Comparing Generic and Community-Situated Crowdsourcing for Data Validation in the Context of Recovery from Substance Use Disorders

Sabirat Rubya
rubya001@umn.edu
University of Minnesota
Minneapolis, Minnesota

Joseph Numainville
numai004@umn.edu
University of Minnesota
Minneapolis, Minnesota

Svetlana Yarosh
lana@umn.edu
University of Minnesota
Minneapolis, Minnesota

ABSTRACT

Targeting the right group of workers for crowdsourcing often achieves better quality results. One unique example of targeted crowdsourcing is seeking community-situated workers whose familiarity with the background and the norms of a particular group can help produce better outcome or accuracy. These community-situated crowd workers can be recruited in different ways from generic online crowdsourcing platforms or from online recovery communities. We evaluate three different approaches to recruit generic and community-situated crowd in terms of the time and the cost of recruitment, and the accuracy of task completion. We consider the context of Alcoholics Anonymous (AA), the largest peer support group for recovering alcoholics, and the task of identifying and validating AA meeting information. We discuss the benefits and trade-offs of recruiting paid vs. unpaid community-situated workers and provide implications for future research in the recovery context and relevant domains of HCI, and for the design of crowdsourcing ICT systems.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in collaborative and social computing**.

KEYWORDS

Crowdsourcing, Alcoholics Anonymous, community-situated crowd

ACM Reference Format:
Sabirat Rubya, Joseph Numainville, and Svetlana Yarosh. 2021. Comparing Generic and Community-Situated Crowdsourcing for Data Validation in the Context of Recovery from Substance Use Disorders. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3411764.3445399

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445399

1 INTRODUCTION

In recent years crowdsourcing systems have been widely used in both industry and academia for collecting human-labeled data. Generic crowdsourcing platforms like Amazon Mechanical Turk attract a large number of workers. However, since enrollment on such platforms does not require any particular skill set from the workers, ensuring the quality of the crowdsourced data is often a challenge. Attracting workers who are likely to have the skills needed for the target task may be useful to overcome this problem. Previous research has shown effectiveness of crowdsourcing with the "right workers" (e.g., who are present at a particular location or time, or have familiarity with or expertise in performing particular tasks) [5, 17, 42, 86]. Building and expanding on this idea, we aim to understand if the unique perspectives and contextual knowledge from members of a community relevant to the target task can help achieve better quality results in crowdsourcing.

Potential community members can be recruited from different online platforms or communities, with or without payment. Unpaid crowdsourcing leads to a potentially unpredictable workforce and indeterminable task completion time due to the lack of financial incentive [8]. Paid tasks, on the other hand, can attract crowd workers who may falsely report their skill, confidence, or expertise required to complete the tasks [13, 87]. In other words, there are performance trade-offs in crowdsourcing in terms of task completion time, cost, and accuracy, while applying different techniques to recruit generic vs. community members from the same paid platform, and community-situated crowd through paid vs. unpaid platforms. We contribute an empirical understanding of these trade-offs that bear significance in a variety of contexts where applying specific recruitment techniques or targeting specific platforms to access potential community crowd workers can provide better accuracy for completing tasks. For instance, members from an online community of autism caregivers can provide more concise and useful information and advice to people with autism to cope with everyday challenges, due to having more experience than the non-members. Similarly, citizen science projects can benefit from crowdsourcing data validation from online platforms consisting of people who have domain expertise. In this paper, we analyze these performance trade-offs by answering the following research questions:

- ***RQ1:*** How do self-reported community-situated crowd workers perform compared to generic crowd workers in terms of task completion time, accuracy, and cost?
- ***RQ2:*** What are the benefits and trade-offs of recruiting potential community-situated crowd workers with or without payment and screening techniques?

We analyze these trade-offs in the context of data validation tasks to create a peer-support meeting list for Alcoholics Anonymous (AA). AA has over 1.5 million members and hosts over 100 thousand weekly groups worldwide [22]. While attending meetings is particularly important for people who are newly sober, newcomers often have trouble finding reliable information about meeting locations and times due to a lack of a global up-to-date meeting list, a problem that arises due to a preference for regional autonomy in AA's organizational structure [72, 83]. Previously, researchers attempted to make the meeting information available and up-to-date in a "global meeting list" through detection of regional AA websites containing meetings and the extraction of day, time, and address of meetings from those sites. They adopted a human-aided information retrieval approach for this purpose where automated machine learning and pattern detection approaches extracted information about possible meetings listed on different AA websites. Then, these extracted results were validated through paid human workers [73]. We refer to this crowdsourcing technique, where a wide variety of "crowds" (without a particular set of assumed skills) can be recruited online or offline to complete tasks, as "generic crowdsourcing." However, a major source of error in this technique came from the workers providing poor quality answers and failing to identify webpages with meetings listed on them and content that provided information about other AA events rather than actual AA meetings.

Members of AA, who attend the meetings, may have a unique perspective on particular norms and traditions rooted in the program. In addition, they are familiar with the organization and content of the AA websites. However, approximately 1.3 million US residents (0.41% of the US population) are AA members, and a random pool of workers from an online platform may contain very few community members. The poor performance of the generic crowd workers may be attributed to the lack of contextual knowledge. We aim to understand if the unique perspectives and contextual knowledge of the AA members and their intrinsic motivation to help their peers can achieve better quality results to crowdsource accurate meeting information. We define "the technique of seeking potential workers whose membership in particular communities of practice provides them with a unique perspective on relevant background and norms of that community" as "community-situated crowdsourcing".

To compare generic and community-situated crowdsourcing, we recruit workers with three different techniques: 1) crowd workers from a paid online crowdsourcing platform (Amazon Mechanical Turk) without any particular screening techniques (paid unfiltered generic), 2) crowd workers from the same paid platform who self-report as 12-step fellowship members and are able to answer a few basic screening questions (paid filtered self-reported community-situated), and 3) members from an online community for recovery from substance abuse (unpaid filtered special community members), and evaluate these techniques of filtering crowdworkers in terms of time, accuracy and cost. We found that community-situated workers recruited from an online community can achieve better accuracy, though it may take substantially longer to recruit community-situated crowd workers from such a community, particularly if they are recruited as unpaid volunteers. Additionally, we found evidence in our data showing that the community-situated workers' expertise and familiarity with the context may have helped them achieve better accuracy.

HCI research often seeks out crowd workers with relevant experiences to address empirical questions (e.g., [17, 34]) or to create novel crowd-based systems (e.g., [38, 46]). Using the specific context of crowd workers in 12-step recovery, we empirically compare the effectiveness and costs of three different approaches for situated crowd work, providing concrete recommendations for which approach may be most appropriate based on a particular task. Our work provides implications for effectively filtering and utilizing community-situated crowd workers that may be applicable in other HCI contexts and discusses these implications for research and design of crowd-based ICT systems.

2 RELATED WORK

2.1 Situated Crowdsourcing

While online crowdsourcing markets make it convenient to pay for workers willing to solve a range of different tasks, they suffer from limitations such as not attracting enough workers with desired background or skills [4, 21, 54]. For example, it can be a challenge to recruit workers for a task that requires workers who speak a specific language or who live in a certain city [10, 42] . Situated crowdsourcing can help fill in the gaps in these scenarios where the crowd needs to be associated with some context. Situated crowdsourcing consists of embedding input mechanisms (e.g., public displays, tablets) into a physical space and leveraging users' serendipitous availability [30] or idle time ("cognitive surplus") [77]. It allows for a geo-fenced and more contextually controlled crowdsourcing environment, thus enabling targeting certain individuals, leveraging people's local knowledge or cognitive states, or simply reaching an untapped source of potential workers [27–29, 34]. Researchers have discussed benefits of targeting geographically or temporally situated crowds over generic crowds in scenarios like providing emergency services in disasters [56], geotagging photos [42], etc. An experiment by Ipeirotis et al. automatically identified "situated crowds" with desired competence to complete a task and demonstrated that the cost of hiring workers through their platform is less than that of hiring workers through paid crowdsourcing platforms [38]. However, these previous works do not focus on situated crowdsourcing in terms of the competence of the workers for the target task resulting from being members of a particular community, nor do they consider recruiting the targeted workers from different platforms. We extend the idea of situated crowdsourcing to capture the idea of selecting the "right crowd" better suited for a task (e.g., with a particular skill or quality) due to their familiarity with the context, increased reliability, and better quality results [15, 54].

Most prior work focuses on temporally-situated or locally-situated participants, who are at a time or place to best be able to do the task. Building on these ideas, we refer to the process of recruiting potential crowd-workers who are members of particular communities who may be motivated to produce better results without monetary incentives as "community-situated" crowdsourcing. Some examples include detecting and reporting predatory publishers through crowdsourcing from the community of authors and

researchers [2], or getting feedback for course projects on a particular course from a community of freelance experts on that topic [86]. In the context of this paper, members of 12-step communities like AA are familiar with the format of recovery meetings and governing structure meetings. A major source of error in previous work with crowdsourcing AA meeting validation [73] came from workers being unable to distinguish area or district meetings from open weekly AA meetings. However, we can assume that actual members of this community would not make the same mistakes, thus being the "right crowd" with the required skill (familiarity with the context) for the tasks of 12-step meeting identification and validation.

The primary contribution of our work is empirically comparing two alternative approaches for community-situated crowdsourcing. The secondary contribution is quantitatively establishing the benefits of community-situated crowdsourcing.

2.2 Community-Situated Crowdsourcing: Generic Platforms versus Targeted Online Communities

There are multiple reasonable approaches for recruiting a community-situated crowd. When using a generic platform like Amazon Mechanical Turk, membership in a particular community can be self-reported, which can serve as a qualification for receiving the task, although this process does not ensure that all the recruited participants are community members. Alternatively, one may be able to solicit members of a desired group directly through online spaces dedicated to those communities (e.g., Facebook recovery groups, InTheRooms.com).

Prior work has incorporated elements of both approaches, but without systematically separating and comparing the two. For example, friendsourcing leverages one's social network to gather information that might be unavailable or less trustworthy if obtained from other sources [6, 7, 58, 75]. Researchers have attempted to build systems that, for example, use friendsourcing for social tagging of images and videos [7], to seek out personalized recommendations, opinions, or factual information [7, 62], to help blind communities get answers to questions about the world around them captured by cameras [12], or to provide cognitive aids to people with dementia [58]. Friendsourcing removes the financial cost of the service and improves the quality and trustworthiness of the answers received [31, 63]. Several studies applying friendsourcing in health contexts suggest that friendsourcing can encourage engagement and provide more emotional and informational support compared to generic crowdsourcing [6]. This is a specific example of leveraging a known community and soliciting members through existing or novel online channels to complete desired tasks.

As another example, altruistic crowdsourcing refers to cases where unpaid tasks are carried out by a large number of volunteer contributors [10]. This form of crowdsourcing often utilizes members of the same community to complete collective tasks or getting better quality and more trustworthy information [10, 27]. However, some situations require altruistic contribution from outgroup crowd workers, for example, in providing directions to people with visual impairments [11], or in generating valuable daily life advice for people with autism [33]. Altruistic crowdsourcing may leverage existing social networks or may serve as a filter when seeing contributions from the larger community (in the sense that those willing to complete unpaid tasks in a particular context are likely to have a personal connection or interests in that context). Both friendsourcing and altruistic crowdsourcing approaches rely heavily on social motivators such as social reciprocity, the practice of returning positive or negative actions in kind, reinforcing social bonds, the opportunity of showing off expertise, etc. [7, 12, 15].

Systems that utilize members of the same community or known channels as the crowd often leverage their intrinsic interest in a particular domain and their sense of belonging to the community [46, 76]. This form of crowdsourcing has proven to receive better quality feedback for class projects from classroom peers [16, 80]. Moreover, researchers have discussed the efficacy of crowdsourcing peer-based altruistic support in critical contexts, such as to reduce depression and to promote engagement [64], for generating behavior change plans [1, 17], to exchange health information [71], etc. Although friendsourced answers often contain personal or contextual information that improves their quality, in some cases people do not consider social network as an appropriate venue for asking questions due to high perceived social costs, limiting the potential benefit of friendsourcing [6, 12, 75].

This body of work provides compelling examples of benefits of community-situated crowdsourcing but is largely opportunistic about how such a situated crowd is recruited. Obviously, friendsourcing could not be reasonably accomplished through filtering workers on generic platforms. However, in most other cases, both qualifying members on existing platforms and targeting specific online communities may be reasonable approaches and may have different costs and benefits. Our primary contribution is providing an empirical comparison of accuracies, costs, and time trade-off in these two approaches to community-situated crowdsourcing, by comparing tasks completed by workers on MTurk who self-identify as members of 12-step programs and those done by members of the targeted online recovery community InTheRooms. Based on our findings, we provide recommendations for community-situated crowdsourcing in other contexts.

2.3 Comparison of Different Types of Workers

Both CHI and CSCW communities have studied the comparison among different types of crowdsourcing based on magnitudes of financial incentives [40, 57, 84], worker motivations [41, 69], and worker expertise [17, 70, 82].

Many initial crowdsourcing studies focused on finding out how the magnitude of financial incentives affects the work produced. While some of the studies suggested that worker quantity may increase with higher incentives for the same task but the quality of the results do not improve [55, 57, 59], others pointed out that the amount of monetary incentives can produce better quality work if it is performance-contingent (e.g., rewards or penalties) [84]. On the other hand, unpaid crowd workers provide a workforce without labor cost and can work as an economical alternative for individuals and organizations who are concerned about a budget [8, 9]. Researchers pointed out that volunteers often provide as reliable and high quality answers as paid workers, though the turnaround time may be higher for unpaid workers and they are more likely to

not complete the tasks when compared to their paid counterparts [8, 45, 57, 68], making the use of crowdsourcing through volunteers questionable for time-sensitive tasks.

Since non-expert crowd workers are often more cost-effective than expert ones, another body of research studied the performances of crowdsourcing by experts and non-experts. They suggested that non-expert work quality may be comparable to the experts [44, 60] in many contexts, including tasks as difficult as identifying a particular type of bio signal (sleep-spindle) from raw data [79]. Moreover, research on leveraging these types of workers in citizen science and web security suggested that their complementary roles and different potentials in different tasks should be better capitalized by community-based systems across different domains [14, 32, 82].

While some of these research works discuss the benefits of recruiting expert and reliable volunteers from specific communities, they do not explicitly compare performances of these participants recruited with different approaches or from different platforms. In peer-support communities for critical health conditions, there may be urgency regarding time and accuracy of the produced results for crowdsourcing tasks [33, 81]. We explicitly compare community-situated paid and unpaid crowdsourcing in the high-impact context of recovery and discuss trade-offs that can be applied to other similar contexts.

3 METHODS

We conducted an experiment to understand the trade-offs of generic vs. community-situated crowdsourcing. For generic crowdsourcing, we recruited workers from Amazon Mechanical Turk (MTurk) and for community-situated crowdsourcing we solicited potential community members from both a filtered subset of MTurk workers and an online community for people in recovery.

3.1 AA Meeting Dataset

The steps of developing a comprehensive AA meeting list include identifying different regional webpages on different domains that contain one or multiple AA meetings and locating all the meetings and their corresponding information (e.g., day, time, and address) on those meeting pages. To create such a list, researchers previously paired information retrieval with human computation [73] and collected *a*) ground truth data labels for 964 webpages, each with a label of 1 or 0 indicating whether it contains a list of meetings or not, and *b*) location of 1892 meetings from a subset of the meeting pages. This subset was selected from 18 regional websites from three different states in the USA. We got access to this data from prior work and saved it in a MySQL database on a secure server.

3.2 Participants

We recruited three types of participants: 1) crowd workers from MTurk without any particular screening techniques (generic MTurk), 2) crowd workers from MTurk who self-report as 12-step fellowship members and are able to answer our screening questions (community-situated MTurk), and 3) volunteers from an online community for recovery from substance abuse (community-situated InTheRooms). Since we already had data for the first population (generic MTurk) from our previous study and we wanted to minimize the cost of recruiting workers, we performed a power analysis. We found that for our statistical comparisons the desired sample sizes (the minimum number of participants) for the other two populations (community-situated MTurk and community-situated InTheRooms) should be 400 to achieve a power of 0.8 at the 95% confidence level.

3.2.1 Generic Crowd Workers from MTurk. Amazon Mechanical Turk or MTurk[1] is a popular online crowdsourcing marketplace where "requesters" can hire remotely located "crowd workers" (MTurkers) to perform discrete on-demand tasks known as Human Intelligence Tasks (HITs). The "generic crowd" sample in our previous work consisted of a total of 1060 individuals recruited from MTurk [73]. A HIT with the background and description of the study was created with a link of the website to view and agree to an online informed consent and to complete the tasks. After they completed all three tasks, they were provided with a unique survey code which they had to copy and paste on the MTurk platform so their responses could be tracked and associated with corresponding worker IDs to approve or disapprove their HITs. Participants were recruited during the month of May 2018.

3.2.2 Filtered Community-situated Crowd Workers from MTurk. For recruiting and filtering MTurk workers who are practicing their recovery through one or more 12-step fellowships, we designed a screening survey and explicitly mentioned that only members of 12-step programs should attempt the questions. The screening survey consisted of three basic questions about recovery practices in 12-step fellowships (see below). 660 workers attempted the survey, and 480 of them answered all three questions correctly. We accepted all the HITs for the screening survey, but only these 480 workers were marked as "qualified" workers. The actual HIT with the link to the tasks (described in Section 3.3) was made visible only to the qualified workers. Participants were recruited during the month of December 2019.

Screening Survey Questions: The screening survey included the following three questions about recovery in 12-step fellowships (Fig 2 in supplementary document):

- What is your primary 12-step fellowship?
- How many meetings do you attend weekly?
- Fill in the blank: The 12th tradition says, "_____ is the spiritual foundation of all our Traditions, ever reminding us to place principles before personalities."

Ensuring Quality of the Workers: Previous work has pointed out that paid crowd workers recruited from online platforms often produce low quality results [25] to maximize their earnings by completing tasks quickly. In order to assure high quality recruitment of participants, we used several strategies. First, the recruitment was limited to MTurkers whose average HIT acceptance rate was 90% or higher. We embedded a custom script on the HIT page that limits the number of times that a single worker may work on this study. MTurkers could take part in this survey only once, reducing the effect of noise in the results from worker experience over repeated tasks. Lastly, all the response patterns were reviewed on a daily basis. If a respondent completed our survey too quickly

[1] https://mturk.com

(i.e., speeders who took less than five minutes in total starting from reading the consent form to finishing the last task) without carefully reading the question items, the responses from that respondent were excluded from the analysis. Also, if a respondent rushed through the tasks by clicking on the same response to multiple questions (i.e., straight-liners for all ten different answers for a particular task), those cases were also filtered out from data analysis. Such data cleaning procedures resulted in an unequal number of participants for three different tasks. For community-situated MTurkers, the sample sizes for page validation, meeting validation, and meeting identification were 423, 406, and 435 respectively.

It is an important point to note again that the term "community-situated crowdsourcing" is a shorthand for the process of seeking workers who are members of particular groups or communities related to the context of the research questions. All the community-situated workers in our study have *self-reported* themselves to be members of 12-step fellowships. We cannot, however, make a certain claim about all of them actually being in recovery. We kept the screening survey questions simple so it does not take too long to recruit the target number of "self-reported filtered community-situated crowd workers" from MTurk.

3.2.3 Community-situated Crowd Workers from ITR. InTheRooms.com (ITR)[2] is the largest online community for recovering addicts, and their friends and families, hosting over 500 thousand members. It hosts over 100 weekly online video-meetings and provides social features for members to connect with and provide support to each other (e.g., profiles, wall posts, and comments).

We worked with the ITR website founders and owners to reach out directly to ITR members to participate in the study. A link to the survey was distributed as a paid banner advertisement on the ITR homepage for two months (October 2019-November 2019), as well as advertised in their weekly ITR newsletter with a message from the researchers explaining the purpose of the study (Fig. 1 of supplementary document). Similar to the workers from MTurk, all ITR members who responded to the advertisement viewed and agreed to an online informed consent.

Again, the community-situated crowd workers from ITR were recruited based on the assumption that the online community members are in fact 12-step fellowship members, which may not be true for all members of ITR, since online communities may have spammers or non-members [24, 78]. However, based on a few assumptions from our research experience with this community for past couple years, we think that unlike MTurk, ITR participants are more likely to be actual 12-step fellows, and hence we did not ask the filtering questions to the ITR members. First, ITR is primarily an online peer support network for 12-step fellowship members and their friends and families. A spammer in this community does not have much to gain from frequent interactions with others. Second, even if we assume that there is a considerable number of non-12 step fellowship members in the community, the probability of them responding to a study without any monetary reward should be very low. Finally, since we were not providing any reward to the ITR members for participating in the survey, providing additional questions for screening might have reduced the turnover of the responses, increasing the time and the cost of recruitment. According to [18], we ideally reached approximately 10,000-25,000 MTurk workers. We accepted 440 responses for the screening, which is about 1.8-4.4% of the estimated active Turkers. Hence, the estimated percentage of potential false reports is 1.4-3.9%. With a conservative calculation, it converts to a maximum of 18 participants who falsely self-reported to be AA members.

It is critical to protect the ITR members from exploitation. In the context of 12-step fellowships, members carry out service to their programs without any financial reward and payment for such services would be in violation of the program traditions. The eventual goal is to build a self-updating meeting list system that would provide value to people using it, and therefore engender people to volunteer small units of their time to keep it available for themselves and others. This is in line with how 12-step programs are currently structured.

3.2.4 Ethical Considerations. Recovery from substance use is a personal and private undertaking for most people and we took steps to consider the ethical implications of our work and to protect the rights of the community-situated workers who are members of different 12-step programs. All research activities on this project were reviewed and approved by our university IRB as an investigation where potential benefits of the scientific work outweighed the risks to the participants. The approval covered two phases. First, all users who completed the tasks viewed and responded to an online informed consent form, but documentation of informed consent was waived in order to preserve participants' anonymity. Second, we made sure that the screening survey and the tasks do not ask the participants for any personally identifiable information.

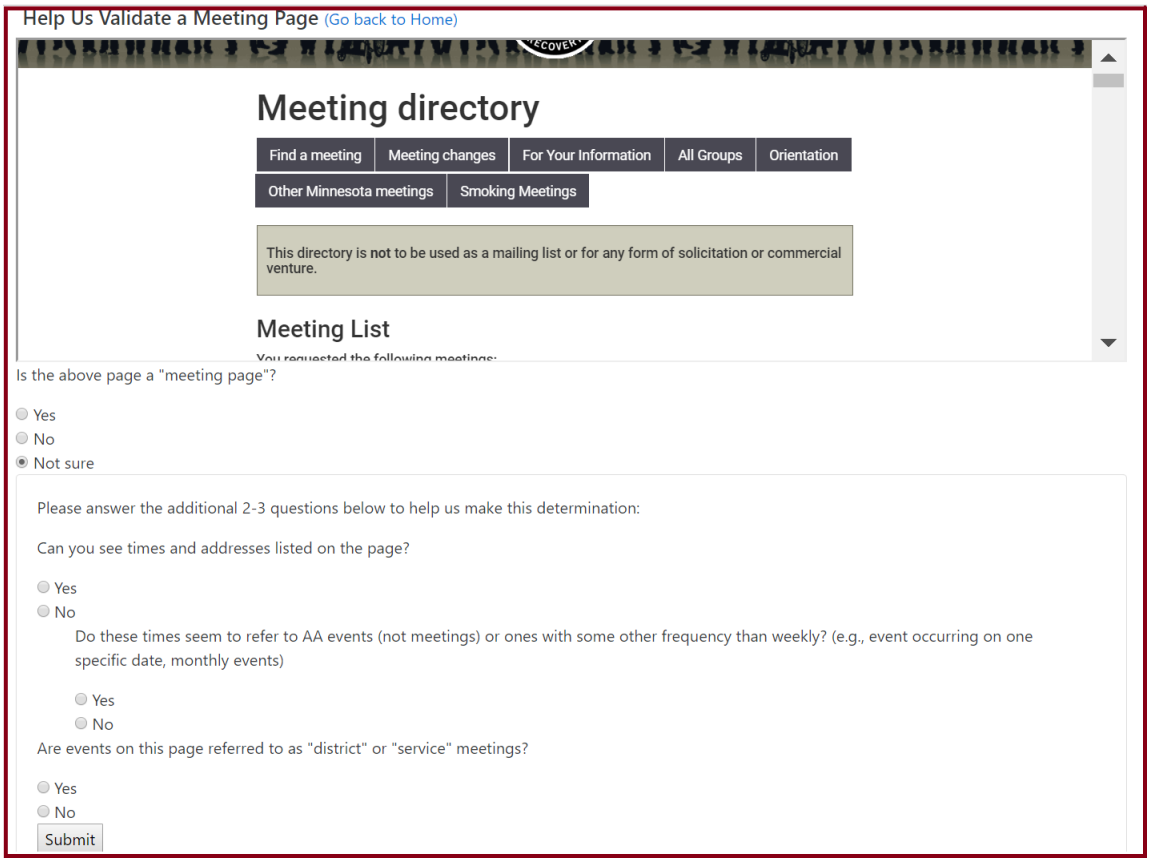
3.3 Tasks

We created responsive web interfaces for three different tasks using the Python Flask framework. All workers were assigned to complete all three tasks, but we randomized the order of the tasks shown to them to reduce the learning effect on the results. From both MTurk and ITR platforms, they were redirected to the task website and were shown a consent form. Documentation of informed consent was waived in order to preserve participant anonymity. Prior to starting each task, there was a popup page with detailed instructions and examples (See supplementary document).

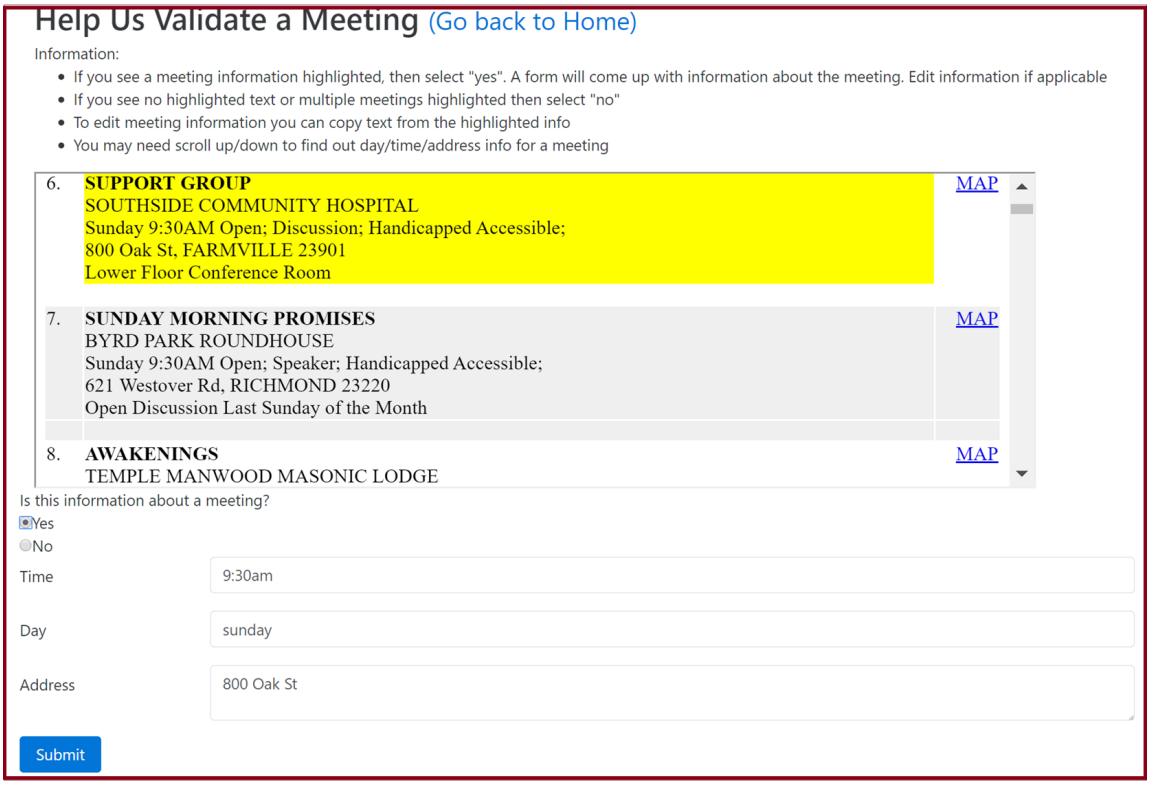
3.3.1 Meeting Page Validation. The interface for page validation showed 10 webpages (selected randomly from the ground truth data set of pages) sequentially and asked the worker if it was a meeting page with "yes", "no", and "not sure" options. If workers selected "not sure," they had to answer two or three additional questions to help us determine the label (Fig. 1a).

3.3.2 Meeting Information Validation. The interface sequentially showed 10 meetings (selected randomly from the ground truth data set of meetings) highlighted in yellow on corresponding webpages and asked the worker if it was a meeting record, with "yes" and "no" options (Fig. 1b). If workers selected "no" for a meeting, they advanced to the next record. Otherwise, they were prompted to edit the automatically extracted details of the meeting if these did not match with the highlighted record.

[2]https://intherooms.com


Help Us Validate a Meeting Page (Go back to Home)
Meeting directory
Find a meeting
Meeting changes
For Your Information
All Groups
Orientation
Other Minnesota meetings
Smoking Meetings
This directory is not to be used as a mailing list or for any form of solicitation or commercial venture.
Meeting List
Is the above page a "meeting page"?
Yes
No
Not sure
Please answer the additional 2-3 questions below to help us make this determination:
Can you see times and addresses listed on the page?
Yes
No
Do these times seem to refer to AA events (not meetings) or ones with some other frequency than weekly? (e.g., event occurring on one specific date, monthly events)
Yes
No
Are events on this page referred to as "district" or "service" meetings?
Yes
No
Submit

(a)


Help Us Validate a Meeting (Go back to Home)
Information:
If you see a meeting information highlighted, then select "yes". A form will come up with information about the meeting. Edit information if applicable
If you see no highlighted text or multiple meetings highlighted then select "no"
To edit meeting information you can copy text from the highlighted info
You may need scroll up/down to find out day/time/address info for a meeting
6. SUPPORT GROUP
SOUTHSIDE COMMUNITY HOSPITAL
Sunday 9:30AM Open; Discussion; Handicapped Accessible;
800 Oak St, FARMVILLE 23901
Lower Floor Conference Room
MAP
7. SUNDAY MORNING PROMISES
BYRD PARK ROUNDHOUSE
Sunday 9:30AM Open; Speaker; Handicapped Accessible;
621 Westover Rd, RICHMOND 23220
Open Discussion Last Sunday of the Month
MAP
8. AWAKENINGS
TEMPLE MANWOOD MASONIC LODGE
MAP
Is this information about a meeting?
Yes
No
Time
9:30am
Day
sunday
Address
800 Oak St
Submit

(b)

(c)

Figure 1: Interfaces of the three tasks (a) meeting page validation, (b) meeting validation, and (c) meeting identification

3.3.3 Meeting Information Identification. The interface sequentially showed 10 meeting pages (selected randomly from the ground truth data set of pages) with the retrieved meetings highlighted in yellow and asked the workers if they noticed any meeting as not highlighted, with "yes" and "no" options. If workers selected "no" for a page, they were advanced to the next page. Otherwise, they were asked to draw a rectangle around any one unhighlighted meeting (Fig. 1c).

Prior to recruitment, we conducted several pilot experiments: two on MTurk and two with undergraduate volunteers, to refine the interfaces of the tasks and task instructions and descriptions, to determine the appropriate pay per worker ($2.5), and to define a reasonable minimum time of task completion (five minutes).

3.4 Platform Cost

Our research questions are related to calculating and comparing the cost of task completion for different worker types, that is the total cost paid to the workers only for completing the tasks, since this fraction of the total cost would more likely influence the workers' task completion time and accuracy. However, in the interest of keeping the discussion about cost of the task completion consistent, we separate the platform cost that was associated with the use of particular crowdsourcing platforms to recruit workers, from the actual worker cost that was the amount of money that went to the workers who completed all the tasks. We acknowledge that in reality the required number of participants might not have been obtained without the advertising that caused the platform cost for recruiting from ITR, and we discuss the total and average worker cost including all the expenses spent for the study (Section 4.1).

On MTurk, the price a requester has to pay for a HIT is comprised of two components: the amount to pay workers plus a fee to pay MTurk. The usual MTurk fee is 20% of the worker fee for any HIT. For each added worker qualification in this study (i.e., workers with greater than 90% HIT acceptance and workers with correct answers on the screening survey) an additional 5% of the worker fees was added to the platform cost.

3.4.1 MTurk cost for generic crowdsourcing. For generic crowdsourcing on MTurk, the researchers paid a total platform cost of $543.75 (25% of the total worker cost described in the Results Section). This was a sum of the usual 20% of the worker cost and the 5% fee for an additional criterion of making the HIT visible only to workers having average HIT acceptance rate of 90% or higher.

3.4.2 MTurk cost for community-situated crowdsourcing. For community-situated crowdsourcing on MTurk, the platform cost was a total of $274.95, including the MTurk fees of $4.95 for the screening survey and $270 for the actual HIT.

3.4.3 ITR cost for community-situated crowdsourcing. We paid $6000 to the founders of InTheRooms to run the advertisement for two months. In addition to putting up the banner on their homepage and on the website's weekly newsletters, the website founders guided us in improving the banner design and the task interfaces to attract more participants, and provided continuous support in recruiting participants through batch emailing community members about the research asking for their participation. For other domains, however, the platform cost would largely depend on the type of platform to recruit participants from and would vary based

on the type of the tasks in question and the required community to perform them.

3.5 Measures

For comparing the performance of different types of crowdsourcing, the variables we calculated and the statistical tests we conducted are as follows:

- Cost: We calculated unit and total costs paid to the workers and to the platforms.
- Time of recruitment: We report the amount of time required to recruit the required number of valid participants from each platform.
- Average accuracy and average task completion time: For the three different groups of workers, generic crowd, community-situated crowd from MTurk, and community-situated crowd from ITR, we at first conducted Kruskal-Wallis tests for mean accuracy and mean completion time for page validation, meeting validation, and meeting identification tasks to find out if there was a difference between at least one group and the other groups of workers. Afterwards, we performed between-subjects Wilcoxon Rank-Sum Tests (as some of those distributions were not normal) for only the tasks that had p-values less than 0.05 from the Kruskal-Wallis tests to calculate the difference in mean accuracy and mean completion time for the three pairs: generic crowd and community-situated crowd from MTurk, generic crowd from MTurk and community-situated crowd from ITR, and community situated crowd from MTurk and ITR. To account for multiple comparisons, we adjusted the p-values using the Holm-Bonferroni method. Since the type of data to identify and validate was different in each task, we calculated the accuracy differently.

$$\text{Acc}_{pg\_validation} = \frac{\text{\# of pages correctly answered with yes/no}}{\text{total\#of pages validated}}$$

$$\text{Acc}_{mtng\_validation} = \frac{\text{\# of meetings correctly validated}}{\text{total \#of meetings attempted}}$$

$$\text{Acc}_{mtng\_identification} = \frac{\text{\# of meetings correctly identified}}{\text{total\#of meeting pages attempted}}$$

- To compare the performances for specific portions of the tasks that require workers' contextual knowledge, we conducted appropriate statistical analyses. For instance, to calculate the difference in average number of meetings validated with a "not sure" option, we conducted a Chi-Squared test, where the categories were represented as whether a particular page was classified with a "not sure" option or not by a particular type of crowd worker.

4 RESULTS

We analyzed the performance of generic crowdsourcing and community-situated crowdsourcing with workers from two different platforms and discuss the trade-offs in this section in terms of recruitment time and cost, task completion time, and accuracy of the workers. Additionally, we were interested in understanding if community-situated crowd workers performed better in detecting and validating particular types of information due to their familiarity with the context.

4.1 Cost of the Workers

For the MTurk HIT to complete the tasks, $2.50 was paid to each worker. For generic crowdsourcing in the previous work, 870 HITs were accepted and we accepted 426 HITs from the community-situated crowd answers. Since the total number of participants in these two samples are different and there was an additional cost of $19.80 for the 660 accepted HITs of the screening survey in the MTurk community-situated crowdsourcing, there is a difference in the total worker cost between generic workers and community-situated MTurk workers. The total worker cost is $2175 ($2.50 per worker) and $1099.80 ($2.57 per worker) for generic MTurk and community-situated MTurk workers respectively. We did not pay any monetary rewards to the ITR workers. Therefore, considering the *average cost per worker*, community volunteers cost the least (zero), followed by generic crowdsourcing from MTurk, and community-situated crowdsourcing from MTurk respectively. Table 1 reports these worker costs.

Ideally, if the same number of paid generic and community-situated crowd workers are recruited to perform the same number of tasks, the cost would be about the same. In contrast to the paid community-situated MTurkers, the community-situated worker cost can potentially be very low or even zero, irrespective of the number of recruited workers, if we choose to recruit volunteers from online groups/communities relevant to the context in question. However, these calculations disregard the platform costs which can significantly increase the average cost per worker for community-situated workers in many studies similar to this one.

4.1.1 Total Cost. Even though in our study the platform cost ideally did not impact the workers' task completion time and accuracy, many studies may require high advertising costs to ensure enough participation and/or to obtain quality results. The third column of Table 1 represents the total and the average costs including both worker and platform costs. The average cost per worker for generic and community-situated crowdsourcing were $3.125 and $3.18 respectively. The additional average worker cost for community-situated MTurk is due to the HIT used for filtering workers. The average worker cost for ITR participants was $13.33 considering the platform cost, which is very expensive compared to the other two populations. Including platform cost in fact reorders the worker cost by type required for this study. However, for tasks related to other contexts/domains, or even in the context of AA, the platform cost can be reduced by recruiting workers from other platforms with large user bases. We discuss it in more detail in Section 5.3.

4.2 Time to Recruit Workers

MTurk and ITR both have over 500,000 members, although not all of them are active [66]. Paid online platforms like MTurk have been shown to be an effective way of quickly recruiting workers (both experts and novices) [9]. We found evidence for this in our study, as it took the least time to recruit generic workers from MTurk.

For generic crowdsourcing, the researchers aimed to recruit participants until they achieved validation results for a predetermined number of meeting pages and meetings [73]. They published HITs

	Worker cost only (average cost per worker)	Total worker cost including platform cost (average cost per worker)
Generic: MTurk	\$2175 (\$2.5)	\$2718.75 (\$3.125)
Community-situated: MTurk	\$1125 (\$2.5) + \$19.80 (\$0.03) = \$1099.80	\$1431 (\$3.18)
Community-situated: ITR	\$0 (\$0)	\$6000 (\$13.33)

Table 1: Worker costs and total costs for generic and community-situated crowdsourcing. Worker cost is the portion of the total cost paid to only the workers for their task completion, and not to the platform.

	Time	Accuracy
Page validation	<.001 ***	<.001 ***
Meeting validation	<.001 ***	<.001 ***
Meeting identification	.003 **	.172

Table 2: p-values from the Kruskal-Wallis tests for the differences in task completion time and accuracy among the three tasks

in batches and received the desired number of responses in about a week ($n = 1060$). For community-situated MTurkers, it took two weeks to recruit 450 participants, starting from the time the screening survey HIT was published. For the ITR advertisement, we achieved a total of 460 complete responses which was about the same number as the targeted number of participants for this comparison. It took, however, additional time for us prior to recruitment to negotiate with the website founders about the logistics of the advertisement. Therefore, time to recruit generic crowd workers was substantially less than both types of community-situated workers. Additionally, recruitment time of unpaid volunteers from ITR was about four times slower than paid community-situated workers on MTurk.

In general for other domains, recruiting volunteers from online communities might consist of similar steps and might take substantially more time than recruiting paid community-situated workers. On the contrary, other platforms with more volunteers (e.g., Facebook groups) may be leveraged to reduce this time. Therefore, the recruiting organizations have to consider the trade-off between time and number of participants for different platforms based on the emergency and frequency of the crowdsourcing task in question (further explored in the discussion section).

4.3 Task Completion Accuracy of Workers

InTheRooms volunteers completed the tasks with highest accuracy regardless of the type of the task, followed by community-situated MTurkers and generic MTurkers, respectively (refer to the boxplots in Fig. 2). Kruskal-Wallis tests showed significant difference in mean accuracy except for the task of meeting identification (Table 2). For page validation and meeting validation, the differences between generic crowd workers and the community-situated crowd workers were statistically significant at the 95% confidence level. Additionally, mean accuracy of the community-situated workers from ITR was significantly higher for page validation than community-situated workers from MTurk. For meeting validation, accuracy of both types of community-situated workers were significantly higher than generic crowd workers. We discuss these accuracies and their interpretations in more details in the following subsections.

4.3.1 Page Validation. The mean accuracy of community-situated crowdsourcing from ITR ($M = 72.1\%$, $SD = 19.18\%$) was significantly higher than both the mean accuracy of community-situated crowdsourcing from MTurk ($M = 60.52\%$, $SD = 18.95\%$) ($p < .0001$) and generic crowdsourcing from MTurk ($M = 56.40\%$, $SD = 16.24\%$)($p < .0001$) (Fig. 2a). This difference in accuracy in the context of creating an AA meeting list means that recruiting participants from online communities for recovering alcoholics can increase the accuracy of meeting page validation by 16% compared to a generic platform like MTurk, resulting in the correct label for about one in six additional meeting pages in the data set.

4.3.2 Meeting Validation. Similar to the task of page validation, the highest mean accuracy ($M = 76.94\%$, $SD = 17.33\%$) was achieved by the community-situated workers from ITR, followed by community-situated MTurkers ($M = 73.54\%$, $SD = 15.45\%$) and generic MTurkers ($M = 71.54\%$, $SD = 20.51\%$). This accuracy was significantly higher than generic crowdsourcing from MTurk ($p < .0001$) (Fig. 2b). In addition, community-situated MTurkers' accuracy was significantly higher (p=.002) than the accuracy of generic MTurkers. In the context of creating an AA meeting list, this finding implies that the ITR members can provide accurate day, time, and address information for about 5% more meetings than the generic crowd workers. AA hosts more than 50,000 meetings throughout the United States and community-situated volunteers can substantially improve the reliability of the meeting list by validating these meetings accurately.

4.3.3 Meeting Identification. Community situated ITR workers achieved the highest mean accuracy for the meeting identification task ($M = 81.80\%$, $SD = 21.24\%$), followed by community-situated MTurk workers ($M = 81.36\%$, $SD = 25.39\%$) and generic MTurk workers ($M = 79.76\%$, $SD = 23.36\%$) respectively (Fig. 2c). There was no significant difference in accuracy between any of the three pairs. We assume that, since meeting identification involved only segmenting a part of a webpage if there was a meeting that was not highlighted (i.e., drawing a rectangle around an unhighlighted

	Page validation		Meeting validation		Meeting identification	
	M	SD	M	SD	M	SD
Generic: MTtuk	237.61	162.11	318.37	255.50	268.50	212.78
Community: MTurk	315.12	181.70	555.73	332.46	309.49	313.62
Community: ITR	343.23	265.59	507.27	360.37	329.26	316.08
Generic and community:MTurk	<.001***		<.001***		.18	
Generic and community:ITR	<.001***		<.001***		.030*	
Community:MTurk and community:ITR	.97		.163		.25	

Table 3: Top three rows represent the means and standard deviations of task completion time (in seconds) of generic and community-situated crowd workers for three different tasks.Bottom three row show the statistical comparisons of average task completion time between: 1) Generic crowdsourcing and community-situated crowdsourcing from MTurk, 2) Generic crowdsourcing and community-situated crowdsourcing from MTurk, and 3) Community-situated crowdsourcing from MTurk and ITR. p-values are reported using Wilcoxon Rank-Sum Tests and adjustment with Holm-Bonferroni method. (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$)

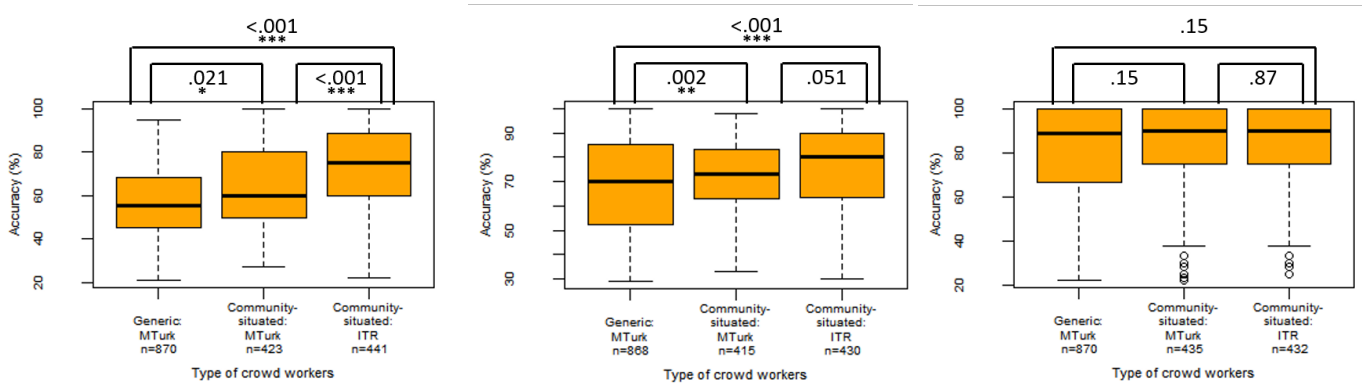


Figure 2: Boxplots of accuracy achieved by three different types of crowd workers for: (a) page validation, (b) meeting validation, and (c) meeting identification. The figure also shows the statistical comparisons of average task completion accuracy between: 1) Generic crowdsourcing and community-situated crowdsourcing from MTurk, 2) Generic crowdsourcing and community-situated crowdsourcing from MTurk, and 3) Community-situated crowdsourcing from MTurk and ITR. p-values are reported using Wilcoxon Rank-Sum Tests and adjustment with Holm-Bonferroni method. (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$)

meeting record), it is comparable to the task of segmenting images for object detection and is a common type of HIT on crowdsourcing platforms. In fact, generic crowd workers completed this task with comparable accuracy in less time on average.

4.4 Evidence for Reasons of Difference in Accuracy

We were interested in further investigation of the possible reasons for the significant difference in accuracy between community-situated crowdsourcing and generic crowdsourcing. Through post-hoc analyses and a closer look into the data, we found out that this difference in accuracy could be contributed by one or more of the following factors:

4.4.1 Task Completion Time of Workers. The average time taken to complete the tasks by generic MTurkers (page validation: $M = 237.61s$, $SD = 162.11s$, meeting validation: $M = 318.37s$, $SD = 255.50s$, meeting identification: $M = 268.50s$, $SD = 212.78s$) was less than both the community-situated MTurkers (page validation: $M = 315.12s$, $SD = 181.70s$, meeting validation: $M = 555.73s$, $SD = 332.46s$, meeting identification: $M = 309.49s$, $SD = 313.62s$), and the ITR workers (page validation: $M = 343.23s$, $SD = 265.59s$, meeting validation: $M = 555.73s$, $SD = 332.46s$, meeting identification: $M = 329.26s$, $SD = 316.08s$). Generic crowd workers completed the tasks of page validation and meeting validation significantly faster than both the paid ($p < .001$) and unpaid ($p < .001$) community-situated crowds. For meeting identification, this difference was not significant. There was no significant difference in completion time of any of the three tasks between paid and unpaid

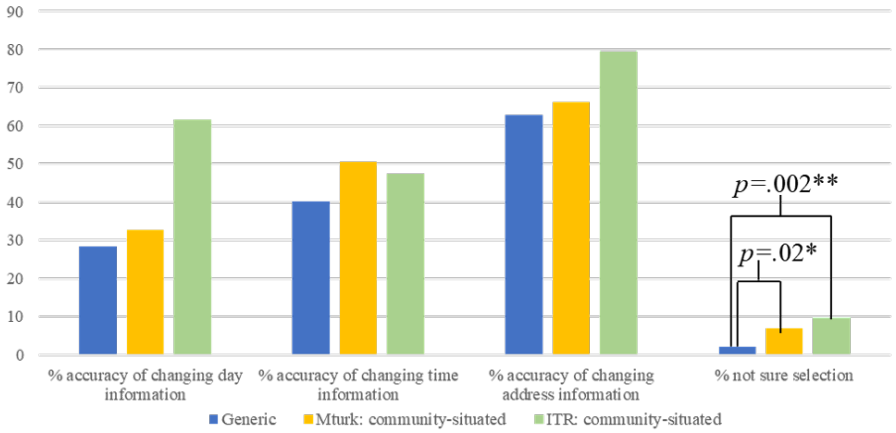


Figure 3: Different types of errors in the page validation and the meeting validation tasks; the first three groups from the left show, in order, the average accuracy of editing the day, time, and address information correctly in the meeting validation task. The rightmost group shows the the average percentage (and the differences between groups) of selecting the "not sure" option in the page validation task

community-situated worker samples (see Time columns in Table 3).

In terms of average completion time by type of task, page validation took the least amount of time for all types of workers and meeting validation took the most. This is expected, since for validating a meeting workers often have to scroll through the webpage shown to them to confirm that the time and location of the highlighted meeting is accurate, and edit the information if necessary. On the other hand, for validating a meeting page they mostly need to skim through the page shown to them for a list of meetings and determine if those are open AA meetings. Unlike editing and typing text for the meeting validation task, for page validation they had to answer one multiple choice question (or a maximum of 2-3 if they selected "not sure").

These results indicate that workers being paid for task completion were probably performing the tasks very quickly and paying minimal attention to each one, causing the lowest accuracy for generic MTurkers. This finding also resonates with results from many previous studies on paid crowdsourcing [9, 57]. For the community-situated crowd samples, volunteers from ITR usually took slightly more time on average than the paid MTurkers. This could probably be attributed to the different motivations of the community-situated crowd from these platforms: For ITR members the only reason of participation was their personal motivations to be of service to the community, whereas MTurk members were both members of 12-step fellowships (supposedly) who wanted to help other community members, and also belonged to online crowd workers seeking to maximize their earnings through completing HITs quickly.

4.4.2 Knowing When "Not Sure". We calculated the reported accuracy of page validation considering only those workers who were certain in labeling most of the webpages shown to them. To be more specific, if a worker selected "not sure" for more than three pages, we did not include his/her response to calculate the accuracy for this particular task. We assumed that the accuracy difference may occur partly because of the willingness of different types of workers in choosing the "not sure" option when they were actually not sure. To find out if this assumption was true, we calculated the average number of pages where workers selected the "not sure" option, and this was highest for the ITR members (M = 9.48%) and lowest for the generic MTurkers (M = 2.1%) (Fig. 3). This measure was calculated by counting the number of pages where the "not sure" option was selected by a particular type of worker and dividing that number by the total number of pages shown to all workers of that type. Since ITR workers achieved the highest average accuracy for page detection, we assume that the community members are more confident of what they do not know and they did not want to provide incorrect answers (more in Section 5.3). A Chi-Square test showed that this number was significantly higher for ITR than generic MTurkers (p = .002).

When workers were not sure, they could answer 2-3 additional questions to guide us in determining the probable label of the webpage. Through an analysis of the answers to those additional questions we identified that ITR workers were more willing and accurate in answering those. In other words, to calculate average accuracy, if we consider the pages where we can obtain a label through workers' answers to additional questions, we get a 1% increase in the accuracy for ITR workers, whereas the accuracy of generic and community-situated MTurkers does not change at all.

4.4.3 Better Context Interpretations. We suspected that the community members could perform better in differentiating meeting pages from other pages due to their background knowledge of the AA program's structures and norms. For example, the findings from the previous work with generic crowdsourcing suggested for

page validation that many generic workers labeled webpages with events or office hours (that have times and locations) inaccurately as meeting pages [73]. We assumed that the community-situated workers would label pages more accurately in these cases. However, we did not have explicit ground truth data for a set of pages that listed events or office hours, and therefore, we cannot claim that ITR workers were significantly more accurate in differentiating meeting and event pages than other types for workers. Here, we provide the findings from manual observations:

- Some webpages consisting of events (and not meetings) were correctly classified as non-meeting pages by community-situated workers only, possibly contributing to a higher accuracy of the ITR workers.
- Community-situated workers performed better than generic crowd workers in differentiating single event information (e.g., day, time, address of picnics or service meetings, or office hours) from meeting information.
- Fig. 3 shows the average percentage of day, time, and address information edited by the workers. The higher accuracy of the community workers than the generic workers in editing the day information might occur due to workers being more familiar with the structure of the webpages and taking more time to skim through the pages. This is because from our observations, we had seen many meeting websites that listed the meetings per day of the week. For these types of pages, workers had to scroll through the top of the page to edit the day information.

5 DISCUSSION

The implications of these results point to next steps in our research on providing a reliable information source for meetings for people in recovery. In addition, we discuss practical implications for technology design given the impacts of worker type on the outcome of data validation, as well as research opportunities for the CHI community.

5.1 Implications for the Context of Recovery

Our investigations point to the importance of considering the trade-offs of time, cost, and accuracy while adapting crowdsourcing for AA meeting data validation. The list of AA meetings in an area largely depends on local AA groups continually updating changes to meetings, which makes the current meeting finders outdated for different regions. Crowdsourcing, when combined with automated information retrieval techniques, can help periodically extract and validate meeting information from different websites and provide a reliable up-to-date source of information. We used the context of recovery as an example to broadly answer our research questions regarding community-situated crowdsourcing applicable to other domains as well. Consequently, our results had many interesting implications for the next step in utilizing crowdsourcing effectively in this specific domain.

InTheRooms volunteers achieved the highest accuracy for most of the tasks (including the comparatively longer task of meeting validation) among the three different worker samples. This finding implies that we can rely on volunteers for data validation for this particular problem. This resonates with the findings of many other studies showing that community-based crowdsourcing results in more accurate and reliable outcomes [32, 82, 85]. Keeping the meeting lists up-to-date would require help from workers in periodic intervals and community-situated crowdsourcing can ensure a stream of continuous workers, potentially reducing the worker cost to zero or close to that. Although the platform cost does not impact worker performance directly, this is part of the total cost required for the study. Cost for recruiting community-situated workers can be significantly reduced both by using other online platforms and apps, and by developing our own system to recruit community members. For example, with the meeting data initially extracted and validated, we can build an up-to-date meeting finder application targeted for AA members. We can potentially ask the app users to volunteer in helping us improve the meeting list through validating meeting information. However, relying on a system or app in its initial phase with a small number of users may be questionable to recruit volunteers, since depending on the volume of data requiring validation, recruiting the desired number of participants can take a substantial amount of time. For example, we recruited from an online community with more than 500,000 members, but the recruitment time was about two months for 450 participants. As a counter-argument, our study limited the participants to perform the tasks only once in order to remove the bias of learning by repetition, whereas in a real-time system, each worker would be able to perform the tasks as many times as possible, minimizing the trade-off in recruitment time, and improving worker expertise at the same time.

In summary, volunteers from recovery communities can be relied upon for validating AA meeting webpages and meetings, and the cost of recruitment can be substantially reduced by careful consideration in adapting different approaches to attract more volunteers.

5.2 Implications for Design of Crowdsourcing ICT Systems

Our findings suggest that many design decisions of crowdsourcing ICT systems, such as the task interfaces, the choice of platform, etc. should consider the trade-offs in cost and accuracy. We found statistically significant difference in the mean percentage of accuracy between the generic and the community-situated crowd workers. However, "statistically significant" differences in time, accuracy, or cost between different types of crowdsourcing may not be practically significant for many other domains. Careful considerations have to be made regarding these trade-offs to select the type of workers and the platform depending on the monetary budget, urgency of retrieving crowdsourced answers, and the accepted threshold of errors in those.

For example, the choice of a platform to solicit the community-situated workers from largely depends on the budget for the platform cost, existing collaboration with the platform's owners, and the sensitivity of the context. In the case of recovery, members are particularly vulnerable and anonymity is of utmost importance to them [74]. We made sure to design the tasks in a way such that no identifiable personal information has to be revealed by the participants. If the information being validated requires crowdsourcing from people with a stigmatized health condition, members of online communities where real identity is associated (e.g., Facebook health

support groups) may not show enough interest in participation. Additionally, if the data requires validation in frequent intervals like the context of this study, requesters have to choose a platform from where they can recruit the expected number of workers within the time constraints. While designing tasks in this type of recruitment, one or more existing techniques of minimizing errors while getting rapid answers [48, 51, 65] should be adopted.

About 4% of our participants from ITR completed the tasks partially (e.g., completed one or two tasks). The order of the tasks was randomized but a close look at our data revealed that many of them left in the middle when the task of meeting validation was shown to them as the last task. We found out that validating meetings took the longest time on an average among the three tasks for both types of crowdsourcing. Making the overall task more granular (i.e., one task per worker instead of three) would probably yield more workers completing the assigned task and increase engagement, as suggested by previous work [19, 52]. However, in the case of ITR workers, who were redirected to the task interfaces from clicking on the advertisement, dividing the tasks would require three times the number of interested participants (about 450 participants per task). These trade-offs are important to consider.

5.3 Implications for Research

Findings from our study provide practical implications for further research regarding the trade-offs and benefits of generic vs. paid community-situated vs. unpaid community-situated crowdsourcing in different contexts.

5.3.1 Trade-offs in community-situated crowdsourcing. Generic MTurkers in this study achieved about the same accuracy for meeting identification, but were significantly quicker in completing the task. Piloting the tasks with different types of workers prior to running the actual experiments may provide a guideline for determining which tasks are more suitable for a particular type of crowdsourcing vs. the others. Similarly, the choice of recruiting community-situated workers from online marketplaces like MTurk rather than online communities can be influenced by the actual proportion of the community members in the general population. For instance, about 1% Americans are members of AA [61], and MTurk typically represents similar proportions of particular community members as the general US population. In many other domains, however, the required community members may represent a larger proportion of the general us population (e.g., students). Researchers conducting experiments related to those communities may seek community-situated crowd workers from MTurk or other online marketplaces, where they can recruit a large number of people very quickly, instead of selecting online groups with a comparatively smaller number of people.

Using the wisdom and motivations from community-situated crowds in this context produced high quality results. This provides implications to develop approaches for collaborative crowdsourcing. In the context of paid microtasks, collective crowdsourcing with the paired-worker model has led to better accuracy and increased output, which, in turn, has translated into lower costs [23, 53]. Moreover, different types of workers exhibit different potentials [14, 36]. Researchers can build on the findings of these previous works and our study to create coordination, where community-situated workers help set directions and guide the non-expert generic workers.

The number of ITR members selecting the "not sure" option for the page validation task was significantly higher than the other two populations. This finding echoes the well known Dunning–Kruger effect, which is defined in psychology as a cognitive bias in which low-ability individuals suffer from illusory superiority, mistakenly assessing their ability as much higher than it really is [20]. In other words, experts tend to know what they do not know. In the context of crowdsourcing, this may affect answer quality tremendously if the task comprises many difficult or confusing questions. Therefore, further research needs to be done to find out ways to minimize this effect. Overall, our findings and collective observations highlight multiple opportunities and directions to pursue deeper understanding of effective applications of community-situated crowdsourcing in different domains, while not compromising the quality of the crowdsourced results.

5.3.2 Consideration of platform cost. Researchers can optimize the platform cost by carefully choosing a platform that has no or little impact on other performance measures (e.g., time of recruitment, answer quality). Take Facebook advertising as an example. The average cost per click for Facebook ads is $1.72 [39], and thus using Facebook ads instead of ITR would possibly reduce the platform cost for this study. However, it might have impacted the response rate, as AA members are particularly concerned about their anonymity [74], and there may not be a large number of targeted 12-step members on Facebook who would participate. This may not be an issue for studies requiring other types of community-situated workers, and those may benefit from recruiting workers through platforms with lower advertising costs. Future research should aim to understand the relationship between platform cost and worker performance, and provide a comprehensive guideline on selecting the right platform for a study.

5.3.3 Impact of motivation and incentives. Research focused on online community-based crowdsourcing should address particular motivations and expertise of the community members to figure out the feasibility and effectiveness of recruiting such workers. HCI researchers have pointed out that community-based intrinsic motivations such as altruism, collectivism, and reciprocity play an important role in worker participation and engagement [3, 5, 26, 47, 49, 50]. Similarly, worker expertise and knowledge about the problem domain have proven to be directly related with the quality of crowdsourced data [17, 79, 86]. In many other contexts like recovery, the expertise can come from particular community members who are willing to volunteer for social reasons, such as to achieve recognition in the community or to provide service. Researchers should investigate generic and community-situated crowdsourcing in those domains to understand the benefits and the trade-offs in time, quality, and cost. For example, members of an online community of cancer caregivers may provide better answers to a care-giving related question than a non-caregiver.

While community members are often considered as experts for the target task, recruiting such participants from paid online platforms has the drawback of workers falsely reporting their membership. Although screening questions can reduce the effect, there is always a possibility that some workers have made their way to

bypass the screening techniques. Further research can focus on analyzing the impact of different screening techniques on the worker performance while recruiting from paid platforms.

ITR members in our study did not receive any monetary incentives. We assume a major motivation behind their participation was to provide service to the community by helping us create a useful resource for the newcomers in the community. We provided an option for the workers to leave us comments on the task descriptions and the design, and some of the comments from the ITR participants actually revealed this "sense of providing service" to the community by helping others (e.g., "thank you for allowing me to help out.I believe it's our purpose to help the newcomer if anyone needs help."). This intrinsic motivation probably also encouraged them to take sufficient time to select accurate answers for as many questions as they could. One might expect the community-situated MTurkers to perform better than ITR members, as they should have the same intrinsic motivation along with monetary remuneration for task completion. However, they might not have paid enough attention in order to complete more tasks in less time, resulting in lower accuracy of task completion than the ITR participants. Prior research has established the impact of incentives and motivations on not only who participates in a particular task [35], but also on their attention, dropouts, and performance [37, 43]. Based on prior work and our findings that different types of motivations and incentives may affect workers' attention and question-answering behavior differently, we recommend future research to use motivation and incentives to analyze reliability of the answers. As an example, whenever possible, researchers can compute worker motivation using established scales [67] and figure out if workers with particular motivation types provide quality answers and further target workers with similar motivations.

5.4 Broader Impact and Takeaways

Our study results have practical implications regarding performance trade-offs for different types of crowdsourcing that may be helpful for other HCI and CSCW researchers. As a concrete example, our findings can inform crowdsourcing research in other health contexts. Previous research has discussed the effectiveness of friendsourcing to generate behavior change plans, but also pointed out concerns about sharing personal information with friends, or offending friends by not following their advice [4]. Community-situated crowdsourcing can be applied to offer advice from other unknown individuals who may be similar on some experiential dimensions. Our paper provides guidance on whether a "filtered generic crowd" or "recruiting from a particular community" crowdsourcing approach would be more appropriate by demonstrating time, cost, and accuracy trade-offs between these two reasonable approaches to seeking out such individuals.

6 LIMITATIONS

We assumed that the community members from InTheRooms are in fact members of 12-step fellowships and did not provide the screening questions for them. Although seemingly there is no particular reason for a non-member to join this online community and take part in the study (as discussed in Section 3.2.3), using the same screening criterion for both platforms would make the study conditions fairer. Similarly, even though the three questions we used for screening were not too easy to bypass and we adopted additional filtering criteria to approve answers, we cannot certainly claim that all of the community-situated MTurkers were in fact 12-step fellowship members. In addition, the motivation and the incentive (i.e., pay for the MTurk community-situated workers, altruism and sense of belonging to the community for ITR workers, etc.) may have impacted the outcome of the study, and therefore, follow-up experiments should be conducted to rule out the impact of incentive and motivation. In particular, comparisons of task completion time, accuracy, and cost should be carried out with sample workers recruited from ITR with a monetary incentive and generic workers recruited for free.

7 CONCLUSION

In the context of crowdsourcing data validation for Alcoholics Anonymous meetings, we provided an empirical comparison of accuracy, cost, and time trade-offs in generic crowdsourcing from MTurk, community-situated crowdsourcing from MTurk, and comm-unity-situated crowdsourcing from an online recovery community, InTheRooms. Our results show that community-situated workers from the online recovery community achieved significantly higher accuracy in the validation tasks than the other two types of crowd workers. We further investigated the possible reasons for this difference in accuracy and found that task completion time, knowing when to be "not sure", and better context interpretations may have influenced the accuracy of different types of crowd workers. From our findings, we provide practical implications for crowdsourcing in the recovery context, as well as for research and design in other domains. We discuss the factors that should be considered while selecting community-situated workers from a particular platform vs. other platforms, including platform cost, urgency and frequency of crowdsourcing, and type of the crowdsourced tasks.

ACKNOWLEDGMENTS

Thank you to the participants for their time and contribution. We thank Zachary Levonian and Daniel Kluver for their useful feedback on the quantitative analysis. This work was funded by the NSF grants (1464376 and 1651575).

REFERENCES

[1] Elena Agapie, Lucas Colusso, Sean A. Munson, and Gary Hsieh. 2016. PlanSourcing: Generating Behavior Change Plans with Friends and Crowds. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 119–133. https://doi.org/10.1145/2818048.2819943 event-place: San Francisco, California, USA.

[2] Rawan N. Al-Matham and Hend S. Al-Khalifa. 2017. A Crowdsourcing Web-Based System for Reporting Predatory Publishers. In *Proceedings of the 19th International Conference on Information Integration and Web-Based Applications & Services (iiWAS '17)*. Association for Computing Machinery, New York, NY, USA, 573–576. https://doi.org/10.1145/3151759.3151844

[3] Sultana Lubna Alam and John Campbell. 2012. Crowdsourcing Motivations in a not-for-profit GLAM context: The Australian Newspapers Digitisation Program. In *ACIS 2012 : Proceedings of the 23rd Australasian Conference on Information Systems*. Australasian Conference on Information Systems. https://openresearch-repository.anu.edu.au/handle/1885/154077

[4] Omar Alonso and Matthew Lease. 2011. Crowdsourcing 101: Putting the WSDM of Crowds to Work for You. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, 1–2. https://doi.org/10.1145/1935826.1935831 event-place: Hong Kong, China.

[5] D. Auferbauer and H. Tellioundefinedlu. 2017. Centralized Crowdsourcing in Disaster Management: Findings and Implications. In *Proceedings of the 8th International Conference on Communities and Technologies (C&T '17)*. Association for Computing Machinery, New York, NY, USA, 173–182. https://doi.org/10.1145/3083671.3083689

[6] Daniel Robert Bateman, Erin Brady, David Wilkerson, Eun-Hye Yi, Yamini Karanam, and Christopher M. Callahan. 2017. Comparing Crowdsourcing and Friendsourcing: A Social Media-Based Feasibility Study to Support Alzheimer Disease Caregivers. *JMIR Research Protocols* 6, 4 (2017), e56. https://doi.org/10.2196/resprot.6904

[7] Michael S. Bernstein, Desney Tan, Greg Smith, Mary Czerwinski, and Eric Horvitz. 2008. Personalization via Friendsourcing. *ACM Trans. Comput.-Hum. Interact.* 17, 2 (5 2008), 6:1–6:28. https://doi.org/10.1145/1746259.1746260

[8] Ria Mae Borromeo, Thomas Laurent, and Motomichi Toyama. 2016. The Influence of Crowd Type and Task Complexity on Crowdsourced Work Quality. In *Proceedings of the 20th International Database Engineering & Applications Symposium (IDEAS '16)*. Association for Computing Machinery, New York, NY, USA, 70–76. https://doi.org/10.1145/2938503.2938511

[9] Ria Mae Borromeo and Motomichi Toyama. 2016. An investigation of unpaid crowdsourcing. *Human-centric Computing and Information Sciences* 6, 1 (2016), 11.

[10] Soumia Bougrine, Hadda Cherroun, and Ahmed Abdelali. 2017. Altruistic crowdsourcing for arabic speech corpus annotation. *Procedia Computer Science* 117 (2017), 137–144.

[11] Erin Brady. 2015. Getting Fast, Free, and Anonymous Answers to Questions Asked by People with Visual Impairments. *SIGACCESS Access. Comput.* 112 (7 2015), 16–25. https://doi.org/10.1145/2809915.2809918

[12] Erin L. Brady, Yu Zhong, Meredith Ringel Morris, and Jeffrey P. Bigham. 2013. Investigating the Appropriateness of Social Network Question Asking As a Resource for Blind Users. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 1225–1236. https://doi.org/10.1145/2441776.2441915 event-place: San Antonio, Texas, USA.

[13] Florian Brühlmann, Serge Petralito, Lena F. Aeschbach, and Klaus Opwis. 2020. The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology* 2 (Nov. 2020), 100022. https://doi.org/10.1016/j.metip.2020.100022

[14] Pern Hui Chia and John Chuang. 2012. Community-Based Web Security: Complementary Roles of the Serious and Casual Contributors. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. Association for Computing Machinery, New York, NY, USA, 1023–1032. https://doi.org/10.1145/2145204.2145356

[15] Franco Curmi, Maria Angela Ferrario, Jon Whittle, and Florian 'Floyd' Mueller. 2015. Crowdsourcing Synchronous Spectator Support: (Go on, Go on, You're the Best)N-1. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 757–766. https://doi.org/10.1145/2702123.2702338 event-place: Seoul, Republic of Korea.

[16] Luca de Alfaro and Michael Shavlovsky. 2014. CrowdGrader: A Tool for Crowdsourcing the Evaluation of Homework Assignments. *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, 415–420. https://doi.org/10.1145/2538862.2538900 event-place: Atlanta, Georgia, USA.

[17] Roelof A. J. de Vries, Cristina Zaga, Franciszka Bayer, Constance H. C. Drossaert, Khiet P. Truong, and Vanessa Evers. 2017. Experts Get Me Started, Peers Keep Me Going: Comparing Crowd- Versus Expert-designed Motivational Text Messages for Exercise Behavior Change. *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 155–162. https://doi.org/10.1145/3154862.3154875 event-place: Barcelona, Spain.

[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*. 135–143.

[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 (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 238–247. https://doi.org/10.1145/2736277.2741685

[20] David Dunning. 2011. Chapter five - The Dunning–Kruger Effect: On Being Ignorant of One's Own Ignorance. In *Advances in Experimental Social Psychology*, James M. Olson and Mark P. Zanna (Eds.). Vol. 44. Academic Press, 247–296. https://doi.org/10.1016/B978-0-12-385522-0.00005-6

[21] Lee Erickson, Irene Petrick, and Eileen Trauth. 2012. Hanging with the right crowd: Matching crowdsourcing need to crowd characteristics. *AMCIS 2012 Proceedings* (29 7 2012). https://aisel.aisnet.org/amcis2012/proceedings/VirtualCommunities/3

[22] M. Ferri, L. Amato, and M. Davoli. 2006. Alcoholics Anonymous and other 12-step programmes for alcohol dependence. *The Cochrane Database of Systematic Reviews* 3 (19 7 2006), CD005032. https://doi.org/10.1002/14651858.CD005032.pub2 PMID: 16856072.

[23] Oluwaseyi Feyisetan and Elena Simperl. 2017. Social Incentives in Paid Collaborative Crowdsourcing. *ACM Trans. Intell. Syst. Technol.* 8, 6, Article Article 73 (July 2017), 31 pages. https://doi.org/10.1145/3078852

[24] Karen E Fisher, Kenton T Unruh, and Joan C Durrance. 2003. Information communities: Characteristics gleaned from studies of three online networks. *Proceedings of the American Society for Information Science and Technology* 40, 1 (2003), 298–305.

[25] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1631–1640. https://doi.org/10.1145/2702123.2702443

[26] Victor Girotto. 2016. Collective Creativity through a Micro-Tasks Crowdsourcing Approach. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion (CSCW '16 Companion)*. Association for Computing Machinery, New York, NY, USA, 143–146. https://doi.org/10.1145/2818052.2874356

[27] Jorge Goncalves, Denzil Ferreira, Simo Hosio, Yong Liu, Jakob Rogstadius, Hannu Kukka, and Vassilis Kostakos. 2013. Crowdsourcing on the Spot: Altruistic Use of Public Displays, Feasibility, Performance, and Behaviours. *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 753–762. https://doi.org/10.1145/2493432.2493481 event-place: Zurich, Switzerland.

[28] Jorge Goncalves, Simo Hosio, Denzil Ferreira, Theodoros Anagnostopoulos, and Vassilis Kostakos. 2015. Bazaar: A Situated Crowdsourcing Market. *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, 1385–1390. https://doi.org/10.1145/2800835.2800974 event-place: Osaka, Japan.

[29] Jorge Goncalves, Simo Hosio, Denzil Ferreira, and Vassilis Kostakos. 2014. Game of Words: Tagging Places Through Crowdsourcing on Public Displays. *Proceedings of the 2014 Conference on Designing Interactive Systems*, 705–714. https://doi.org/10.1145/2598510.2598514 event-place: Vancouver, BC, Canada.

[30] Jorge Goncalves, Simo Hosio, Yong Liu, and Vassilis Kostakos. 2016. Worker Performance in a Situated Crowdsourcing Market. *Interacting with Computers* 28, 5 (08 2016), 612–624. https://doi.org/10.1093/iwc/iwv035

[31] Daniel González, Regina Motz, and Libertad Tansini. 2013. Recommendations Given from Socially-Connected People, Yan Tang Demey and Hervé Panetto (Eds.). *On the Move to Meaningful Internet Systems: OTM 2013 Workshops*, 649–655.

[32] Kota Gushima, Mizuki Sakamoto, and Tatsuo Nakajima. 2016. Community-Based Crowdsourcing to Increase a Community's Well-Being. In *Proceedings of the 18th International Conference on Information Integration and Web-Based Applications and Services (iiWAS '16)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3011141.3011173

[33] Hwajung Hong, Eric Gilbert, Gregory D. Abowd, and Rosa I. Arriaga. 2015. In-group Questions and Out-group Answers: Crowdsourcing Daily Living Advice for Individuals with Autism. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 777–786. https://doi.org/10.1145/2702123.2702402 [Online; accessed 2016-04-13].

[34] Simo Hosio, Jorge Goncalves, Vili Lehdonvirta, Denzil Ferreira, and Vassilis Kostakos. 2014. Situated Crowdsourcing Using a Market Model. *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, 55–64. https://doi.org/10.1145/2642918.2647362 event-place: Honolulu, Hawaii, USA.

[35] Gary Hsieh and Rafał Kocielnik. 2016. You Get Who You Pay for: The Impact of Incentives on Participation Bias. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 823–835. https://doi.org/10.1145/2818048.2819936

[36] Yun Huang, Yifeng Huang, Na Xue, and Jeffrey P. Bigham. 2017. Leveraging Complementary Contributions of Different Workers for Efficient Crowdsourcing of Video Captions. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 4617–4626. https://doi.org/10.1145/3025453.3026032

[37] Kazushi Ikeda and Michael S. Bernstein. 2016. Pay It Backward: Per-Task Payments on Crowdsourcing Platforms Reduce Productivity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4111–4121. https://doi.org/10.1145/2858036.2858327

[38] Panagiotis G. Ipeirotis and Evgeniy Gabrilovich. 2014. Quizz: Targeted Crowdsourcing with a Billion (Potential) Users. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*. Association for Computing Machinery, New York, NY, USA, 143–154. https://doi.org/10.1145/2566486.2567988

[39] Mark Irvine. [n.d.]. Facebook Ad Benchmarks for Industry Data. https://www.wordstream.com/blog/ws/2017/02/28/facebook-advertising-benchmarks Library Catalog: www.wordstream.com.

[40] Steve T. K. Jan, Chun Wang, Qing Zhang, and Gang Wang. 2018. Analyzing Payment-Driven Targeted Q8A Systems. *Trans. Soc. Comput.* 1, 3, Article Article 12 (Dec. 2018), 21 pages. https://doi.org/10.1145/3281449

[41] Steve T. K. Jan, Chun Wang, Qing Zhang, and Gang Wang. 2018. Analyzing Payment-Driven Targeted Q8A Systems. *Trans. Soc. Comput.* 1, 3, Article Article 12 (Dec. 2018), 21 pages. https://doi.org/10.1145/3281449

[42] Christopher Jonathan and Mohamed F. Mokbel. 2018. Stella: Geotagging Images via Crowdsourcing. *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 169–178. https://doi.org/10.1145/3274895.3274902 event-place: Seattle, Washington.

[43] Eunice Jun, Gary Hsieh, and Katharina Reinecke. 2017. Types of Motivation Affect Study Selection, Attention, and Dropouts in Online Experiments. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 56 (Dec. 2017), 15 pages. https://doi.org/10.1145/3134691

[44] Hernisa Kacorri, Kaoru Shinkawa, and Shin Saito. 2014. Introducing Game Elements in Crowdsourced Video Captioning by Non-Experts. In *Proceedings of the 11th Web for All Conference (W4A '14)*. Association for Computing Machinery, New York, NY, USA, Article Article 29, 4 pages. https://doi.org/10.1145/2596695.2596713

[45] Keiko Katsuragawa, Qi Shu, and Edward Lank. 2019. PledgeWork: Online Volunteering through Crowdwork. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article Paper 311, 11 pages. https://doi.org/10.1145/3290605.3300541

[46] Nicolas Kaufmann and Thimo Schulze. 0. Worker Motivation in Crowdsourcing and Human Computation. http://www.humancomputation.com/Program.html [Online; accessed 2019-05-13].

[47] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. 2011. More than fun and money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk. *AMCIS 2011 Proceedings - All Submissions* (6 8 2011). https://aisel.aisnet.org/amcis2011_submissions/340

[48] Aniket Kittur, E Chi, and Bongwon Suh. 2008. Crowdsourcing for usability: Using micro-task markets for rapid, remote, and low-cost user measurements. *Proc. CHI 2008* (2008).

[49] Masatomo Kobayashi, Shoma Arita, Toshinari Itoko, Shin Saito, and Hironobu Takagi. 2015. Motivating Multi-Generational Crowd Workers in Social-Purpose Work. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 1813–1824. https://doi.org/10.1145/2675133.2675255

[50] MIIA KOSONEN, CHUNMEI GAN, MIKA VANHALA, and KIRSIMARJA BLOMQVIST. 2014. USER MOTIVATION AND KNOWLEDGE SHARING IN IDEA CROWDSOURCING. *International Journal of Innovation Management* 18, 05 (2014), 1450031. https://doi.org/10.1142/S1363919614500315

[51] Ranjay A. Krishna, Kenji Hata, Stephanie Chen, Joshua Kravitz, David A. Shamma, Li Fei-Fei, and Michael S. Bernstein. 2016. Embracing Error to Enable Rapid Crowdsourcing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 3167–3179. https://doi.org/10.1145/2858036.2858115

[52] Anand Kulkarni, Matthew Can, and Björn Hartmann. 2012. Collaboratively Crowdsourcing Workflows with Turkomatic. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. Association for Computing Machinery, New York, NY, USA, 1003–1012. https://doi.org/10.1145/2145204.2145354

[53] Mucahid Kutlu, Tyler McDonnell, Aashish Sheshadri, Tamer Elsayed, and Matthew Lease. 2018. Mix and Match: Collaborative Expert-Crowd Judging for Building Test Collections Accurately and Affordably. *In Proceedings of the 1st Biannual Conference on the Design of Experimental Search & Information REtrieval Systems (DESIRES)* (2018), 42–46.

[54] Hongwei Li, Bo Zhao, and Ariel Fuxman. 2014. The Wisdom of Minority: Discovering and Targeting the Right Group of Workers for Crowdsourcing. *Proceedings of the 23rd International Conference on World Wide Web*, 165–176. https://doi.org/10.1145/2566486.2568033 event-place: Seoul, Korea.

[55] Qi Li, Fenglong Ma, Jing Gao, Lu Su, and Christopher J. Quinn. 2016. Crowdsourcing High Quality Labels with a Tight Budget. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM '16)*. Association for Computing Machinery, New York, NY, USA, 237–246. https://doi.org/10.1145/2835776.2835797

[56] Thomas Ludwig, Christoph Kotthaus, Christian Reuter, van Sören Dongen, and Volkmar Pipek. 2017. Situated crowdsourcing during disasters: Managing the tasks of spontaneous volunteers through public displays. *International Journal of Human-Computer Studies* 102 (1 6 2017), 103–121. https://doi.org/10.1016/j.ijhcs.2016.09.008

[57] Andrew Mao, Ece Kamar, Yiling Chen, Eric Horvitz, Megan E. Schwamb, Chris J. Lintott, and Arfon M. Smith. [n.d.]. Volunteering Versus Work for Pay: Incentives and Tradeoffs in Crowdsourcing. In *First AAAI Conference on Human Computation and Crowdsourcing* (2013). https://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7497 Retrieved January 9, 2020 from.

[58] João Martins, José Carilho, Oliver Schnell, Carlos Duarte, Francisco M. Couto, Luís Carriço, and Tiago Guerreiro. 2014. Friendsourcing the Unmet Needs of People with Dementia. *Proceedings of the 11th Web for All Conference*, 35:1–35:4. https://doi.org/10.1145/2596695.2596716 event-place: Seoul, Korea.

[59] Winter Mason and Duncan J. Watts. 2009. Financial Incentives and the "Performance of Crowds". In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '09)*. Association for Computing Machinery, New York, NY, USA, 77–85. https://doi.org/10.1145/1600150.1600175

[60] Bart Mellebeek, Francesc Benavent, Jens Grivolla, Joan Codina, Marta R. Costa-jussà, and Rafael Banchs. 2010. Opinion Mining of Spanish Customer Comments with Non-Expert Annotations on Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10)*. Association for Computational Linguistics, USA, 114–121.

[61] AA membership. 2019. Worldwide A.A. Individual and Group Membership. https://www.aa.org/assets/en_US/aa-literature/smf-132-estimates-worldwide-aa-individual-and-group-membership. Accessed: 2020-09-30.

[62] Meredith Ringel Morris, Kori Inkpen, and Gina Venolia. 2014. Remote Shopping Advice: Enhancing In-store Shopping with Social Technologies. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 662–673. https://doi.org/10.1145/2531602.2531707 event-place: Baltimore, Maryland, USA.

[63] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. 2010. A Comparison of Information Seeking Using Search Engines and Social Networks. *Fourth International AAAI Conference on Weblogs and Social Media*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1518 [Online; accessed 2019-05-13].

[64] Robert R. Morris, Stephen M. Schueller, and Rosalind W. Picard. 2015. Efficacy of a Web-based, crowdsourced peer-to-peer cognitive reappraisal platform for depression: randomized controlled trial. *Journal of Medical Internet Research* 17, 3 (30 3 2015), e72. https://doi.org/10.2196/jmir.4167 PMID: 25835472 PMCID: PMC4395771.

[65] Swaprava Nath, Pankaj Dayama, Dinesh Garg, Y. Narahari, and James Zou. 2012. Threats and Trade-Offs in Resource Critical Crowdsourcing Tasks Over Networks. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*. https://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/4807

[66] 1615 L. St NW, Suite 800Washington, and DC 20036USA202-419-4300 | Main202-857-8562 | Fax202-419-4372 | Media Inquiries. [n.d.]. How Mechanical Turk Works. https://www.pewresearch.org/internet/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/pi_2016-07-11_mechanical-turk_0-01/

[67] Lisa Posch, Arnim Bleier, Clemens M. Lechner, Daniel Danner, Fabian Flöck, and Markus Strohmaier. 2019. Measuring Motivations of Crowdworkers: The Multidimensional Crowdworker Motivation Scale. *Trans. Soc. Comput.* 2, 2, Article 8 (Sept. 2019), 34 pages. https://doi.org/10.1145/3335081

[68] Judith Redi and Isabel Povoa. 2014. Crowdsourcing for Rating Image Aesthetic Appeal: Better a Paid or a Volunteer Crowd?. In *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia (CrowdMM '14)*. Association for Computing Machinery, New York, NY, USA, 25–30. https://doi.org/10.1145/2660114.2660118

[69] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. [n.d.]. An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. In *Fifth International AAAI Conference on Weblogs and Social Media* (2011). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2778 Retrieved April 11, 2019 from.

[70] Dana Rotman, Jenny Preece, Jen Hammock, Kezee Procita, Derek Hansen, Cynthia Parr, Darcy Lewis, and David Jacobs. 2012. Dynamic Changes in Motivation in Collaborative Citizen-Science Projects. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. Association for Computing Machinery, New York, NY, USA, 217–226. https://doi.org/10.1145/2145204.2145238

[71] Ellen L. Rubenstein. 2013. Crowdsourcing Health Literacy: The Case of an Online Community. *Proceedings of the 76th ASIST Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries*, 120:1–120:5. http://dl.acm.org/citation.cfm?id=2655780.2655900 event-place: Montreal, Quebec, Canada.

[72] Sabirat Rubya. 2017. Facilitating Peer Support for Recovery from Substance Use Disorders. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 172–177. https://doi.org/10.1145/3027063.3048431

[73] Sabirat Rubya, Xizi Wang, and Svetlana Yarosh. 2019. HAIR: Towards Developing a Global Self-Updating Peer Support Group Meeting List Using Human-Aided Information Retrieval. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR '19)*. Association for Computing Machinery, New York, NY, USA, 83–92. https://doi.org/10.1145/3295750.3298933

[74] Sabirat Rubya and Svetlana Yarosh. 2017. Interpretations of Online Anonymity in Alcoholics Anonymous and Narcotics Anonymous. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article Article 91 (Dec. 2017), 22 pages. https://doi.org/10.1145/3134726

[75] Jeffrey M. Rzeszotarski and Meredith Ringel Morris. 2014. Estimating the Social Costs of Friendsourcing. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2735–2744. https://doi.org/10.1145/2556288.2557181 event-place: Toronto, Ontario, Canada.

[76] Eric Schenk and Claude Guittard. 2011. Towards a characterization of crowdsourcing practices. *Journal of Innovation Economics Management* n°7, 1 (6 4 2011),

93–107.

[77] Clay Shirky. 2010. *Cognitive Surplus: How Technology Makes Consumers into Collaborators*. Penguin. Google-Books-ID: PjeTO822t_4C.

[78] Tammara Combs Turner, Marc A Smith, Danyel Fisher, and Howard T Welser. 2005. Picturing Usenet: Mapping computer-mediated collective action. *Journal of Computer-Mediated Communication* 10, 4 (2005), JCMC1048.

[79] Simon C. Warby, Sabrina L. Wendt, Peter Welinder, Emil G.S. Munk, Oscar Carrillo, Helge B.D. Sorensen, Poul Jennum, Paul E. Peppard, Pietro Perona, and Emmanuel Mignot. [n.d.]. Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. *Nature Methods* 11, 4 ([n. d.]), 385–392. https://doi.org/10.1038/nmeth.2855

[80] Helen Wauck, Yu-Chun (Grace) Yen, Wai-Tat Fu, Elizabeth Gerber, Steven P. Dow, and Brian P. Bailey. 2017. From in the Class or in the Wild?: Peers Provide Better Design Feedback Than External Crowds. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 5580–5591. https://doi.org/10.1145/3025453.3025477 event-place: Denver, Colorado, USA.

[81] Kerri Wazny. [n.d.]. Applications of crowdsourcing in health: an overview. *Journal of Global Health* 8, 1 ([n. d.]). https://doi.org/10.7189/jogh.08.010502

[82] Andrea Wiggins and Yurong He. 2016. Community-Based Data Validation Practices in Citizen Science. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1548–1559. https://doi.org/10.1145/2818048.2820063

[83] Svetlana Yarosh. 2013. Shifting Dynamics or Breaking Sacred Traditions?: The Role of Technology in Twelve-step Fellowships. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3413–3422. https://doi.org/10.1145/2470654.2466468 [Online; accessed 2016-03-29].

[84] Ming Yin, Yiling Chen, and Yu-An Sun. [n.d.]. The Effects of Performance-Contingent Financial Incentives in Online Labor Markets. In *Twenty-Seventh AAAI Conference on Artificial Intelligence* (2013). https://www.aaai.org/ocs/index.php/AAAI/AAAI13/paper/view/6301 Retrieved January 9, 2020 from.

[85] Z. Yu, D. Zhang, D. Yang, and G. Chen. 2012. Selecting the Best Solvers: Toward Community Based Crowdsourcing for Disaster Management. In *2012 IEEE Asia-Pacific Services Computing Conference*. 271–277. https://doi.org/10.1109/APSCC.2012.20

[86] Alvin Yuan, Kurt Luther, Markus Krause, Sophie Isabel Vennix, Steven P Dow, and Bjorn Hartmann. 2016. Almost an Expert: The Effects of Rubrics and Expertise on Perceived Value of Crowdsourced Design Critiques. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1005–1017. https://doi.org/10.1145/2818048.2819953

[87] Dongqing Zhu and Ben Carterette. 2010. An analysis of assessor behavior in crowdsourced preference judgments. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*. 17–20.