "Poor" for certain HIT Types in comparison to others. For example, unsurprisingly, Information Finding HITs have Poor usability for the Off-site Knowledge Demands characteristic. Information Finding typically, though not always, involves searching through external web pages. As another example, nearly all Image Transcription and Image Classification tasks have a High Task Volume. At the same time, almost half of Image Transcription and nearly all of Image Classification are *Divisible*, meaning they could be designed in smaller chunks. This represents evidence

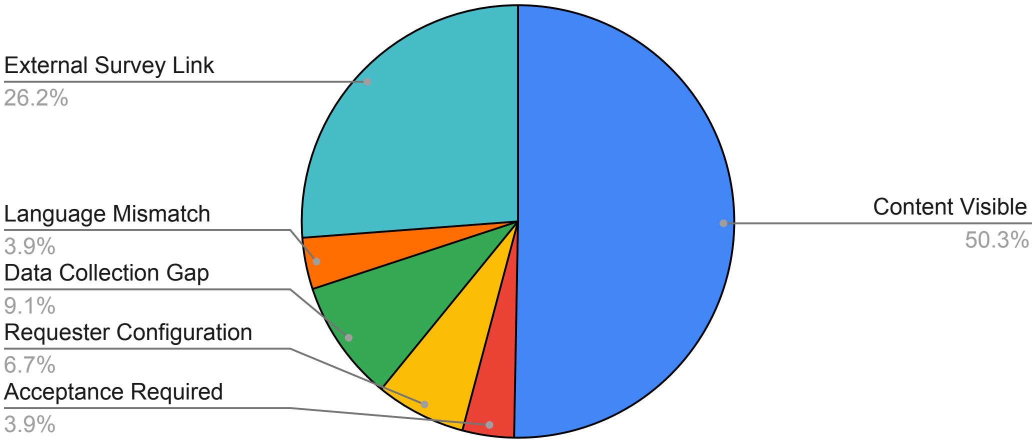
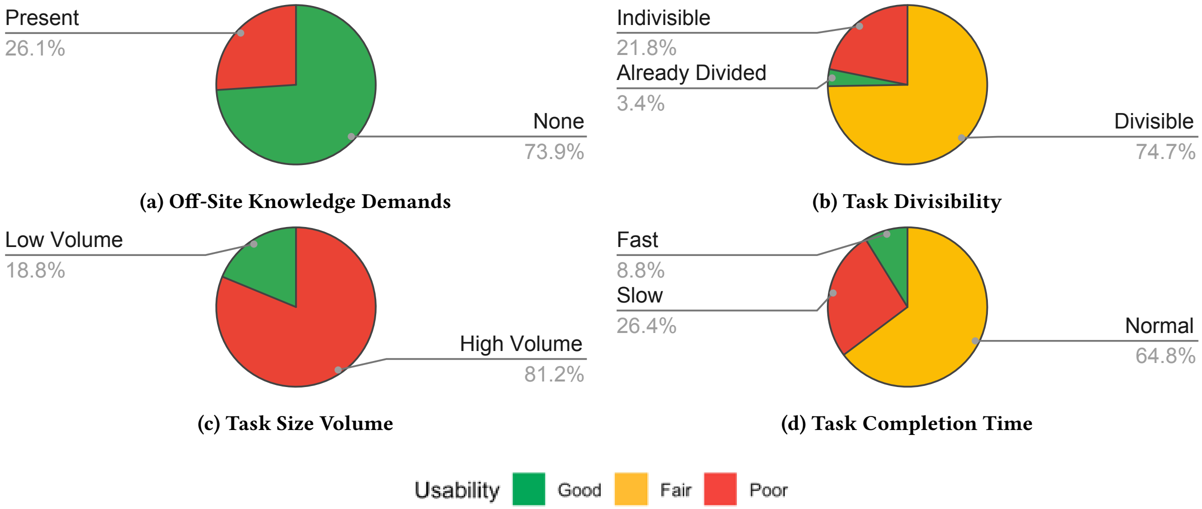


Figure 4: Representation of labels for "Content Visibility" across the dataset of 519 HITs.



(a) Off-Site Knowledge Demands

(b) Task Divisibility

(c) Task Size Volume

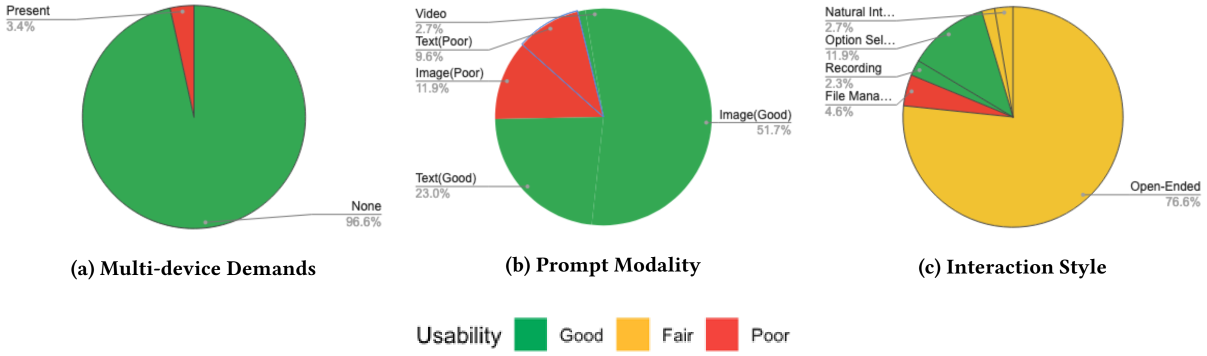
(d) Task Completion Time

Figure 5: Representation of attributes for task characteristics: (a) Off-Site Knowledge Demands with 26.1% Present and 73.9% None, (b) Task Divisibility with 74.7% divisible, 21.8% Indivisible, and 3.4% Already Divided, (c) Task Size Volume with Low Volume with 81.2% High Volume and 18.8% Low Volume, and (d) Task Completion Time with 8.8% fast, 64.8% Normal, and 26.4% Slow.

that requesters could, for example, redesign these two specific HIT types – Image Transcription and Image Classification – in support of generally improving their mobile usability.

By rank via Figure 7, the least usable HIT Types include Information Finding, Qualification, and Content Generation. These show more Poor characteristics relative to other HIT Types, with at least some Poor usability for all characteristics except Multi Device Demands. The remaining HIT Types have mixed usability, having some Good, Fair, and Poor more evenly (though differently) distributed.

To quantitatively examine how the distribution of usability characteristics vary between HIT types, we conducted a series of Chi-squared tests to test for significant differences. To create task profiles, we binned each unique value-characteristic pair and tallied their occurrences, creating 8 vectors of tallies with 16 dimensions each. We employed a version of the Chi-squared test χ^2 to eradicate issues related to limited observations in our data, that generated multiple simulations to account for smaller data to compare distributions. Each test compared the given distribution against the distribution of the sum of the other (7) profiles. In this case, the null hypothesis is that the given HIT Type's profile is not significantly different from the rest of the HIT corpus' profile. We adjust p-values by applying Bonferroni correction to account for multiple



(a) Multi-device Demands

(b) Prompt Modality

(c) Interaction Style

Figure 6: Representation of attributes for interaction characteristics: (a) Multi-device Demands, (b) Prompt Modality, (c) Interaction Style.

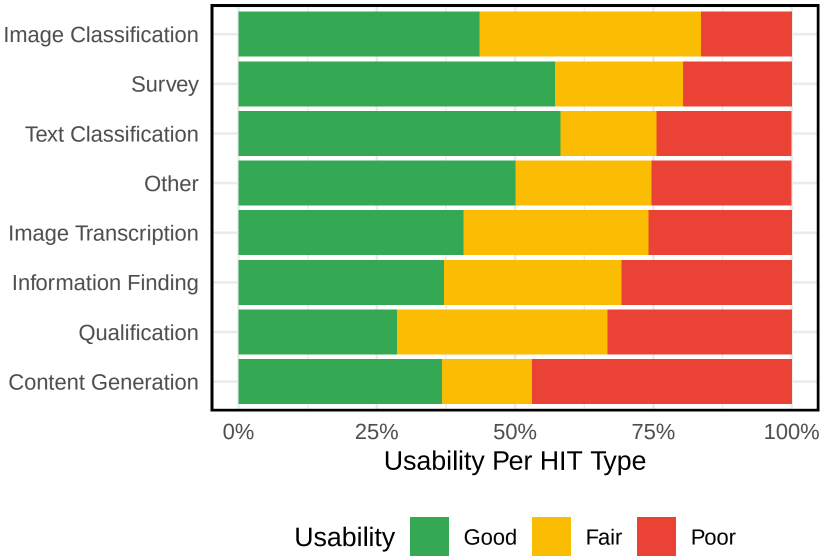


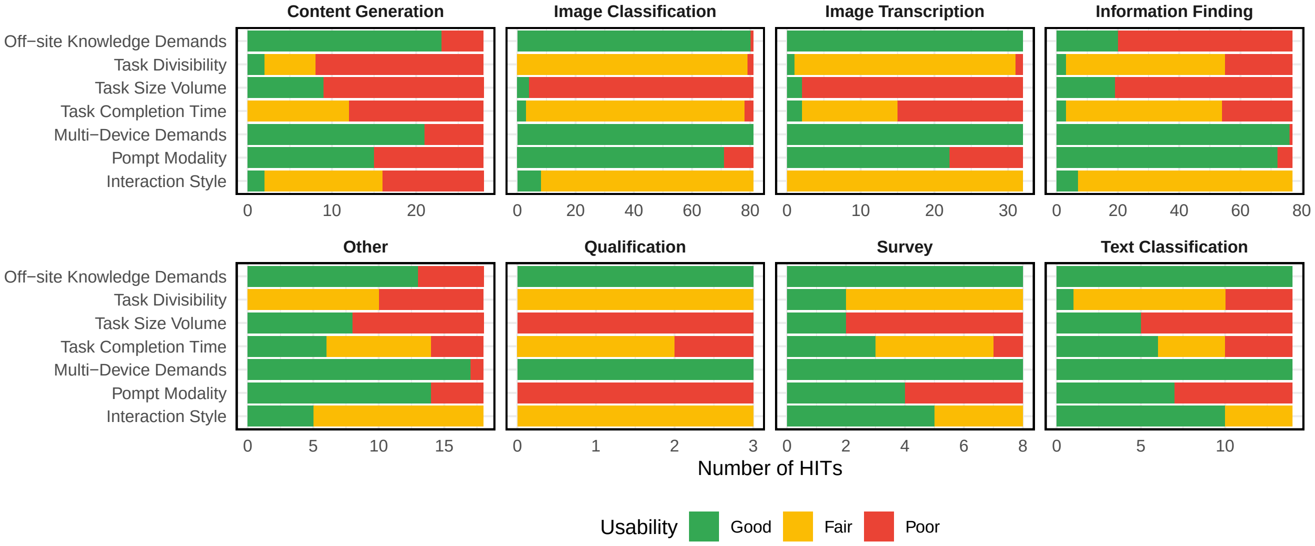
Figure 7: Distribution of Good, Fair, and Poor across each HIT type. Here, we order by the least amount of total Poor usability values. This provides a ranking of Tasks, ordered by most to least usable according to our taxonomy.

hypothesis testing and report effect size χ^2 (Cramer's V), and the χ^2 statistic. Table 3 shows the results of our Chi-squared tests [66].

Overall, six of the eight Chi-squared tests on HIT Types yielded a result that suggests they exhibit usability characteristic distributions that are statistically significant. As shown in Table 3, each of the five tests yielded a p-value less than 0.05 with effect sizes ranging from 0.24 to 0.93. Text Classification (p<.005), Information Finding (p<.005), Image Transcription (p<.005), Image Classification (p<.005), Survey (p<.05), and Content Generation (p<.005) all have significantly different usability profiles compared to sum of the other HITs. This suggests that our intuition is correct, that the usability distributions vary based on the type of HIT because of their nature or how requesters tend to design them. These findings, tables, and charts address RQ3 directly, establishing the distribution of characteristics across HIT Types (Figure 8) and provides a ranking of hits most conducive to mobile interaction (Figure 7).

6 DISCUSSION

Our study provides insight on the state of mobile crowdwork. Our work began by developing a taxonomy of characteristics that reflect the usability of HITs on smartphone devices. In demonstrating the taxonomy's utility, we observe that some HIT types – namely Image Classification HITs and Survey HITs – are generally more usable on smartphone devices than other HIT types. Further, we observe that six of the eight examined HIT types exhibit characteristic profiles that are significantly unique from other profiles.



HIT Type	χ^2	β	p	
Content Generation	229.60	0.938	0.004	**
Image Classification	143.26	0.741	0.004	**
Image Transcription	49.41	0.435	0.004	**
Information Finding	163.87	0.792	0.004	**
Other	32.94	0.356	0.095	
Qualification	15.041	0.240	1.000	
Text Classification	79.19	0.550	0.004	**
Survey	44.829	0.414	0.020	*

Table 3: Results from Chi-Squared tests across task types. These test compare the distribution each HIT Type's distribution of usability values (MUR) to the other HITs in the set. Significance indicates a relatively distinct profile of usability. (*: p<0.05; **: p<0.01)

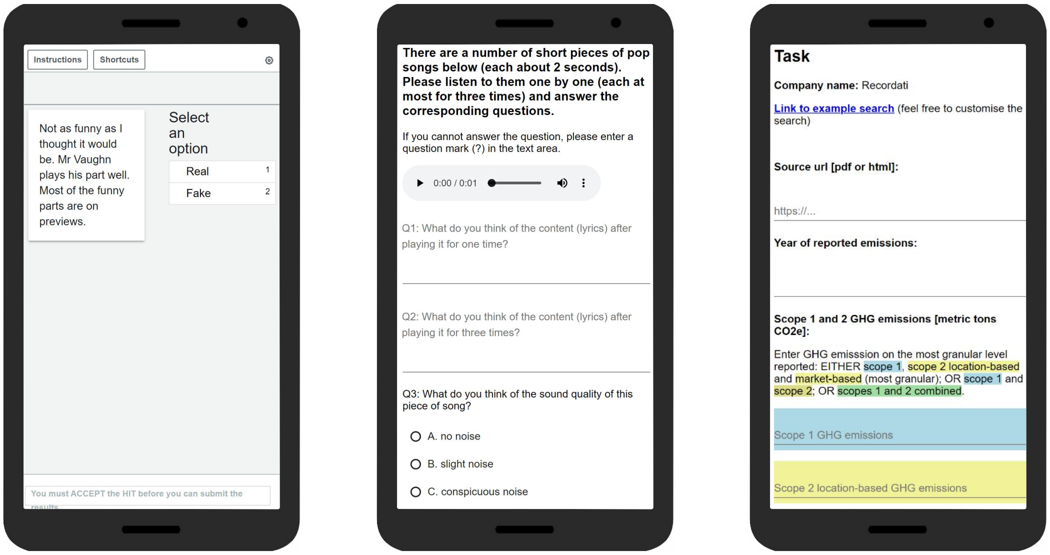
Collectively, our work demonstrates that many HIT opportunities on crowdsourcing platforms, such as Amazon Mechanical Turk, are significantly limited in their mobile usability. These conclusions are further supplemented by trends in preference and practice as self-described by crowdworkers who work on the platform today.

An important consideration for interpreting our results is the characterization of mobile usability. The novelty of the developed taxonomy is grounded in practicality and it ability to surface trade-offs. More precisely, we view the taxonomy as a tool for measuring and tweaking the mobile usability of HITs on mobile devices. For example, in Figure 9, the Content Generation task has High Volume, but is divisible. This means the Requester could make it more suitable for mobile smartphone devices by having fewer subtasks per HIT. The same HIT includes an Open-ended+ Interaction Style that can support more nuanced answers at the cost of having less mobile usability. To switch to an Option Selection would make this task faster and more usable, but at the potential cost of data collection needs of the requester.

As shown in Figure 4, the taxonomy can be used to assess HITs that are diverse in their nature, spanning a multitude of task types, requirements, and constraints. While each characteristic is designed to represent a particular dimension of a HIT's design, the taxonomy's characteristics are intended to be used in unison. Despite being significantly thorough, our examination of "mobile usability" was conducted to understand usability in the specific context of mobile smartphone devices. We now discuss the implications of the developed taxonomy in the context of mobilizing crowdwork both within and beyond smartphone devices.

6.1 Implications for Design: A Taxonomy for Practicing "Good" Mobile Usability

Our research demonstrates how our taxonomy provides a *framework for designing* HITs that are "mobile-optimized". In HCI, Grudin and Poltrock refer to taxonomies as "pretheoretical constructs that characterize cooperative work and identify the technologies that support different types of work" [28]. In our taxonomy none of the characteristics are independently functional, all of them need to be taken together to support mobile suitability. This provides a language for HIT designers potentially strong and weak aspects of designing HITs and balancing both aspects. Requesters (i.e., HIT designers) can utilize our taxonomy and its characteristics as a checklist for ensuring the design of their interfaces are usable on smartphone devices. For example, creating an idealized mobile HIT would avoid *Off-Site Knowledge Demands*, maintain *Already Divided content*, and use *Option Selections* for its interaction style. While some tasks are inevitably more inclined to be less conducive to mobile experiences, our work establishes a design space for future



Not as funny as I thought it would be. Mr Vaughn plays his part well. Most of the funny parts are on previews.

Select an option

Real 1

Fake 2

You must ACCEPT the HIT before you can submit the results

There are a number of short pieces of pop songs below (each about 2 seconds). Please listen to them one by one (each at most for three times) and answer the corresponding questions.

If you cannot answer the question, please enter a question mark (?) in the text area.

Q1: What do you think of the content (lyrics) after playing it for one time?

Q2: What do you think of the content (lyrics) after playing it for three times?

Q3: What do you think of the sound quality of this piece of song?

○ A. no noise

○ B. slight noise

○ C. conspicuous noise

Task

Company name: Recordati

Link to example search (feel free to customise the search)

Source url [pdf or html]:

https://...

Year of reported emissions:

Scope 1 and 2 GHG emissions [metric tons CO2e]:

Enter GHG emisssion on the most granular level reported: EITHER scope 1, scope 2 location-based and market-based (most granular); OR scope 1 and scope 2; OR scopes 1 and 2 combined.

Scope 1 GHG emissions

Scope 2 location-based GHG emissions

Task Type	*Text Classification*	*Content Generation*	*Information Finding*
Off-site Knowledge Demands	🟩 None	🟥 Present	🟥 Present
Task Divisibility	🟩 Already Divided	🟨 Divisible	🟥 Indivisible
Task Size Volume	🟩 Low Volume	🟥 High Volume	🟥 High Volume
Task Completion Time	🟩 Fast	🟨 Normal	🟥 Slow
Multi-device Demands	🟩 None	🟩 None	🟩 None
Prompt Modality	🟩 Text (No Scrolling)	🟩 Audio	🟥 Text (Extensive Scrolling)
Interaction Style	🟩 Option Selection	🟨 Open-ended+	🟨 Open-ended+

Figure 9: Three example HITs from our scraped corpus with their following Mobile Usability Rating (MUR) based on our taxonomy of characteristics (Figures 1 and 8). Each has a different Task Type and corresponding MUR ratings. The Sentiment Analysis / Text Classification HIT has an excellent MUR rating. In particular, the Interaction Style and low Task Size Volume is well-aligned for completing on a mobile phone. In contrast, the Information Finding HIT has a poor MUR rating. It requires switching away from the main HIT interface to a website, searching for URLs, and entering the answers with an Interaction Style of Open-ended+. In the middle, the Content Generation HIT can be performed on mobile phone, listening to short clips of pop-music from the same page, but still uses and Interaction Style of Open-ended+.

research that seeks to instantiate different HIT designs that leverage the insights of our taxonomy.

Alongside its use as a guideline for mobile design, our taxonomy can also be applied as a *tool for assessing existing HIT design*. Our studies suggest that crowdworkers experience challenges in finding and managing the HITs they accept to be well-matched to their device, setting, and, in some circumstances, their abilities [101]. Uzor et al. suggests that workers that identify as having an impairment would benefit from better market-provider enforced metadata that specifies whether, for example, people with visual impairment would be well-matched to a HIT [89]. There exists a fundamental opportunity for translating our taxonomy of characters into an automated tool as prior studies have previously done in their own usability contexts [1, 91]. Amazon Mechanical Turk and other crowdsourcing platforms could incorporate such a tool to provide requesters with feedback about improving their HIT designs in the same way that interfaces for password creation provide feedback about the "strength" of a password. Further, reviewing platforms, such as Turkopticon [44], could stand to benefit from attaching systematic metadata about the mobile usability of HITs alongside the standard reviewing metadata. The collection of such data could facilitate the creation of data sets of mobile usability ratings, which could be utilized toward automated tools for detecting HITs that have "good" mobile usability.

Alongside general-purpose usability frameworks, there exists an opportunity to explore convergent perspectives of usability and design as we move toward a practice of cross-device crowdwork. For example, Nakatsu et al. [61] developed a taxonomy of crowdsourcing tasks based on task complexity across two dimensions: structure and independence. Our taxonomy presents a thorough examination of mobile HIT usability through well-structured and independent tasks that Nakatsu et al. refer to as "contractual hiring".

There remains a significant opportunity for further characterizing mobile usability in crowdwork that, according to Nakatsu et al., involve multiple parties and significantly less structure.

Finally, our taxonomy can serve as a useful complement to current and future frameworks alike for other niche contexts. For example, significant attention has been given to understanding pathways for improving the accessibility of work opportunities in crowdwork [83]. In a study of accessibility on Amazon Mechanical Turk, Zyskowski et al. found that 39% of respondents reported using an assistive device to engage in crowdwork [101]. Reporting a similar percentage of assistive device usage, Uzor et al. found that these users gravitate toward completing survey HITs in comparison to other HIT types (e.g., information finding) [89]. Though our focus was limited to mobile usability, our taxonomy may be useful to individuals within the accessibility community who have struggled, or continue to struggle, with characterizing accessible HITs. In the same way that "most accessibility fixes actually make products better for all users" [13], we expect that making HITs more usable will positively impact all Mechanical Turk workers.

6.2 Toward a Usable Practice of Cross-Device Crowdwork

Our research takes an important step toward defining and evaluating mobile usability for crowdwork. A wealth of prior research at the intersection of crowdwork and mobile computing has repeatedly reinforced the importance and opportunities of smartphone devices [23, 30, 45], which collectively drove us to focus explicitly on understanding usability within this type of device's context. Our study serves as a foundation for developing and conducting future usability assessments across other types of devices or specific situational contexts. Recent studies suggest that crowdworkers have an interest in using more devices than their smartphone alone to support their work [35, 96]. For example, Hettiachchi et al. studied task acceptance rates across desktop computers, smartphones, and voice assistants, observing that preferences of task acceptance for smartphones and voice assistants are only slightly smaller than preferences for the desktop [36]. As non-desktop devices continue to become increasingly more important to crowdwork, there exists a growing need to understand how usability should be defined within each of their respective contexts.

By studying, understanding, and assessing device-related usability, crowdsourcing researchers, crowdworkers, and platform can begin to develop new "cross-device" systems, tools, and experiences [64]. *Information work*, for example, encompasses a range of computer-based professions (e.g., programming, design, writing) – many of which are recognized as desktop-centric practices [68]. Among these traditionally desktop-centric professions, studies have shown that people use the smartphone to facilitate communication [39], transmit information across devices [64], and continue tasks more generally across devices (e.g., web browsing [48, 63]). Interactive systems research has focused explicitly on designing new cross-device systems and tools to better understand the benefits that arise from cross-device experiences (e.g., in software development [38] and in every-day experiences [67]). Aligned to our discussion of mobile usability, Mercury [95] and PlayWrite [43] are two mobile microtasking systems that employ microtasks that would, by our own taxonomy's assessment, be deemed as having "Good" mobile usability. Despite being prototypes, each system's evaluation demonstrated its potential for impacting the nature of their respective work context substantially.

We argue that our work provides a framework for defining, understanding, and assessing "cross-device" in the scope of crowdwork: How is it technically feasible for a HIT to be completed across multiple devices? Are there administrative tasks related to crowdwork (e.g., finding HITs) that should also be facilitated across devices? Are there combinations of devices that are more usable with one another than other combinations? Each of these questions poses a particular challenge that collectively hinge on the fundamental characterization of what mobile usability means for a particular device. The landscape of research on cross-device crowdwork remains relatively small, and we therefore encourage researchers, crowdworkers, and marketplace platforms to recognize the area as one that is fruitful for innovation.

6.3 The Frontier of Mobile Tooling in Crowdwork

Our study provides insight into the role that tooling can play in facilitating mobile experiences in crowdwork that are both efficient and productive. We specifically find that crowdworkers have a desire to mobilize aspects of their work, but are limited by the tooling that exists today. A wealth of prior research has reinforced the role that workstation-based tools play in enabling efficiency in crowdwork [47, 80, 96]. Kaplan et al. specifically noted that many tools for efficient crowdwork are facilitated through platforms (e.g., Turkopticon [44]) or browser extensions (e.g., MTurk Suite) [47]. Today, Firefox for Android remains the only smartphone-based browser that allows users to load and employ browser extensions. Many of these tools rely on browser APIs that are only supported on desktop browser implementations and are therefore incompatible on mobile devices.

Within the purview of tool development, the role of the crowdworker continues to remain an important consideration. Many of the most long-standing tools in crowdwork, such as Turkopticon [44] and MTurkSuite, are not only worker-developed, but also worker-sustained. For crowdworkers, tools are "the glue that makes their work possible" [96], and this indicates a need to engage with them as researchers that continue to build tools to support them. Further, state-of-the-art workstation tools, such as Otto[4], require its users to pay a monthly fee of $10.00 per month in order to access the tool's core features. As tool development continues among researchers, crowdworkers, and other participates, there exists a broader challenge of ensuring that tools – whether they be for the desktop, for the smartphone, or any other device – remain available to the public in order to ensure facilitate work experiences that are not only productive, but also fair to crowdworkers at large. This is particularly relevant for crowdworkers (e.g, in rural areas) who may use their mobile device more often than a workstation computer [10, 23, 30, 92].

[4] https://www.ourhitstop.net/membership-tier-information

6.4 Limitations and Future Work

Our study has several limitations. First, our study is an examination of mobile crowdwork through the contemporary lens of Amazon Mechanical Turk. Though the platform is recognized as significant marketplace for crowdwork, our study cannot draw conclusions about crowdwork or gig work that occurs on other platforms under a different work structure (e.g., Upwork[5]). Second, the presented taxonomy was reinforced by an online survey deployed to 111 crowdworkers that work on Amazon Mechanical Turk in which participants described the HITs they currently engage with, or desire to engage with, on their smartphone. Participants' responses may be biased by the recency of their activities or by an inability to recall specific information. To reduce our concerns with this limitations, we supported the observations from our survey by conducting a follow-up study in which a dataset of HITs scraped directly from Amazon Mechanical Turk was qualitatively analyzed. However, a limitation of this follow-up study is that our analysis was limited to a total of 519 HITs. In general, our analysis suggests findings that reaffirm the observations made from our online survey, thus providing reliability to both choices in methodology. We would caution researchers that a direct MUR evaluation without additional context will not capture all aspects of HIT design relevant to mobile usability. The characteristics in our taxonomy apply most strongly to the context of current mobile phone devices and HIT types. Future studies can add nuance to these findings by employing more fine-grained methods and tools (e.g., activity logging apps for mobile devices) toward the targeted goal of assessing mobile task preference in practice.

7 CONCLUSION

There is a growing interest in extending crowdwork beyond traditional desktop-centric design to include mobile devices (e.g., smartphones). In this paper, we present the iterative development of the taxonomy, highlighting the observed practices and preferences around mobile crowdwork. We first establish an initial design of our taxonomy through a targeted literature analysis. We then support and extend the taxonomy through an online survey with Amazon Mechanical Turk crowdworkers. Finally, we demonstrate the taxonomy's utility by applying it to analyze the mobile usability of a dataset of scraped HITs. We conclude with a discussion around the implications of both the presented taxonomy and our study's findings as it relates to making crowdwork more usable on mobile devices.

REFERENCES

[1] Ritu Agarwal and Viswanath Venkatesh. 2002. Assessing a firm's web presence: a heuristic evaluation procedure for the measurement of usability. *Information systems research* 13, 2 (2002), 168–186.

[2] Alan Aipe and Ujwal Gadiraju. 2018. Similarhits: Revealing the role of task similarity in microtask crowdsourcing. In *Proceedings of the 29th on Hypertext and Social Media.* 115–122.

[3] Reham Alabduljabbar and Hmood Al-Dossari. 2019. A dynamic selection approach for quality control mechanisms in crowdsourcing. *IEEE Access* 7 (2019), 38644–38656.

[4] Lisa Anthony, Quincy Brown, Jaye Nias, Berthel Tate, and Shreya Mohan. 2012. Interaction and recognition challenges in interpreting children's touch and gesture input on mobile devices. In *Proceedings of the 2012 ACM international conference on Interactive tabletops and surfaces.* 225–234.

[5] Michael S Bernstein, David R Karger, Robert C Miller, and Joel Brandt. 2012. Analytic methods for optimizing realtime crowdsourcing. *arXiv preprint arXiv:1204.2995* (2012).

[6] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology.* 313–322.

[7] Michael S Bernstein, Jaime Teevan, Susan Dumais, Daniel Liebling, and Eric Horvitz. 2012. Direct answers for search queries in the long tail. In *Proceedings of the SIGCHI conference on human factors in computing systems.* 237–246.

[8] Nigel Bevan and Miles Macleod. 1994. Usability measurement in context. *Behaviour & information technology* 13, 1-2 (1994), 132–145.

[9] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. (2012).

[10] Pew Research Center. 2021. Mobile fact sheet. *Pew Research Center* (2021).

[11] Justin Cheng, Jaime Teevan, Shamsi T Iqbal, and Michael S Bernstein. 2015. Break it down: A comparison of macro-and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.* 4061–4064.

[12] Pei-Yu Chi, Anurag Batra, and Maxwell Hsu. 2018. Mobile crowdsourcing in the wild: Challenges from a global community. In *Proceedings of the 20th international conference on human-computer interaction with mobile devices and services adjunct.* 410–415.

[13] Elizabeth F Churchill. 2018. Putting Accessibility First. *ACM/SIGCSE Seek* 25 (2018), 24.

[14] Constantinos K Coursaris and Dan J Kim. 2011. A meta-analytical review of empirical mobile usability studies. *Journal of usability studies* 6, 3 (2011), 117–171.

[15] André de Lima Salgado, Leandro Agostini do Amaral, Renata Pontin de Mattos Fortes, Marcos Hortes Nisihara Chagas, and Ger Joyce. 2017. Addressing mobile usability and elderly users: Validating contextualized heuristics. In *International Conference of Design, User Experience, and Usability.* Springer, 379–394.

[16] Andre L Delbecq, Andrew H Van de Ven, and David H Gustafson. 1975. *Group techniques for program planning: A guide to nominal group and Delphi processes.* Scott, Foresman,.

[17] Vincenzo Della Mea, Eddy Maddalena, and Stefano Mizzaro. 2013. Crowdsourcing to Mobile Users: A Study of the Role of Platforms and Tasks.. In *DBCrowd.* Citeseer, 14–19.

[18] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and Dynamics of Mechanical Turk Workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM '18.* ACM Press, New York, New York, USA, 135–143. https://doi.org/10.1145/3159652.3159661

[19] 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 *Proceedings of the 24th international conference on world wide web.* 238–247.

[20] Evan W Duggan. 2003. Generating systems requirements with facilitated group techniques. *Human-Computer Interaction* 18, 4 (2003), 373–394.

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

[22] Haakon Faste, Nir Rachmel, Russell Essary, and Evan Sheehan. 2013. Brainstorm, Chainstorm, Cheatstorm, Tweetstorm: new ideation strategies for distributed HCI design. In *Proceedings of the sigchi conference on human factors in computing systems.* 1343–1352.

[23] Claudia Flores-Saviaga, Yuwen Li, Benjamin Hanrahan, Jeffrey Bigham, and Saiph Savage. 2020. The Challenges of Crowd Workers in Rural and Urban America. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 159–162.

[24] 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.* 218–223.

[25] Benjamin M Good and Andrew I Su. 2013. Crowdsourcing for bioinformatics. *Bioinformatics* 29, 16 (2013), 1925–1933.

[26] Catherine Grady and Matthew Lease. 2010. Crowdsourcing document relevance assessment with mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's mechanical turk.* 172–179.

[27] Mary L Gray and Siddharth Suri. 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass.* Eamon Dolan Books.

[28] Jonathan Grudin and Steven Poltrock. 2012. Taxonomy and theory in computer supported cooperative work. *The Oxford handbook of organizational psychology* 2 (2012), 1323–1348.

[29] 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.* 1843–1852.

[30] Benjamin V Hanrahan, Anita Chen, JiaHua Ma, Ning F Ma, Anna Squicciarini, and Saiph Savage. 2021. The Expertise Involved in Deciding which HITs are Worth Doing on Amazon Mechanical Turk. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.

[5] https://www.upwork.com/

[31] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. 2018. A data-driven analysis of workers' earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–14.

[32] Edgar Hassler, Jeffrey C Carver, Nicholas A Kraft, and David Hale. 2014. Outcomes of a community workshop to identify and rank barriers to the systematic literature review process. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*. 1–10.

[33] DM Hegedus and RV Rasmussen. 1986. Task effectiveness and interaction process of a modified nominal group technique in solving an evaluation problem. *Journal of Management* 12, 4 (1986), 545–560.

[34] Christothea Herodotou, Maria Aristeidou, Grant Miller, Heidi Ballard, and Lucy Robinson. 2020. What do we know about young volunteers? An exploratory study of participation in Zooniverse. *Citizen Science: Theory and Practice* 5, 1 (2020).

[35] Danula Hettiachchi, Zhanna Sarsenbayeva, Fraser Allison, Niels van Berkel, Tilman Dingler, Gabriele Marini, Vassilis Kostakos, and Jorge Goncalves. 2020. " Hi! I am the Crowd Tasker" Crowdsourcing through Digital Voice Assistants. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[36] Danula Hettiachchi, Senuri Wijenayake, Simo Hosio, Vassilis Kostakos, and Jorge Goncalves. 2020. How Context Influences Cross-Device Task Acceptance in Crowd Work. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 53–62.

[37] Mahmood Hosseini, Keith Phalp, Jacqui Taylor, and Raian Ali. 2014. The four pillars of crowdsourcing: A reference model. In *2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS)*. IEEE, 1–12.

[38] Maria Husmann, Alfonso Murolo, Nicolas Kick, Linda Di Geronimo, and Moira C Norrie. 2018. Supporting out of office software development using personal devices. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–11.

[39] Heather M Hutchings and Jeffrey S Pierce. 2006. Understanding the whethers, hows, and whys of divisible interfaces. In *Proceedings of the working conference on Advanced visual interfaces*. 274–277.

[40] Ursula Huws and Simon Joyce. 2016. *Crowd working survey: size of the UK's 'Gig Economy' revealed for the first time*. Technical Report. http://www.feps-europe.eu/assets/a82bcd12-fb97-43a6-9346-24242695a183/crowd-working-surveypdf.pdf

[41] Ursula Huws and Simon Joyce. 2017. *First survey results reveal high levels of crowd work in Switzerland*. Technical Report. http://unieuropaprojects.org/content/uploads/2017-09-13-factsheet-ch.pdf

[42] Kazushi Ikeda and Keiichiro Hoashi. 2017. Crowdsourcing go: Effect of worker situation on mobile crowdsourcing performance. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 1142–1153.

[43] Shamsi T Iqbal, Jaime Teevan, Dan Liebling, and Anne Loomis Thompson. 2018. Multitasking with Play Write, a mobile microproductivity writing tool. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 411–422.

[44] Lilly C Irani and M Six Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 611–620.

[45] Jason T Jacques and Per Ola Kristensson. 2017. Design strategies for efficient access to mobile device users via Amazon Mechanical Turk. In *Proceedings of the First ACM Workshop on Mobile Crowdsensing Systems and Applications*. 25–30.

[46] Matt Jones, George Buchanan, and Harold Thimbleby. 2002. Sorting out searching on small screen devices. In *International Conference on Mobile Human-Computer Interaction*. Springer, 81–94.

[47] Toni Kaplan, Susumu Saito, Kotaro Hara, and Jeffrey P Bigham. 2018. Striving to earn more: a survey of work strategies and tool use among crowd workers. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

[48] Amy K Karlson, Brian R Meyers, Andy Jacobs, Paul Johns, and Shaun K Kane. 2009. Working overtime: Patterns of smartphone and PC usage in the day of an information worker. In *International Conference on Pervasive Computing*. Springer, 398–405.

[49] Mohammad Taha Khan, Maria Hyun, Chris Kanich, and Blase Ur. 2018. Forgotten but not gone: Identifying the need for longitudinal data management in cloud storage. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[50] 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 *Proceedings of the 2013 conference on Computer supported cooperative work*. 1301–1318.

[51] Siou Chew Kuek, Cecilia Paradi-Guilford, Toks Fayomi, Saori Imaizumi, Panos Ipeirotis, Patricia Pina, and Manpreet Singh. 2015. The global opportunity in online outsourcing. (2015).

[52] Pilar Pazos Lago, Mario G Beruvides, Jiun-Yin Jian, Ana Maria Canto, Angela Sandoval, and Roman Taraban. 2007. Structuring group decision making in a web-based environment by using the nominal group technique. *Computers & Industrial Engineering* 52, 2 (2007), 277–295.

[53] Walter S Lasecki, Juho Kim, Nick Rafter, Onkur Sen, Jeffrey P Bigham, and Michael S Bernstein. 2015. Apparition: Crowdsourced user interfaces that come to life as you sketch them. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1925–1934.

[54] Zohar Laslo, Baruch Keren, and Hagai Ilani. 2008. Minimizing task completion time with the execution set method. *European Journal of Operational Research* 187, 3 (2008), 1513–1519.

[55] Edith Law and Luis von Ahn. 2011. Human computation. *Synthesis lectures on artificial intelligence and machine learning* 5, 3 (2011), 1–121.

[56] Songil Lee, Gyouhyung Kyung, Jungyong Lee, Seung Ki Moon, and Kyoung Jong Park. 2016. Grasp and index finger reach zone during one-handed smartphone rear interaction: effects of task type, phone width and hand length. *Ergonomics* 59, 11 (2016), 1462–1472.

[57] Jizi Li, Ying Wang, Dengku Yu, and Chunling Liu. 2021. Solvers' committed resources in crowdsourcing marketplace: do task design characteristics matter? *Behaviour & Information Technology* (2021), 1–20.

[58] David Martin, Benjamin V Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 224–235.

[59] Ian McGraw, Chia-ying Lee, I Lee Hetherington, Stephanie Seneff, and James R Glass. 2010. Collecting Voices from the Cloud.. In *LREC*. 1576–1583.

[60] Brian Mullen, Craig Johnson, and Eduardo Salas. 1991. Productivity loss in brainstorming groups: A meta-analytic integration. *Basic and applied social psychology* 12, 1 (1991), 3–23.

[61] Robbie T Nakatsu, Elissa B Grossman, and Charalambos L Iacovou. 2014. A taxonomy of crowdsourcing based on task complexity. *Journal of Information Science* 40, 6 (2014), 823–834.

[62] 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, 11 (2011), 45.

[63] Michael Nebeling and Anind K Dey. 2016. XDBrowser: user-defined cross-device web page designs. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5494–5505.

[64] Michael Nebeling, Theano Mintsi, Maria Husmann, and Moira Norrie. 2014. Interactive development of cross-device user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2793–2802.

[65] 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 the 2016 CHI conference on human factors in computing systems*. 3834–3846.

[66] MS Nikulin. 1973. Chi-squared test for normality. In *Proceedings of the International Vilnius Conference on Probability Theory and Mathematical Statistics*, Vol. 2. 119–122.

[67] Elnaz Nouri, Adam Fourney, Robert Sim, and Ryen W White. 2019. Supporting complex tasks using multiple devices. In *Proceedings of WSDM'19 Task Intelligence Workshop (TI@ WSDM19)*.

[68] Antti Oulasvirta and Lauri Sumari. 2007. Mobile kits and laptop trays: managing multiple devices in mobile information work. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1127–1136.

[69] Veljko Pejovic and Artemis Skarlatidou. 2020. Understanding interaction design challenges in mobile extreme citizen science. *International Journal of Human–Computer Interaction* 36, 3 (2020), 251–270.

[70] Xin Peng, Jingxiao Gu, Tian Huat Tan, Jun Sun, Yijun Yu, Bashar Nuseibeh, and Wenyun Zhao. 2016. CrowdService: serving the individuals through mobile crowdsourcing and service composition. In *Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering*. 214–219.

[71] Jenny Preece, Yvonne Rogers, Helen Sharp, David Benyon, Simon Holland, and Tom Carey. 1994. *Human-computer interaction*. Addison-Wesley Longman Ltd.

[72] Alexander J Quinn and Benjamin B Bederson. 2011. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1403–1412.

[73] Veronica A Rivera and David T Lee. 2021. I Want to, but First I Need to: Understanding Crowdworkers' Career Goals, Challenges, and Tensions. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–22.

[74] João GP Rodrigues, Ana Aguiar, and João Barros. 2014. Sensemycity: Crowdsourcing an urban sensor. *arXiv preprint arXiv:1412.2070* (2014).

[75] Holly Rosser and Andrea Wiggins. 2018. Tutorial designs and task types in zooniverse. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 177–180.

[76] Virpi Roto. 2005. Browsing on mobile phones. *Nokia Research Center* 10 (2005), 2005.

[77] Saiph Savage, Chun Wei Chiang, Susumu Saito, Carlos Toxtli, and Jeffrey Bigham. 2020. Becoming the super turker: Increasing wages via a strategy from high earning workers. In *Proceedings of The Web Conference 2020*. 1241–1252.

[78] N Sadat Shami, Gilly Leshed, and David Klein. 2005. Context of use evaluation of peripheral displays (CUEPD). In *IFIP Conference on Human-Computer Interaction*. Springer, 579–587.

[79] Kim Bartel Sheehan. 2018. Crowdsourcing research: data collection with Amazon's Mechanical Turk. *Communication Monographs* 85, 1 (2018), 140–156.

[80] M Six Silberman, Lilly Irani, and Joel Ross. 2010. Ethics and tactics of professional crowdwork. *XRDS: Crossroads, The ACM Magazine for Students* 17, 2 (2010), 39–43.

[81] Robert Simpson, Kevin R Page, and David De Roure. 2014. Zooniverse: observing the world's largest citizen science platform. In *Proceedings of the 23rd international conference on world wide web*. 1049–1054.

[82] Mads Soegaard. 2015. Interaction styles. *The glossary of human computer interaction* (2015).

[83] Saiganesh Swaminathan, Kotaro Hara, and Jeffrey P Bigham. 2017. The crowd work accessibility problem. In *Proceedings of the 14th International Web for All Conference*. 1–4.

[84] Jaime Teevan. 2016. The future of microwork. *XRDS: Crossroads, The ACM Magazine for Students* 23, 2 (2016), 26–29.

[85] Jaime Teevan, Shamsi T Iqbal, and Curtis Von Veh. 2016. Supporting collaborative writing with microtasks. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2657–2668.

[86] Peter Thomas and Robert D Macredie. 2002. Introduction to the new usability.

[87] Carlos Toxtli, Angela Richmond-Fuller, and Saiph Savage. 2020. Reputation Agent: Prompting Fair Reviews in Gig Markets. In *Proceedings of The Web Conference 2020*. 1228–1240.

[88] Khai N Truong, Thariq Shihipar, and Daniel J Wigdor. 2014. Slide to X: unlocking the potential of smartphone unlocking. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 3635–3644.

[89] Stephen Uzor, Jason T Jacques, John J Dudley, and Per Ola Kristensson. 2021. Investigating the Accessibility of Crowdwork Tasks on Mechanical Turk. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[90] Aditya Vashistha, Pooja Sethi, and Richard Anderson. 2017. Respeak: A voice-based, crowd-powered speech transcription system. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 1855–1866.

[91] Viswanath Venkatesh and Venkataraman Ramesh. 2006. Web and wireless site usability: Understanding differences and modeling use. *MIS quarterly* (2006), 181–206.

[92] E Vogels. 2021. Digital divide persists even as Americans with lower incomes make gains in tech adoption. *Pew Research Center* (2021).

[93] Liang Wang, Zhiwen Yu, Qi Han, Dingqi Yang, Shirui Pan, Yuan Yao, and Daqing Zhang. 2020. Compact Scheduling for Task Graph Oriented Mobile Crowdsourcing. *IEEE Transactions on Mobile Computing* (2020).

[94] Matthijs J Warrens. 2011. Cohen's kappa is a weighted average. *Statistical Methodology* 8, 6 (2011), 473–484.

[95] Alex C Williams, Harmanpreet Kaur, Shamsi Iqbal, Ryen W White, Jaime Teevan, and Adam Fourney. 2019. Mercury: Empowering Programmers' Mobile Work Practices with Microproductivity. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 81–94.

[96] Alex C Williams, Gloria Mark, Kristy Milland, Edward Lank, and Edith Law. 2019. The perpetual work life of crowdworkers: How tooling practices increase fragmentation in crowdwork. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–28.

[97] Xiaocan Wu, Danlei Huang, Yu-E Sun, Xiaofei Bu, Yu Xin, and He Huang. 2017. An efficient allocation mechanism for crowdsourcing tasks with minimum execution time. In *International Conference on Intelligent Computing*. Springer, 156–167.

[98] 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*. 347–348.

[99] Dongsong Zhang and Boonlit Adipat. 2005. Challenges, methodologies, and issues in the usability testing of mobile applications. *International journal of human-computer interaction* 18, 3 (2005), 293–308.

[100] Guido Zuccon, Teerapong Leelanupab, Stewart Whiting, Emine Yilmaz, Joemon M Jose, and Leif Azzopardi. 2013. Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems. *Information retrieval* 16, 2 (2013), 267–305.

[101] Kathryn Zyskowski, Meredith Ringel Morris, Jeffrey P Bigham, Mary L Gray, and Shaun K Kane. 2015. Accessible crowdwork? Understanding the value in and challenge of microtask employment for people with disabilities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 1682–1693.