Task Assignment using Worker Cognitive Ability and Context to Improve Data Quality in Crowdsourcing

Danula Hettiachchi

ORCID: 0000-0003-3875-5727

Doctor of Philosophy
February, 2021

School of Computing and Information Systems
Faculty of Engineering and Information Technology
The University of Melbourne, Australia

Submitted in total fulfilment of the requirements for the degree of Doctor of Philosophy.

Abstract

While crowd work on crowdsourcing platforms is becoming prevalent, there exists no widely accepted method to successfully match workers to different types of tasks. Previous work has considered using worker demographics, behavioural traces, and prior task completion records to optimise task assignment. However, optimum task assignment remains a challenging research problem, since proposed approaches lack an awareness of workers' cognitive abilities and context. This thesis investigates and discusses how to use these key constructs for effective task assignment: workers' cognitive ability, and an understanding of the workers' context. Specifically, the thesis presents 'CrowdCog', a dynamic online system for task assignment and task recommendations, that uses fast-paced online cognitive tests to estimate worker performance across a variety of tasks. The proposed task assignment method can achieve significant data quality improvements compared to a baseline where workers select preferred tasks. Next, the thesis investigates how worker context can influence task acceptance, and it presents 'CrowdTasker', a voice-based crowdsourcing platform that provides an alternative form factor and modality to crowd workers. Our findings inform how to better design crowdsourcing platforms to facilitate effective task assignment and recommendation, which can benefit both workers and task requesters.

Declaration

This is to certify that

1. the thesis comprises only my original work towards the PhD,
2. due acknowledgement has been made in the text to all other material used,
3. appropriate ethics procedure and guidelines have been followed to conduct this research,
4. the thesis is less than the 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Danula Hettiachchi
February 2021

Preface

This thesis is submitted in total fulfilment of the requirements for the degree of *Doctor of Philosophy* at the University of Melbourne. The research presented here was primarily conducted at the School of Computing and Information Systems, The University of Melbourne under the supervision of Dr. Jorge Goncalves and Prof. Vassilis Kostakos. The work was supported by the Melbourne Research Scholarship, awarded by The University of Melbourne.

The thesis comprises four peer-reviewed articles referred by roman numerals (Article I - IV), in accordance with The University of Melbourne guidelines for a *Thesis with Publication*[1]. All articles are included in full, prefaced by a brief introduction situating each article within the context of the thesis. While several collaborators have contributed to the articles, I declare that I am the primary author and have more than 50% contributions in each of the following publications:

Article I

Hettiachchi D., van Berkel N., Hosio S., Kostakos V., Goncalves J. (2019) Effect of Cognitive Abilities on Crowdsourcing Task Performance. *In: Human-Computer Interaction – INTERACT 2019. Lecture Notes in Computer Science*, vol 11746. Springer, Cham. https://doi.org/10.1007/978-3-030-29381-9_28.

Ethics ID: 1852019, The University of Melbourne Human Ethics Advisory Group.

Published by Springer, Cham on August 2019.

Article II

Hettiachchi, D., van Berkel, N., Kostakos, V., Goncalves, J. (2020). CrowdCog: A Cognitive Skill based System for Heterogeneous Task Assignment and Recommendation in Crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2), 1–22. https://doi.org/10.1145/3415181

Ethics ID: 1853314, The University of Melbourne Human Ethics Advisory Group.

Published by ACM on October 2020.

[1]The University of Melbourne. (2009). Graduate Research Training Policy (MPF1321). Retrieved from https://policy.unimelb.edu.au/MPF1321

Article III

Hettiachchi, D., Wijenayake, S., Hosio, S., Kostakos, V., Goncalves, J. (2020). How Context Influences Cross-Device Task Acceptance in Crowd Work. *In Proceedings of the Eighth AAAI Conference on Human Computation and Crowdsourcing - HCOMP 2020* (pp. 53–62). AAAI Press. https://ojs.aaai.org//index.php/HCOMP/article/view/7463

Ethics ID: 2056409, The University of Melbourne Human Ethics Advisory Group.

Published by AAAI on October 2020.

Article IV

Hettiachchi, D., Sarsenbayeva, Z., Allison, F., van Berkel, N., Dingler, T., Marini, G., Kostakos, V., Goncalves, J. (2020). "Hi! I am the Crowd Tasker" Crowdsourcing through Digital Voice Assistants. *In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems - CHI 2020* (pp. 1–14). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3313831.3376320

Ethics ID: 1954377, The University of Melbourne Human Ethics Advisory Group.

Published by ACM on April 2020.

I would like to acknowledge the contribution of the following sources of funding that supported my research:

- Melbourne Research Scholarship
- Google PhD Travel Grant
- Melbourne Faculty of Engineering and Information Technology Travel Scholarship
- Research Internship at Amazon

Article I and Article II received minimal editorial editing in the regular publication process. No other third-party editorial assistance was provided in preparation of the thesis. The studies presented in Article I - IV received ethics approval from The University of Melbourne Human Ethics Advisory Group under the ethics applications listed above.

Acknowledgements

I am thankful to everyone who accompanied me during this journey, making it a rewarding experience.

First and foremost, I express my gratitude to my supervisors, Dr Jorge Goncalves and Professor Vassilis Kostakos. I am incredibly grateful for your impeccable guidance and the many career and life lessons learned. Working with excellent scholars like you mainly contributed to all my wins and personal growth during these years. Jorge, thank you for all the wisdom and for always getting the best out of me. I highly value that you are a reliable and very supportive supervisor, a mentor and a friend. I thoroughly enjoyed and learned from our encounters, let it be our regular meetings or casual meetups. Perhaps, what I like the most is your genuine competitiveness when playing games. Vas, I really appreciate how ambitious and yet practical you are. I am fortunate to receive your constant feedback, support, and profound questions that often shaped my research direction and this thesis. Thank you for always encouraging me to aim high, giving responsibility and connecting with us as a humble person.

I thank Assoc. Prof. Simo Hosio for being an important part of my PhD journey. Your feedback and input have been very helpful in shaping my PhD. Thank you for your continued encouragement, support and friendly chats.

I started my PhD sharing the 4.04 office with three brilliant PhD students. Thank you, Dr Niels van Berkel, Dr Zhanna Sarsenbayeva, and Dr Chu Luo, for welcoming me to the group and providing great company. From random chats to getting feedback on my work, there was always a lot to learn from you, which immensely benefited my work.

Many thanks to everyone at the HCI group and CIS. I should specially mention Dr Tilman Dingler, Dr Eduardo Velloso and Dr Benjamin Tag. I am grateful for the encouragement and wisdom you have shared. I will undoubtedly cherish all the memorable encounters we had throughout the years. Thank you, Assoc. Prof. Jenny Waycott, for your insightful feedback and being a kind and supportive committee chair and Assoc. Prof. George Buchanan, Dr Jarrod Knibbe, Dr Andrew Irlitti, Dr Ronal Singh, Dr Greg Wadley, Dr Steven Baker, Dr Ryan Kelly, and Dr Dana McKay.

I would like to thank Prof. Frank Vetere for providing me with the opportunity to work as the usability lab manager. The role allowed me to work with many HCI researchers and expand my skills. Thank you, Dr Nicole Barbee, for your support and promptly attending to all the troubles. Many thanks to the previous lab manager, Dr Joshua Newn, for suggesting me for the role, helping me throughout and being a great friend. Thanks to Dr Melissa Rogerson, Zaher Joukhadar, and Allen Pilares for your support during this role.

It was a delight to work with many brilliant colleagues who became lifelong friends. Thank you, Dr Fraser Allison, Gabriele Marini, and Senuri Wijenayeke, who also

supported the research presented in this thesis. Special thanks to my wonderful friends at 5.10, Chaofan Wang, Kangning Yang, Difeng Yu, Weiwei Jiang, Jing Wei, Qiushi Zhou, Brandon Syiem, Henrietta Lyons, Martin Reinoso, Ebrahim Babaei, Madeleine Antonellos, and Elsy Garcia, for keeping the excitement of the PhD journey. Thanks to Dr Yousef Kowsar, Dr Sarah Webber, Dr Romina Carrasco, Namrata Srivastava, Prashan Madumal, Unni Krishnan, Ahed Aldwin, and many more. I am glad that I had such great company during my PhD. Thanks to all the visitors, Assoc. Prof. Evangelos Karapanos, Dr Katerina Mangaroska, Hank Lee, Oludamilare Matthews, Iuliia Brishtel and others for sharing exciting stories and adding colour to our PhD lab.

I was fortunate to complete an internship at Amazon during my PhD. Many thanks to all the colleagues at Ground Truth Science Team for making my internship a fruitful experience. I am very grateful to my mentor, Assoc. Prof. Matt Lease and Lei Jin, from whom I learned a lot. Thank you, Matt, for your unwavering support. Thanks to Prof. Pietro Perona, Dr Tristan Mckinney, Dr Mike Schaekermann, Dr Fedor Zhadonov, Chris Zazzi, Jonathan Buck and others. It was a privilege to work with such a talented group.

Next, I would like to thank Prof. Martin Tomitsch, Prof. Jeni Paay, Dr Frederik Brudy, Dr Joel Fredericks, Dr Barrett Ens, Dr Maxime Cordeil, Assoc. Prof. Tuck Leong, Assoc. Prof. Yolande Strengers, Kadek Satriadi, Jessica Rahman, and many others I worked closely with during conference organizing activities at CHI 2019, CHI Downunder 2020 and OzCHI 2020. Thank you for the opportunities. I am looking forward to seeing and perhaps working with you again. I thank Dr Lachlan Hayes from Nothern Hospital Epping and the Medtasker team for their collaboration.

I further acknowledge the University of Melbourne, Amazon and Google for supporting my research during the PhD. Thanks to all participants involved, as my research would not have been possible without you.

I am also grateful to all the teachers who guided and moulded me. Special thanks to Prof. Sanath Jayasena, Dr Chandana Gamage, Dr Dilum Bandara and all the academics at the University of Moratuwa for their continued support during my bachelor's degree. Owing to the free education, I was fortunate to study at the finest educational institutes in Sri Lanka. Hence, my sincere gratitude goes to the people of Sri Lanka.

I am surrounded by a wonderful set of friends in Australia, Sri Lanka and beyond. Many thanks to all my dear friends for your comfort and the great times we shared together.

I am incredibly thankful to my beloved mother, Dhamma and my father, Ranjith. They are amazing parents who worked really hard to give my sister and me the best. Thank you for your comfort and for always motivating me to become the best version of myself. Thank you to my sister Samudi for all the love and for all my relatives. Finally, thank you to my wife, Maneesha Perera. I am blessed to have you by my side in this journey, and it would not have been possible without you. Thank you for your ingenuity, kindness, and love, and I wish you the very best for your own PhD adventure.

Thank you!
Melbourne, February 2021
Danula Hettiachchi

Contents

List of Tables

Chapter 1
Introduction

1.1 Motivation

Crowdsourcing is the process of gathering the 'wisdom of the crowd' or input from a distributed workforce to perform a unified task [81]. It is an economical and efficient way to gather large volumes of data and thus it is widely used within the scientific community as well as industry [42]. Crowdsourcing can be used for various types of tasks, such as labelling the sentiment of a social media post [82] and drawing a bounding box around a target object in an image [159] as well as more complex tasks like software development [158]. Today, data gathered through crowdsourcing can end up moulding critical computing applications such as self-driving vehicles, recommendation systems and digital assistants [30]. For instance, a computer vision model that detects pedestrians requires a large volume of annotated training images, often sourced through crowdsourcing [165].

With the growing popularity of crowdsourcing, ensuring the quality of the collected crowd data has become an increasingly important research challenge. To this end, researchers have introduced numerous methods to enhance data quality [27]. For instance, there are methods that aim to improve task design and presentation, train or provide feedback to workers [43], use questions with known answers as quality checks [84], or process the collected data using different aggregation methods [174]. The applicability of data quality improvement methods can sometimes be limited depending on several factors like the nature of the task, the attributes of the crowd workforce, and the availability of resources. For example, task design approaches typically only work for a specific type of tasks, and it is expensive to generate questions with known answers, which are not available at scale. Therefore, researchers have investigated different quality improvement methods while aiming to maximise their applicability on a broad range of tasks.

Another class of methods aims to match workers with different tasks based a certain worker attribute, which we term as *'task assignment'* methods in this thesis. A basic crowdsourcing platform operates in a market model, where all tasks are visible to the worker pool and workers uptake tasks as they like. Task assignment attempts to change this model by matching workers with compatible tasks. Research shows that task assignment is an effective method that can enhance the accuracy of the task output, uplift the efficiency of the crowdsourcing process, and improve crowd workers' satisfaction [105]. However, due to variations in crowd tasks, the inconsistencies with the availability and the diversity of the worker population, optimal task assignment in crowdsourcing is known to be a challenging task [49].

There are two fundamental steps in task assignment, which are worker performance estimation, and assigning workers to tasks using a particular strategy that uses estimated

performance values. While workers' historical task performance has been used to estimate the performance effectively [132, 141], such historical data is sometimes not readily available, especially for new workers. Therefore, this thesis explores different ways of estimating worker performance and assigning tasks through such estimations without relying on the workers' historical performance data. In particular, we examine how we can use worker cognitive ability and context information for effective task assignment. These attributes have highly desirable properties. We can reliably gauge workers' cognitive ability through brief online cognitive tests, where it is difficult for workers to provide false input or signals. Similarly, context information is readily available and can be sensed through user devices, making them suitable for task assignments with broader applicability. Furthermore, we are interested in exploring how worker context-based task assignment could be beneficial when crowd tasks are available through various work devices. To this end, this thesis also investigates the feasibility of voice-based crowd work through smart speakers, an initial step towards a future where cross-device task assignment is possible.

1.2 Contribution

1.2.1 Research Questions

This thesis contributes towards improving data quality in crowdsourcing. The thesis particularly explores how worker cognitive abilities and context can be used for effective task assignment in crowdsourcing platforms. Researchers and practitioners can adapt our findings to make crowdsourcing platforms more valuable for task requesters by improving output quality, and make them more accessible for workers by elevating their experience and the worker-task compatibility.

Based on the research gaps identified in Section 1.1, the broader research question answered through this these is:

RQ: How can we improve crowdsourcing data quality through worker cognitive ability and context-based task assignment?

More specifically, I have dissected the research question into three main questions and sub-questions, which I discuss across four Chapters in the thesis.

RQ1 How can we improve crowdsourcing data quality by assigning tasks using online cognitive tests?

- 1.1 How can we estimate crowdsourcing task performance for specific tasks using cognitive test outcomes?
- 1.2 What is the data quality gain that we can achieve by assigning tasks based on worker cognitive abilities?

RQ2 How can we achieve dynamic online task assignment using worker cognitive ability?

- 2.1 What are the practical methods to test workers's cognitive ability and assign suitable tasks in an online setting?
- 2.2 How does the data quality gain compare with other mechanisms?

RQ3 How can we use the context of the worker to assign tasks effectively?

- 3.1 What is the impact of worker context on their willingness to accept varying crowd tasks when using different devices?
- 3.2 What are the benefits and implications for task assignment when using voice interaction as a novel modality for crowdsourcing?

These research questions are investigated through four different studies. First, we built our understanding of the relationship between cognitive abilities and crowdsourcing task performance through a controlled crowdsourcing study (Article I). We then leveraged the findings from Article I to create a dynamic online task assignment system that matches workers to compatible tasks using their cognitive test outcomes (Article II). We show that our task assignment method can outperform a baseline where workers select the tasks themselves, while also being comparable to other complex methods in the literature. Regarding RQ3, we investigated crowd workers' willingness to accept tasks presented on different work devices when their context varies (Article III). Furthermore, findings from this work then led us to explore in-depth voice-based crowdsourcing (Article IV). We built a crowdsourcing platform that works through a digital voice assistant, and established the feasibility of voice-based crowd work through a lab and a field study. Our findings highlight that such platforms enable workers to complete crowd work with greater flexibility and pave the way to cross-device task assignment.

1.2.2 Role of the Author

The thesis includes four publications [72, 73, 75, 76] as main contributions presented in Chapters 4, 5, 6 & 7. The articles are published in leading, international, and peer-reviewed conferences/journals in Human-Computer Interaction and Crowdsourcing; ACM CHI, ACM CSCW, AAAI HCOMP and INTERACT. I conducted the majority of the work (greater than 50%) and was the lead author for all the publications presented in this thesis. I initiated the studies by providing the primary concepts behind the work, prepared the experimental studies, and carried out administrative tasks, programming and development tasks, participant recruitment, study deployment and data analysis.

Furthermore, I was the corresponding researcher during the publication process. I handled the process of preparing the articles for submission and revising the articles based on peer-reviews. I received valuable suggestions and feedback from my co-authors on designing and executing the studies and analysing data. Also, the co-authors assisted when preparing the manuscript of the publications. Throughout the main chapters of this thesis, I use the scientific term "we" to reflect and acknowledge my co-authors' contribution.

1.3 Thesis Outline

The remainder of this thesis is organised as follows. Chapter 2 provides an overview of the related work in data quality in crowdsourcing with a specific focus on task assignment methods. Next, Chapter 3 describes the methodology followed in the studies presented in the thesis.

Chapters 4, 5, 6 & 7 present the four articles that investigate the specific research questions. Chapter 4 details a crowdsourcing study that investigates the relationship between online cognitive skills and common crowdsourcing tasks. Chapter 5 builds on the task-test relationship reported in Chapter 4 and presents a dynamic online task assignment and recommendation system.

With the goal of extending task assignment to a cross-device crowdsourcing paradigm, Chapter 6 examines how worker context information such as work location, time of the day can be leveraged for cross-device task assignment. Particularly, the article presents a study that explores whether workers are willing to accept crowd tasks presented on different devices, when their context varies. In Chapter 7, we present and evaluate a voice-based crowdsourcing that works through a digital voice assistant. We anticipate such platforms can provide greater flexibility to workers and enable cross-device task assignment.

Followed by the original research contributions, Chapter 8 discusses our results and findings in relation to the research questions. In addition, we set forth the implications of our findings for crowd platforms and practitioners, and reflect on future directions for crowdsourcing research regarding task assignment. Chapter 9 concludes with a summary of the thesis.

Chapter 2
Background

In this Chapter, we provide an overview of data quality improvement methods in crowdsourcing followed by a detailed literature survey on existing techniques that aim to match workers with compatible tasks and questions. We distinguish and review specific methods that solve task assignment, question assignment and plurality problems and discuss challenges in employing different worker performance estimation and assignment methods in a crowdsourcing platform.

Section 2.1 describes the method we followed to select the literature included in this chapter. Section 2.2 briefly reviews data quality improvement methods in crowdsourcing and Section 2.3 defines the four task assignment problems that we discuss in detail. Section 2.4 elaborates on worker performance modelling and estimation methods, which are two critical steps of task assignment. Then, Section 2.5 summarises task assignment approaches including heterogeneous task assignment, question assignment, the plurality problem and budget allocation methods. Finally, Section 2.6 provides an overview of task assignment methods available in existing crowdsourcing platforms.

The literature survey presented in this chapter is supplemented by related work sections in Articles I, II, III, and IV, where we discuss additional prior work in relation to the individual studies.

2.1 Literature Selection

We conducted an extensive literature search on the ACM Digital Library using a query that includes keywords 'task assignment', 'task routing' or 'data quality' and 'crowd*' in the Abstract. We included articles published from 2010 and retrieved 747 records. We reduced the resulting set of papers by limiting to publications from a list of conference and journals that, to best of our knowledge, publish work on crowdsourcing and related topics. Selected conferences were AAAI, AAMAS, CHI, CIKM, CSCW, ESEM, HCOMP, HT, ICDE, ICML, IUI, JCDL, KDD, SIGIR, SIGMOD, UbiComp, UIST, WI, WSDM and WWW. Selected journals were PACM IMWUT, PACM HCI, TKDE, TSC, VLDB. We also excluded workshops, demo papers, posters, extended abstracts, etc. Literature from specific venues that are not included in the ACM Digital Library (e.g., HCOMP) were manually screened and added to our dataset. Then, we carefully inspected the remaining of the papers and filtered out papers that were deemed to not be relevant. Furthermore, this chapter also includes several additional papers hand-picked due to their relevance to the topic.

2.1.1 Scope

Crowdsourcing extends beyond traditional online crowdsourcing using desktop or laptop computers. Other general types which can overlap include mobile crowdsourcing [136] (e.g., smartphones, tablets), situated crowdsourcing [60, 64, 80] (e.g., public displays), and spatial crowdsourcing [61, 160] (e.g., workers attempt location based tasks including physical tasks). Task assignment in crowdsourcing has also been investigated based on such domains. However, due to wide variations in techniques used in these different settings, we limit our scope to online crowdsourcing.

Crowdsourcing can also be broadly categorised as paid and unpaid crowd work based on the rewards received by workers. Paid work corresponds to crowdsourcing tasks where workers receive monetary rewards typically through a crowdsourcing platform which facilitates the payment process. Unpaid or voluntary crowd work is also completed in popular platforms and projects like Wikipedia[1], Moral Machine [8] and Test My Brain [57]. However, there are key distinctions in how you motivate unpaid and paid crowd work [63, 124, 147]. For example, in Test My Brain, workers get personalised feedback that help them learn more about their mind and brain. In this review, we primarily focus on methods and literature that investigate paid crowdsourcing tasks on commercial crowdsourcing platforms.

When we consider the type of work available on crowdsourcing platforms, they can range from micro tasks [37] such as labelling, ranking and classification to complex and long term tasks like software and web development tasks [158]. Our survey focuses on crowdsourcing techniques concerning tasks that can be completed in a single session, which constitutes the bulk of available crowd work.

2.1.2 Related Surveys

We also note several related survey articles that capture different elements of crowdsourcing. Daniel et al. [27] look at overarching quality enhancement mechanisms in crowdsourcing. Their survey organises literature under three segments: quality model, which describes different quality dimensions, quality assessment methods, and quality assurance actions. While Daniel et al. [27] summarise task assignment methods, they are not analysed in detail due to the broader scope of their survey.

Zheng et al. [174] examine 17 truth inference techniques such as majority vote, Zencrowd [31, 32] and Minimax [176]. The survey also presents an evaluation of different methods using five real work datasets. The primary focus of our survey lies outside truth inference methods. However, we provide a summary of truth inference methods in Section 2.2.3, under post-processing data quality improvement methods.

Li et al. [114] surveys crowdsourced data management with an emphasis on different crowd data manipulation operations such as selection, collection and join. Their survey organises prior work under quality, cost and latency control methods. Vaughan [165] also present a comprehensive review on how crowdsourcing methods can benefit machine learning research.

[1] https://www.wikipedia.org/

Overall, in contrast to prior literature reviews, this chapter sheds light on the task assignment problem in crowdsourcing and discusses related assignment based quality improvement methods.

2.2 Quality Enhancement in Crowdsourcing

As crowdsourcing typically relies on contributions from a diverse workforce where task requesters have limited information on the workers, it is important to employ data quality improvement measures [27]. In this section, we provide an overview of data quality in crowdsourcing.

In crowdsourcing, data quality is typically quantified via different attributes such as task accuracy, the response time of collected data, and cost-efficiency. Different quality improvement methods aim to improve one or more quality attributes. For example, the accuracy of a translation task can be enhanced in a cost-effective manner by employing workflow changes [6].

We note that quality improvement methods can differ from one another based on the following characteristics.

- *Applicability:* A quality improvement method can work for a specific type of task, a broader range of tasks or across all types of tasks. Universal methods are highly desired, yet can be costly and difficult to implement. For example, certain question assignment methods [49, 88] only work for multi-class labelling tasks. In contrast, worker filtering based on approval rate works for most tasks when worker-history is available.

- *Complexity:* Some quality improvement methods involve complex implementations that require substantial time and effort. Such methods are not suitable for one-time jobs. For example, it is not straightforward to implement crowd workflows that facilitate real-time discussions among workers [22, 79].

- *Effectiveness:* The effectiveness of quality improvement methods also varies. Effectiveness of a method can be quantified by measuring the quality attributes.

- *Cost:* There is an inherent cost attached to each quality improvement method. It is explicit for some methods (e.g., issuing bonus payments to workers), while others have indirect costs (e.g., infrastructure cost to capture and analyse worker behaviour data).

Generally, task requesters prefer quality improvement methods that are low in complexity, highly effective, economical and broadly applicable. However, methods that satisfy all these quality needs are scarce, and task requesters typically select quality improvement methods based on the specific task at hand, time and budget constraints, quality requirement and platform compatibility.

While there is a wide array of such quality enhancement techniques, based on the method execution phase, they can be broadly categorised into pre-execution methods,

online methods and post-processing techniques as detailed in Table 2.1. Given the standard crowdsourcing workflow, task requesters consider and employ pre-execution methods before task deployment. Fundamentally, through these methods, requesters specify how the task should be presented and executed in the crowdsourcing platform. Next, online methods alter the crowd task execution by dynamically deciding parameters such as the number of labels to collect, worker-task assignment, and task reward. Finally, post-processing methods examine how we can obtain better outcomes by processing the gathered crowd input. In this survey, we are primarily interested in online methods, however we briefly summarise pre-execution and post-processing methods in the following sub-sections.

Table 2.1: Quality Enhancement Methods

Pre-execution	Improve Task Design Train workers Improve extrinsic and intrinsic motivation
Online Methods	Heterogeneous Task Assignment Question Assignment Plurality Assignment Budget Allocation
Post-processing	Answer Aggregation (Truth Inference) Filtering workers

2.2.1 Pre-execution Methods

Data quality improvement methods employed at the pre-execution phase involve improving how workers interact with the task in terms of task design and crowdsourcing workflows.

2.2.1.1 Task Design and Crowdsourcing Workflows

Improving task design based on design guidelines and crowdsourcing best practices is one of the most well-known quality improvement methods. Research shows that clear task descriptions [55], data semantics or narratives that provide task context [41], and enhanced task user interfaces that improve the usability [1, 5] and reduce cognitive load [4] elevate data quality.

The outcomes of methods relating to task design can vary depending on the task itself. For example, Find-Fix-Verify [13] is a workflow introduced for writing tasks such as proofreading, formatting and shortening text. Iterate and vote is another design pattern where we ask multiple workers to work on the same task in a sequential manner. Little et al. [116] show that iterate and vote method works well on brainstorming and transcription tasks. Similarly, under map-reduce, a larger task can be broken down into discrete sub-tasks and processed by one or more workers. The final outcome is obtained by merging individual responses [24, 106].

Many other complex workflows have been proposed. For instance, the assess, justify & reconsider [45] workflow improves task accuracy by 20% over majority vote for annotation tasks. Several extensions to this method have been proposed such as introducing multiple turns [22, 152]. Annotate and verify is another workflow that includes a verification step. Su, Deng, and Fei-Fei [159] show that data quality in a bounding box task is improved when they employ the annotate and verity method with two quality and coverage assessment tasks followed by the drawing task [159].

More complex workflows that facilitate real time group coordination [12, 22, 152] can be challenging to incorporate into a crowdsourcing platform. Other variants include tools that allow workers [110] and task requesters (e.g., Retool [21], CrowdWeaver [104]) to design custom workflows. There is limited work that explores how to build and manage the crowdsourcing pipeline when employing a task workflow [161]. For example, the reward for each step can be dynamically adjusted to efficiently process the overall pipeline [131]. On the contrary, some work argues that static crowdsourcing workflows are limited in terms of supporting complex work and calls for open-ended workflow adaptation [146].

Other related task design and workflow improvements include gamification [62, 134] and adding breaks or micro-diversions [26].

2.2.1.2 Feedback and Training

Providing feedback to workers based on their work can improve the data quality in crowdsourcing. Dow et al. [43] report that external expert feedback and self-assessment encourages workers to revise their work. Dow et al. [43] highlight three key aspects of feedback for crowd work. 'Timeliness' indicates when the worker gets feedback (i.e., synchronously or asynchronously). The level of detail in the feedback or 'specificity' can vary from a simple label (e.g., approve, reject) to more complex template-based or detailed one to one feedback. Finally, 'source' or the party giving feedback, which can be experts, peer workers, the requester, or the worker themselves.

In a peer-review setup, the process of reviewing others' work has also been shown to help workers elevate their own data quality [177]. Similarly, Whiting et al. [168] show that workers achieve high output quality when they receive feedback from peers in an organised work group setting. While expert and peer feedback are effective in improving data quality, it is challenging to ensure the timeliness of feedback which is important when implementing a scalable feedback system.

It is also possible to deploy a feedback-driven dedicated training task and let workers complete multiple training questions until they achieve a specified data quality threshold. Park, Shoemark, and Morency [140] report that such a mechanism can be effective in crowdsourcing tasks that involve complex tools and interfaces. However, training or feedback may also bias the task outcome depending on the specific examples selected for the training/feedback step [113]. Feedback can also be used to explain unclear task instructions. For example, prior work by Manam and Quinn [123] proposes a Q&A and Edit feature that workers can use to clarify and improve task instructions or questions.

Other similar work tools that can potentially help improve data quality include third-party web platforms, browser extensions and scripts (e.g., Turkopticon [89], Panda

Crazy[2]) [96]. These tools provide additional information for workers to avoid substandard tasks and make their work more efficient.

2.2.2 Online Methods

While pre-execution methods focus on priming the task and workers, online methods aim to increase data quality by dynamically changing task deployment parameters and conditions like matching workers with compatible and relevant tasks. In this survey, we primarily focus on such online assignment methods, that we discuss in detail in the Sections 2.3, 2.4 & 2.5.

2.2.3 Post-processing Methods

Post-processing methods are employed after workers complete the entire batch of tasks in the crowdsourcing platform. A large portion of post-processing methods falls under answer aggregation techniques. We also discuss several other methods including filtering workers.

2.2.3.1 Aggregating Answers

Typically in crowdsourcing, we obtain multiple answers for each question. Once all the answers are collected, we need to aggregate them to create the final answer for each question. This process is also known as truth inference in crowdsourcing. There are many ways to aggregate answers, and task requesters may opt for different strategies depending on the task and data quality needs.

Majority voting is the most simple and naive, yet widely used approach for answer aggregation [174]. However, majority vote can fail when only a handful of highly accurate workers provide the correct answer. Prior work has proposed many extensions to majority voting. For example, instead of calculating the majority vote, the labels can be aggregated to a score that reflects the level of agreement [174]. Then, we can calculate the best threshold value to obtain the final answer. A training set or a gold standard question set can be used when determining the threshold.

Zhuang et al. [178] examined the bias that can be introduced into crowdsourcing when a worker provides answers to multiple tasks grouped into a batch, which is a common mechanism employed to reduce cost and improve convenience for the worker. They proposed an alternative to majority voting which could result in improved accuracy when batching is present. Ma et al. [121] proposed a truth inference method that is able to account for varying expertise of workers across different topics.

For rating and filtering tasks, Das Sarma, Parameswaran, and Widom [28] proposed an algorithm for finding the global optimal estimates of accurate task answers and worker quality for the underlying maximum likelihood problem. They claim their approach outperforms Expectation Maximisation based algorithms when the worker pool is sufficiently large. Further, in an extensive survey on truth inference, Zheng et al. [174] evaluate the performance of different truth inference algorithms.

[2]https://github.com/JohnnyRS/PandaCrazy-Max

2.2.3.2 Clustering

Kairam and Heer [92] proposed an automated clustering-based method as a design pattern for analysing crowd task responses. Using entity annotations of Twitter posts and Wikipedia documents, they identified systematic areas of disagreement between groups of workers that can be used to identify themes and summarise the responses.

2.2.3.3 Filtering Answers

After data collection, we can also remove specific responses to improve the data quality. For example, If we are able to identify malicious workers who may submit purposely inaccurate or incomplete responses, we can filter all the answers provided by such users during the aggregation process. Instead of using worker responses as the sole quality signal, Moshfeghi, Huertas-Rosero, and Jose [135] propose a method that uses task completion time to identity careless workers. Similarly, post-hoc worker filtering is also possible after estimating worker accuracy through different techniques, such as analysing worker behavioural traces [69, 149] and the worker network [109]. In Section 2.4.3, we discuss estimation methods in detail.

Furthermore, data quality can be impacted when workers use bots to provide automated responses or collude with other workers to share information [19, 38]. KhudaBukhsh, Carbonell, and Jansen [102] propose an unsupervised collusion detection algorithm that can help identify such workers and remove corresponding responses. It is also possible to detect colluding worker by analysing contribution similarity [94]. In addition, sybils or bots can be identified by estimating worker similarity and clustering them into groups [171].

2.3 Task Assignment Problems

Before we examine online methods in detail, it is important to identify the different stakeholders and parameters involved. We explain the crowdsourcing workflow, involved entities and different parameters that can be optimised in an online setting for task assignment purposes.

- *Requester:* A person who posts tasks on a crowdsourcing platform. Requesters reward the workers through the platform when they provide answers to their task.
- *Worker:* A person who completes tasks on a crowdsourcing platform in return for a reward. There is a large body of literature that examines characteristics of worker population [36, 148], work practices [170] and challenges faced by workers [151].
- *Task:* A collection of questions of the same task type. Prior work [54] has identified different task categories, such as verification and validation, interpretation and analysis, content creation, surveys, and content access.
- *Question:* An individual question within a task. For example, in an Audio Annotation task, this would be a single audio clip that requires an annotation.

An arbitrary number of answers can be collected for each question. Typically this threshold or the number of answers or labels required for each question is pre-determined by the requester.

- *Answer:* The answer provided by a specific worker to a specific question. Answer could take different forms depending on the task (e.g., a label 'Positive' for a sentiment analysis task). Typically in crowdsourcing, multiple workers provide answers for the same question. Numerous metrics such as accuracy, response time can be used to measure the quality of an answer.

- *Reward:* There can be intrinsic and extrinsic rewards [147]. The main reward mechanism used in crowdsourcing includes providing a pre-specified base payment and bonus payments issued at requesters discretion.

- *Crowdsourcing Platform:* Interaction between workers and task requesters is often managed by a third-party platform. For example, Amazon Mechanical Turk, Appen, Prolific and Toloka are commercial crowdsourcing platforms, that charge a fee from task requesters for managing the crowdsourcing workflow.

As detailed in Table 2.2, we use a consistent notation throughout the survey to describe different assignment problems.

Table 2.2: Notations used in the survey

Set of workers	$W = \{w_1, .., w_n\}$
Set of tasks	$T = \{t_1, .., t_n\}$
Set of questions for task t	$Q_t = \{q_1, .., q_n\}$
A task assignment	$\{t, w\}$
A question assignment of question q and worker w	$QA_{q,w}$
An answer provided by worker w to question q	$A_{q,w}$
Reward or payment for a question q	R_q

While the interaction between entities detailed above can vary depending on the specific crowdsourcing platform, next we summarise a typical crowdsourcing workflow. Task requesters first post their tasks in a crowdsourcing platform, with specific instructions and rewards for successful completion. Workers who have already signed up in the platform can browse and start working on tasks that they are eligible for. Eligibility constraints (e.g., location, skill and quality requirements) are often set by requesters or the crowdsourcing platform itself. Finally, when the work is completed, requesters can obtain the worker input or data contributions from the platform and compute the final output. Optionally, they may indicate whether individual worker answers meet their expectation. For instance, requesters can 'approve' or 'reject' answers. The crowdsourcing platform then transfers the reward to workers. This is similar to a first-come-first-serve or a market model.

Online assignment methods in crowdsourcing aim to alter this market model by directing workers to relevant and compatible tasks in order to increase the overall

data quality. At a high level, we identify and examine four key assignment challenges; heterogeneous task assignment, question assignment, plurality assignment problem and budget allocation.

2.3.1 Heterogeneous Task Assignment Problem

In this thesis, we explore heterogeneous task assignment or simply 'task assignment', which aims to select the best-suited task for a worker when there are different tasks available (e.g., Sentiment Analysis, Entity Resolution, and Classification).

Definition. *Assume that we have a set of tasks $T = \{t_1, .., t_k\}$ and a set of workers $W = \{w_1, .., w_m\}$ where $|T| = k$ and $|W| = m$. Each task t may contain an arbitrary number of questions. In order to maximise the overall quality of the data we gather, for each worker $w \in W$, we aim to assign the task t' where the worker is more likely to produce results of better quality.*

2.3.2 Question Assignment Problem

Select a specific number of questions from a task for a worker. For example, in a Twitter Sentiment Analysis task with 1000 tweets, the aim is to find specific tweets to assign for each worker.

Definition. *Assume that we have a set of questions $Q = \{q_1, .., q_k\}$ for a specific task t and a set of workers $W = \{w_1, .., w_m\}$ where $|Q| = k$ and $|W| = m$. In order to maximise the overall quality of the data we gather, for each worker, we aim to assign one or several questions where the worker is more likely to produce results of better quality.*

2.3.3 Plurality Assignment Problem

Deciding on the optimal number of workers that should be assigned to each sub-task or question is known as plurality assignment problem. Typically in crowdsourcing platforms, requesters manually configure a fixed number as the number of workers to be assigned for each task.

Definition. *Assume that we have a set of questions $Q = \{q_1, .., q_k\}$ for a specific task t and a set of workers $W = \{w_1, .., w_m\}$ where $|Q| = k$ and $|W| = m$. For each question $q \in Q$, multiple workers can provide answers (e.g., $A_{q,w1}$, $A_{q,w2}$, .. $A_{q,wx}$). We want to determine the ideal number of answers needed for each question q.*

2.3.4 Budget Allocation

The wide popularity of crowdsourcing is largely due to its economical nature when compared to other ways to acquiring large volumes of data. Hence, in addition to the data quality, budget allocation is an important factor in crowd work. Certain work considers budget allocation as part of the task or question assignment problem. For example, Assadi, Hsu, and Jabbari [7] investigate task assignment with the aim of maximising the number of tasks allocated with a fixed budget.

2.3.5 Task Recommendation

Instead of directly assigning workers to a compatible task, we can also provide users with compatibility information and let them select the task they wish to work on. Literature investigates this variation of task assignment as 'task recommendation [56]'. While task assignment aims to maximise the overall performance, task recommendation considers the potential benefits of providing workers autonomy and is regarded as a more flexible approach to match workers with compatible tasks. Nevertheless, worker performance estimation techniques related to task assignment are relevant in task recommendation as well.

2.4 Worker Performance Estimation

Worker performance estimation is a critical step in online assignment process. If performance estimations are unreliable, subsequent task, question or budget assignment decisions will not lead to desired quality enhancements. In this section, we discuss different metrics that can be used for estimation, data structures utilised for worker performance modelling and ways of estimating the performance.

2.4.1 Performance Metrics

2.4.1.1 Accuracy

Task accuracy is the most widely used performance metric in crowdsourcing. Accuracy is typically a number between 0 (incorrect) and 1 (correct) and can be defined in different ways depending on the task. For instance, for a classification task with single correct answer, accuracy of each question would be 1 if the worker provides the correct label and 0 otherwise. In contrast, a distant metric can define the similarity between text for translation tasks which results in a fraction. We use accuracy as the primary metric in the studies (Article I, II & IV) presented in this thesis. Other metrics that represent task accuracy include F-score [175], information gain [115] for multiple-choice tasks, mean Intersection over Union (mIoU) for image annotation tasks [139] etc.

2.4.1.2 Cost

While there are different crowd pricing mechanisms discussed in the literature [157], in a typical crowdsourcing platform, there is a pre-specified cost attached to each collected answer. However, other costs such as bonus payments, platform fees (e.g., MTurk[3]) can increase the total cost. Since crowdsourcing is often used for tasks with a large number of questions, cost is considered an important performance metric.

[3]https://www.mturk.com/pricing

2.4.1.3 Task Completion Time

When we consider task completion, there are two key metrics, time that workers spend on completing each question (i.e., work time) and total time that is needed to complete a task job that contains a set of questions (i.e., batch completion time). Both metrics can be optimised in different ways. Minimising work time is particularly helpful for tasks that require workers with specific skills or backgrounds [126]. In addition to task assignment, task scheduling strategies also aim to optimise batch completion time [35]. Crowdsourcing platforms typically provide task time information to requesters and they can also set a maximum time limit for each question.

In the studies (Article I, II & IV) presented in this thesis, we supplement task accuracy data with task completion times. In particular, we analyse task completion times in Chapter 7 to evaluate the feasibility of conducting crowd work through voice interaction. In addition, we leverage task completion time to identify and filter workers who provide non-serious responses.

2.4.1.4 Other Factors

Another indirect performance metric is worker satisfaction. Prior work highlights a relationship between crowd worker satisfaction and turnover [16], which may have an impact on data quality in the long run.

Some task assignment methods also consider special properties depending on the task. For instance, privacy preservation is an important performance metric for audio transcription tasks [18]. Others have considered the fairness [58], worker survival or likelihood to continue on tasks [107] and diversity in terms of worker properties [10].

2.4.2 Worker Performance Modelling

Based on the complexity and requirements of worker performance estimation method and the task or question assignment method, the literature proposes different ways to represent the quality of each worker, which we summarise below.

2.4.2.1 Worker Probability

The quality of each worker is modelled by a single attribute that describes the probability of the worker providing the true answer for any given question. This is a simple and widely adopted method [67, 119]. However, a single probability score is often insufficient to model the quality of the worker due to variations in question difficulty. The basic worker probability model can be extended by including a confidence value along with the probability value [90].

Instead of using a single probability value for all the tasks, worker probability can be modelled for each task (e.g., [132]) or question within the task (e.g., [49]). For example, quality of a specific worker could be 0.5 for sentiment analysis task and 0.8 for classification task. We use this approach in our cognitive ability based heterogeneous task assignment method presented in Chapter 5.

2.4.2.2 Confusion Matrix

Confusion matrix is extensively used to model worker performance for multiple-choice questions where each question has a fixed number of possible answers (e.g., [145, 166, 167]). Each cell (i, j) within the matrix indicates the probability of the worker answering the question with a label i given the true answer of the question is j. For the initialisation each worker could be assumed a perfect worker, values could be drawn from a prior distribution, or values can be estimated using gold standard questions.

2.4.2.3 Graph-based

In a graph-based model, workers or tasks are modelled as nodes in a graph (e.g., [17, 172]). Edges represent possible relationships among them. Different approaches are also possible. For instance, task assignments can be modelled as edges in a bipartite graph with both workers and questions as nodes (e.g., [97, 118]).

2.4.2.4 Tree-based

A tree-based model is a slight variant of the graph-based model. For instance, Mavridis, Gross-Amblard, and Miklós [126] use a skill taxonomy modelled as a tree where nodes represent elementary skills. Each worker also has a set of skills that they possess. A skill distance metric between the required skills for the task and the given skills of a worker is considered as the worker quality value for the particular task.

2.4.3 Performance Estimation Methods

Before assigning tasks or questions to workers, we need to estimate the performance of each worker. Estimations can be obtained by using objective measures such as gold standard questions, past/current task performance data, and qualification tests or by using worker characteristics or behavioural traits that are known to correlate with task performance. Table 2.3 organises prior work based on the performance estimation method.

2.4.3.1 Gold Standard Questions

Gold Standard Questions are questions with a known answer. It is common practice to use gold standard questions to estimate worker performance [119]. Typically, gold questions are injected into the task to appear among regular questions such that workers are unable to anticipate or detect gold questions.

When implementing gold standards, it is essential to know how we can inject these questions systematically. Prior work by Liu, Ihler, and Steyvers [117] investigates the optimum number of gold questions to use in a task. It is not beneficial to use a small number of gold standard questions in a large question batch. Workers could then collectively identify and pay more attention to gold questions making them ineffective as quality checks [19, 20]. Furthermore, creating ground truth data is not straightforward and crowdsourced tasks often do not have ground-truth data. Therefore, scalable and

Table 2.3: An overview of worker performance estimation methods used in online assignment methods.

Method		Literature
Gold Standard Questions & Qualification Tests		[88], [119]
Current Answer Distribution		[175], [101], [9], [144]
Worker Attributes	Demographics	[98], [154], [47], [36]
	Personality Tests	[99], [98], [120]
	Skills	[126], [111]
	Cognitive Tests	[59], Chapter 4, Chapter 5
	Work Device Features	[50], Chapter 7
	Worker Context	[86], Chapter 6
Worker Behaviour	Behavioural Traces	[149], [69], [51], [65]
	Social Media Data	[39], [172]

inexpensive methods of creating good gold data are necessary when using gold standards as a quality improvement method. Oleson et al. [137] present a programmatic approach to generate gold standard data. They report that a programmatic gold method can increase the gold per question ratio, allowing for high-quality data without extended costs.

Instead of populating gold questions before the label collection, we can also validate selected answers using domain experts. For instance, Hung et al. [85] propose a probabilistic model for classification tasks that help us find a subset of answers to validate through experts. The method considers the output accuracy and detection of inaccurate workers to find the most beneficial answer to validate. In addition, we can use domain experts to generate reliable and high-quality gold data [70].

Finally, in addition to measuring worker performance, gold standard questions can function as training questions that provide feedback to workers [52, 113].

2.4.3.2 Qualification Tests

Qualification tests contain a set of questions that workers need to complete before accessing the actual task. A qualification test can contain questions related to worker experience, background or skills that are needed for the actual crowdsourcing task [130]. For instance, a simple language skill test could be an appropriate qualification test for a translation task. A set of gold standard questions can also be presented as a qualification task. As answers are known a-priori, requesters can measure the performance in qualification test and allow a subset of workers to attempt the regular task. Crowdsourcing platforms such as MTurk supports qualification tests.

When using gold standard questions as a qualification test, there should be sufficient coverage of the different questions included in a task. Similarly, the qualification test should be challenging, such that workers are unable to pass is without understanding the task instructions fully.

When employing qualification tests, we can also ask workers to assess their own responses when ground truth data is not available or automated assessment is not feasible. Gadiraju et al. [53] show that self-assessment can be a useful performance indicator when we account for varying levels of accuracy in worker self-assessments.

2.4.3.3 Using Current Answer Distribution

In an ongoing task, we can also use the current answer distribution to estimate worker accuracy. Expectation Maximisation (EM) [29] is one of the most commonly used estimation methods to gauge worker performance for multiple class labelling questions (i.e., multiple choice questions) [175]. The method examines all the current answers and iteratively updates worker quality values and task answers until they converge. Khan and Garcia-Molina [101] used a different approach that uses Marginal Likelihood Estimation. They report that compared to Expectation Maximisation, Marginal Likelihood Estimation significantly reduces root mean squared error (RMSE) in predicting worker accuracy when there are few votes per worker. Raykar and Yu [144] consider a discrete optimisation problem and propose a Bayesian approach that can estimate a binary state that decides whether a worker is a spammer or not.

Estimating worker accuracy from current answer distribution is not exclusive to labelling tasks. Baba and Kashima [9] introduced a two-stage work flow with a creation and a review stage for tasks with unstructured responses, such as content generation and language translation. Their method uses the maximum a posteriori (MAP) inference to estimate the accuracy and model parameters.

2.4.3.4 Worker Attributes

When looking at task or question assignment from the workers' perspective, several worker attributes have been shown to have an impact on crowd task performance.

- *Demographics:* In a study of relevance labelling in crowdsourcing, Kazai, Kamps, and Milic-Frayling [98] reported a strong relationship between the accuracy of the crowd workers and their location. In their study, they used two task categories: full design (involves a number of quality controls such as challenge-response tasks and pre-filtering), and simple design (a simple version of full design tasks without strict quality controls) with 263 workers. They also showed that the majority of the simple design tasks were completed by Asian workers whereas full design tasks were mostly completed by American workers, possibly due to fewer Asian workers meeting the pre-specified qualification requirements in full design tasks. Similar results have shown that US workers perform significantly better than Indian workers in content analysis [154]. However, in an attempt to examine the preference for games over conventional tasks in relevance labelling, Eickhoff et al. [47] reported that there is no significant difference in performance of workers from US and India in Amazon Mechanical Turk. However, the researchers note that language skills of crowd workers and the differences in pay rates could also influence location-based performance variations. While these studies do not

attempt to match workers to tasks based on the said attributes, the results imply that using these approaches is feasible. Other work have also shown that worker demographics can introduce biases to the data collected [36, 74].

- *Personality:* Kazai, Kamps, and Milic-Frayling [99] analysed crowd users based on five personality dimensions introduced by Goldberg [91] known as the 'Big Five'. They further segmented workers into five types: Spammer, Sloppy, Incompetent, Competent and Diligent based on the personality and reported a significant correlation between the worker type and the mean accuracy of the worker. In a subsequent study, Kazai, Kamps, and Milic-Frayling [98] also reported that the Big Five personality traits - openness and conscientiousness - are correlated with higher task accuracy. Lykourentzou et al. [120] also examined the effect of personality on the performance of collaborative crowd work on creative tasks. They created 14 five-person teams: balanced (uniform personality coverage) and imbalanced (excessive leader-type personalities) considering only the outcome of 'DISC' [125] (dominance, inducement, submission, compliance) personality test and reported that balanced teams produce better work in terms of the quality of outcome compared to imbalance teams. They also reported fewer conflicts and higher levels of satisfaction and acceptance in balanced teams.

- *Cognitive Biases:* The study by Eickhoff [46] investigates cognitive biases and shows that cognitive biases negatively impact crowd task performance in relevance labelling. Cognitive biases are known as systematic errors in thinking and can impact peoples everyday judgements and decisions.

- *Cognitive Ability:* Alagarai Sampath, Rajeshuni, and Indurkhya [4] experiment with task presentation designs relating to cognitive features such as visual saliency of the target fields and working memory requirements. The study conducted on MTurk uses a transcription task and reports design parameters that can improve task performance. Goncalves et al. [59] investigated the impact of the cognitive ability of crowd worker performance and demonstrated that performance can be predicted from the results of cognitive ability tests. In their study, they used 8 cognitive tests which included visual and fluency tasks and 8 different crowdsourcing task categories (distance evaluation, item classification, proofreading etc.) attempted by 24 participants in a lab setting. However, they used time-consuming and paper-based cognitive tests from ETS cognitive kit [48] that are not practical for an online setting. In Chapter 4, we investigate the effect of cognitive abilities on crowdsourcing task performance in an online setting. Our work leverages the three executive functions of the brain (inhibition control, cognitive flexibility and working memory) [33] to describe and model the relationship between cognitive tests and crowdsourcing tasks. In Chapter 5, we propose a dynamic task assignment approach that uses cognitive tests.

- *Mood:* Prior work has also investigated if workers' mood has any impact on the crowdsourcing task performance [179]. While there is no evidence that shows a direct link between mood and task accuracy, the study reports that workers in

a pleasant mood exhibit higher perceived benefits from completing tasks when compared to workers in an unpleasant mood.

- *Work Device Features:* Gadiraju et al. [50] show that crowd work device and its characteristics such as screen size, device speed have an impact on data quality. The research also highlights that the negative impact of bad user interfaces is exacerbated when workers use less suitable work devices. In addition, device sensing capabilities and battery level can also impact the quality of crowd contributions [71]. Chapter 7 explores voice-based crowdsourcing, where workers complete crowd tasks through smart speakers and investigates if there is a performance difference compared to regular crowd work through desktop computers.

- *Worker Context:* Other contextual factors concerning the worker's current situation can also impact crowd task performance. Ikeda and Hoashi [86] show that task completion rate decreases when workers are busy or with other people. Also, worker context is a critical performance estimator for task assignment in spatial crowdsourcing, where tasks relate to a specific location [66]. In Chapter 6, we investigate workers' willingness to accept crowd tasks to understand the impact of context when tasks are available through a multitude of work devices.

- *Skills:* Prior work by Mavridis, Gross-Amblard, and Miklós [126] estimates worker performance using a distance measure between the skills of the worker and the skills required for the specific task. They use a taxonomy-based skill model. Similarly, Kumai et al. [111] model each skill with a numeric value. For instance, 1 minus the average word error rate (WER) of a worker's typing results can represent their typing skill.

2.4.3.5 Worker Behaviour

Prior work shows that worker behaviour data can be used to estimate worker performance [51, 65, 69, 149]. Rzeszotarski and Kittur [149] proposed 'task fingerprinting', a method that builds predictive models of task performance based on user behavioural traces. Their method analyses an array of actions (e.g., scrolling, mouse movements, key-strokes) captured while the user is completing crowdsourcing tasks. Task fingerprinting has been shown to be effective for image tagging, part-of-speech classification, and passage comprehension tasks in Amazon Mechanical Turk.

Han et al. [69] also reported that most of the worker behavioural factors are correlated with the output quality in an annotation task. Their method includes several additional features compared to the task fingerprinting method [149] and uses four types of behavioural features: temporal, page navigation, contextual, and compound. In a different approach, Kazai and Zitouni [100] show how we can use the behaviours of trained professional workers as gold standard behaviour data to identify workers with poor performance in relevance labelling.

While other methods [69, 149] aim to classify workers into either 'good' or 'bad' categories, Gadiraju et al. [51] classify workers into five categories using behavioural

traces from completed HITs. The study shows that significant accuracy improvements can be achieved in image transcription and information finding tasks by selecting workers to tasks based on given categories. To predict task and worker accuracy in relevance labelling tasks, Goyal et al. [65] use action-based (e.g., mouse movement in pixels in horizontal direction, total pixel scroll in vertical direction) and time-based (e.g., fraction of the total time that was spent completing the HIT, mean time between two successive logged click events) features in their predictive model. Goyal et al. [65] argue that worker behaviour signals captured in a single session can be used to estimate the work quality when prior work history is unavailable.

Behavioural data like social media interests captured outside the crowdsourcing platforms have also been used to predict task performance [39]. While this can be an interesting direction which attempts to create a global profile of the crowd worker, current strict privacy regulations would make practical implementation almost impossible.

2.4.3.6 Using a Combination of Estimators

Rather than using a single performance estimator, it is also possible to use a combination of different estimators. For instance, most of the expectation maximisation based methods use gold standard questions for initial estimation. Similarly, Barbosa and Chen [10] introduce a framework where the worker pool for each task can be constrained using multiple factors such as demographics, worker experience and skills. Their results show that worker selection with appropriate uniform or skewed populations helps mitigate biases in collected data.

2.4.4 Challenges and Limitations

While prior work reports promising results on using various worker performance estimation methods, there are many limitations when we consider implementation and broader adoption of such method.

Perhaps the most well-known estimation method is the use of gold standard questions. However, there are several fundamental limitations. First, gold standard questions are not broadly available for all tasks (e.g., tasks with subjective responses). Second, it can be costly to generate good gold questions. Third, gold questions are also susceptible to adversarial attacks. In an attack, workers detect and mark gold standard questions through various third-party tools such that subsequent workers can pay more attention to gold standard questions to amplify their measured performance [19]. Despite such limitations, the use of gold standard questions is an effective quality control method applicable to a broader range of tasks.

Worker attributes are also widely used to estimate the worker performance. Attributes like cognitive ability, personality and skills are preferred as they can be extended to estimate task performance across a wider range of tasks. Similarly, task requesters often use demographics (e.g., location, age, education level) as it is straightforward to use them. However, there are notable challenges in integrating certain worker attributes into a task assignment system. For example, attributes like demographics are self-reported by workers, allowing workers to provide incorrect information to gain undue advantages.

Comprehensive personality tests are time-consuming and there is also the possibility for workers to manipulate the outcome. Similarly, less competent crowd workers tend to overestimate their performance in self-assessments [53].

Numerous complications exist when concerning the use of worker skills [111, 126]. Workers need to list down their skills and such information should be available at platform level. We have to either assume that worker input related to skills are accurate or validate such information. Skill assessment can be a lengthy process increasing the barrier of entry for new workers. Also, requesters have to define which skills are required when creating new tasks.

While worker activity tracking [51, 65, 69, 149] has shown promising results, there are several practical limitations. First, such implementations often run as browser-scripts and can make the crowdsourcing platform interface resource intensive. This in turn can limit the accessibility of crowdsourcing platforms, particularly for workers with computing devices with limited capacities and low bandwidth internet connectivity. Second, behavioural data collection, data storage, and performance estimation can be computationally intensive for the back-end infrastructure of the crowdsourcing platforms, thus incurring additional costs. Third, there are privacy concerns with regard to tracking and storing activity data.

2.5 Task Assignment Methods

In this section, we discuss methods or frameworks that actively prevent contributions of sub-par quality by implementing various quality control mechanisms. In contrast to post-processing techniques, task assignment or routing methods can significantly reduce the overall number of answers required to obtain high quality output for crowd tasks. Thus, they can bring a financial benefit to task requesters. Also, task assignment can increase the compatibility between worker capabilities and task needs, potentially leading to increased worker satisfaction.

Literature presents a number of task assignment algorithms or frameworks that can be integrated with, or used in place of existing crowdsourcing platforms. They consider different quality metrics (e.g., accuracy, task completion time) and implement one or more quality improvement techniques (e.g., gold standard questions [44], removing or blocking erroneous workers [101]) to enhance the metrics. The primary motivation behind each assignment method can also be divergent. For example, some methods aim to maximise the quality of the output (e.g., [49, 150, 175]) while other methods attempt to reduce the cost by achieving a reasonable accuracy with a minimum number of workers (e.g., [101]).

We organise prior work under task assignment, question assignment and plurality problems we outlined in Section 2.3. Table 2.4 provides a brief summary of the worker performance estimation and assignment strategy of each method we discuss in this section.

Table 2.4: An overview of worker performance estimation and assignment strategies of assignment methods.

Assignment Problem			Reference	Performance Estimation	Assignment Strategy
Task	Question	Plurality			
✓			[78]	Requesters manually evaluate the answers	Based on the online primal-dual framework
✓			[77]	Using gold standard questions	By extending online primal-dual methods
✓			[7]	Using bids provided by workers	Maximises the number of tasks allocated within a budget
✓			[132]	Estimate using the performance in other tasks	Through a hierarchical Bayesian transfer learning model
✓			[34]	Use historic records to learn quality distributions	Model workers and tasks in a bipartite graph and use an adaptive, non-adaptive or greedy method to assign tasks.
✓			[39]	Using interested topics captured from social media	Rank available workers through category-based, expert-profiling and semantic-based assignment models.
✓			[126]	Through a distance measure between worker skills and the skills required for tasks	Targets skill compatibility. Assigns specialised tasks to workers with fewer skills first.
✓			[40]	Assumes that context-switching reduces worker satisfaction and performance	Scheduling tasks to maximise the likelihood of a worker receiving a task that they have recently worked on.
✓			[35]	Assumes that context-switching reduces worker satisfaction and performance	Schedule tasks prioritising currently running jobs and workers getting familiar work

✓			[18]	Estimate the loss of private information	A graph-based method that maintains privacy without starving the on-demand workforce.
✓			[111]	Estimate worker skills using a qualification task	Form groups of workers based on skill balance and worker re-assignments.
✓			[87]	Worker specified task interest and other factors such as skills	Uses different strategies depending on the task collaboration scheme.
✓			[153]	Assumes that expertise of each worker is a known numerical value	Sequential assignment based on budget, data quality and latency needs
✓			Chapter 5	Estimated using cognitive test outcomes	Select workers to maximise gain in accuracy
	✓	✓	[119] - CDAS	Injecting gold standard questions	Estimate the required answer count and use early termination.
	✓		[101] - CrowdDQS	Marginal likelihood curve estimation	Maximise gain in accuracy
	✓		[49] - iCrowd	Static gold standard questions & task similarity	Save questions for most accurate workers
	✓		[150] - OSQC	Hybrid gold plurality algorithm	Multi-rule quality control
	✓		[23] - OKG	Statistical inference with Beta distribution priors	Maximise gain in accuracy
	✓		[175] - QASCA	Expectation maximisation (EM)	Maximise gain in accuracy or F-score
	✓		[88] - Quizz	Estimate using only gold standard question responses	Maximise information entrophy
	✓		[58]	Estimate using limited gold standard questions	Maximise gain in accuracy while satisfying budget, fairness and diversity constraints.
		✓	[133]	Using gold standard questions	Estimate plurality form a greedy algorithm that assumes that answer quality increases monotonically at a decreasing rate with its plurality

✓	[156]	By modelling task difficulty and worker skills	Through an incremental Bayesian model that re-evaluate answer quality at each stage.
✓	[162]	By iteratively estimating worker expertise and question difficulty	Batch assignment maximising the number of questions completed in each batch.
✓	[2]	Assumes the past performance of a worker is known	Decide on when to stop assigning another worker

2.5.1 Heterogeneous Task Assignment

As crowdsourcing platforms contain a variety of tasks (e.g., sentiment analysis, classification, transcription), heterogeneous task assignment focuses on matching different task types with workers. Heterogeneous task assignment is the primary interest of this thesis and it can be particularly useful in cases where 'expert' workers must be allocated for more difficult tasks [78]. In addition to heterogeneous task assignment, crowdsourcing literature also explores question assignment, where questions within the same task (e.g., different questions of sentiment analysis task) are assigned to different workers to maximise the performance gain. We also review question assignment methods in Section 2.5.2.

Task assignment involves multiple steps. First, worker performance is modelled and estimated using different methods discussed in Section 2.4. Then, the task assignment process is carried to maximise the potential gain in terms of a specific performance criteria. For instance, one task assignment method could achieve modest data quality gains while minimising the overall cost. In contrast, another method could aim to achieve the highest possible data quality with a set budget.

Ho and Vaughan [78] propose a task assignment method based on the online primal-dual framework, which has been previously utilised for different online optimisation problems. The proposed Dual Task Assigner algorithm assumes that workers with unknown skills request tasks one at a time. In the study, researchers use three types of ellipse classification tasks to account for different expertise levels and use a translation task to simulate different skills. However, their approach assumes that the requester can immediately evaluate the quality of completed work. This vastly limits the applicability of their approach in a real-world crowdsourcing problem. Ho, Jabbari, and Vaughan [77] further investigate heterogeneous task assignment in classification tasks with binary labels. However, for the assignment, they use gold standard questions of each task type to estimate the accuracy of the workers.

We can also examine task assignment from the requester perspective. Assadi, Hsu, and Jabbari [7] propose an online algorithm that can be used by a requester to maximise

the number of tasks allocated with a fixed budget. In a different approach for task assignment, Mo, Zhong, and Yang [132] apply a hierarchical Bayesian transfer learning model. They use the historical performance of workers in a similar or different type of tasks to estimate the accuracy of the new tasks. Their experiment with a real-world dataset shows the effectiveness of the proposed approach when transferring knowledge from related but different crowd tasks (e.g., questions on sports vs makeup and cooking). However, their real-world evaluation is limited to a single scenario with one source task and one target task.

While most methods focus on a predefined set of tasks, Dickerson et al. [34] examine task assignment when tasks are not known a-priori. Their work proposes a novel theoretical model, called Online Task Assignment with Two-Sided Arrival (OTA-TSA), where both workers and tasks arrive in an online manner.

Data collected outside crowdsourcing platforms can also be used to match tasks with workers. Difallah, Demartini, and Cudré-Mauroux [39] present a system where tasks are allocated based on worker profile data such as interested topics captured from a social media network. Similarly, Zhao et al. [172] propose 'Social Transfer graph' for task matching. They demonstrate how tasks on Quora can be matched with Quora users' by extracting respective users' Twitter profile data (i.e., tweets and connections). The general applicability of such methods raises numerous practical and ethical considerations.

Mavridis, Gross-Amblard, and Miklós [126] introduced a skill-based task assignment model. Worker performance is estimated using a distance measure between the skills of the worker and the skills required for the specific tasks. The method attempts to assign the most specialised task first to the workers with the lowest number of skills based on the distance measure.

Task assignment can be challenging for more complex and collaborative tasks. Ikeda et al. [87] propose a task assignment framework that can decompose complex tasks and support sequential, simultaneous and hybrid worker collaboration schemes. Their assignment strategy selects a worker based on interests indicated by workers and their eligibility calculated using the project description and worker human factors (e.g., language capabilities). In contrast, Schmitz and Lykourentzou [153] look at non-decomposable macro-tasks like document drafting. They propose a sequential assignment model, where multiple workers attempt a task on a fixed time-slot, one after the other. At the end of each iteration, the next worker is selected if the task does not reach the desired quality threshold.

Instead of assigning tasks on the fly, it is also possible to schedule them when tasks are known apriori. Prior work by Difallah, Demartini, and Cudré-Mauroux [40] investigates task scheduling in crowdsourcing platforms and shows that scheduling can help minimise the overall task latency, while significantly improving the worker productivity captured through average task execution time. Research also highlights that scheduling is useful in ensuring tasks are fairly distributed across workers [35].

Addressing the growing concerns on crowdsourcing sensitive tasks like transcribing audio scripts, Celis et al. [18] examined task assignment with regard to trade-off in privacy. To preserve content privacy, we need to ensure that not too many parts of the same job are assigned to the same worker. They introduced three settings: PUSH, PULL, and a new setting, Tug Of War (TOW), which aims to balance the benefit for both

workers (by ensuring they can attempt a reasonable number of questions) and requesters (by minimising the privacy loss).

Instead of assigning tasks to individual workers, Kumai et al. [111] investigate the worker group assignment problem, where task requesters should select a group of workers for each task. They represent the worker accuracy using skills estimated through a qualification task and then forms groups based on three strategies that consider the skill balance among groups and the number of worker re-assignments.

2.5.2 Question Assignment

The aim of question assignment is to match workers with questions within a task such that we can obtain high quality output. Unlike in heterogeneous task assignment, we need to estimate worker performance and allocate tasks as workers complete submit answers to individual or batches of questions. Zheng et al. [175] present a formal definition of question assignment problem in crowdsourcing and show that optimal question assignment is an NP-hard problem.

Question assignment involves several fundamental steps. First, we should obtain a set of questions that are available to be assigned. Such candidate questions should not have been previously assigned to the current worker and should have available assignments with respect to the maximum number of answers required. Second, we estimate the performance gain (in terms of accuracy, for example) for each candidate question. Third, a subset of questions is selected to be assigned to the given workers.

Baseline approaches for question assignment are random assignment or a round robin assignment. Typical crowdsourcing platforms use these baseline approaches for question assignment.

2.5.2.1 Assigning questions to workers in a sequential manner

The question assignment problem can vary depending on the worker arrival assumption. The most practical problem is how to find a suitable question or a specific number of questions for an individual worker given a set of candidate questions. A naive way to assign questions is to enumerate all feasible assignments, calculate the performance gain for each assignment and then picks the assignment with the maximum performance gain. However, this method is computationally expensive and is not practical for typical crowdsourcing platforms where each task has a large number of questions.

Zheng et al. [175] proposed a question assignment framework (QASCA) which attempt to maximise either accuracy or F-score. For assigning k questions based on accuracy, the paper proposes the Top-K benefit algorithm which calculates the gain in expected number of correct answers for each question in candidate set and pick the questions which have the highest benefits. The algorithm has a time-complexity of $O(n)$ where n is the number of questions in the candidate set. A more complex online algorithm is presented for assigning questions based on F-score.

'CrowdDQS' proposed by Khan and Garcia-Molina [101] is a dynamic question assignment mechanism which examines most recent votes and selectively assigns gold standard questions to workers to identify and removes workers with poor performance in

real-time . They claim the proposed system which integrates seamlessly with Mechanical Turk can drastically reduce (up to 6 times) the number of votes required to accurately answer questions when compared to a round-robin assignment with majority voting. The proposed question assignment method aims to maximise the potential gain. The algorithm greedily chooses a question from the candidate set whose confidence score stands to increase the most if another answer is obtained from the considered worker.

Another dynamic question assignment method proposed by Kobren et al. [107] uses the worker survival metric (a user's likelihood of continuing to work on a task). Survival score is formulated using different measures such as accuracy, response time, difficulty of recently completed questions. The framework assigns questions to workers in order to achieve higher worker engagement and higher value for the task requester. Modelled using the markov decision process, the method aims to assign a question that maximises a combination of worker survival and expected information gain.

Different questions within a task may require knowledge and expertise on various domains. The task assignment method by Zheng, Li, and Cheng [173] attempts to organise questions and workers into different domains by building a knowledge base. Questions with uncertain true labels are then assigned to workers when their expertise overlap with the question's domain.

2.5.2.2 Question Assignment with a batch of workers

Another variant of the question assignment problem is to come up with an optimal assignment scheme given a set of workers and set of questions as opposed to assigning for a sequence of workers (e.g., [101, 178]). Cao et al. [17] termed this as the Jury Selection Problem (JSP) where they aim to select a subset of crowd workers for each question under a limited budget, whose majority voting aggregated answers have the lowest probability of producing an incorrect answer.

Fan et al. [49] introduced dynamic crowdsourcing framework named 'iCrowd' which assigns tasks to workers with a higher chance of accurately completing the task using a graph based estimation model. They consider the task similarity when estimating worker accuracy. The proposed question assignment strategy has three steps. First, it identifies a set of active workers who are ready to work on the task and dynamically finds sets of workers with the highest estimated accuracy for each available question. Then, the framework uses a greedy-approximation algorithm to formulate the optimum assignments ensuring each worker has no more than one question. Then, it strategically assigns gold standard questions to workers who are left without any question assignments.

'AskIt' proposed by Boim et al. [14] is another framework that achieves batch-wise question assignment. The assignment method aims to minimise the global uncertainty of entropy for questions while satisfying general assignment constraints such as maximum number of answers required for each question. Two metrics are proposed to measure global uncertainty that uses the difference between maximum and minimum entropy for individual questions. AskIt uses a greedy-heuristic to come up with the optimum assignment scheme. In addition, the framework employs an initial pre-processing step that uses collaborative filtering to predict missing answers and to identify questions that

are likely to be skipped by a specific worker. However, we note that the paper lacks details of the question assignment algorithm.

Goel and Faltings [58] proposed an algorithm for assigning tasks to workers, that optimises the expected answer accuracy while ensuring that the collected answers satisfy pre-specified notions of error fairness. The algorithm also limits the probability of assigning many tasks to a single worker, thus ensuring the diversity of responses. Question assignment is modelled as a constrained optimisation problem that finds the optimal crowdsourcing policy.

In a different approach, the method proposed by Li, Zhao, and Fuxman [115] assigns a portion of questions to the entire worker pool and estimates the accuracy for sub-groups of workers based on characteristics such as nationality, education level and gender. Then, the framework assigns questions to workers from the specific sub-group with the highest information gain. However, this method is not practical and cost effective when considering implementation on a crowdsourcing platform with a large number of workers from diverse backgrounds [36].

2.5.2.3 Blocking or Removing workers

Question assignment can also be achieved by blocking or removing workers from the pool of eligible workers as opposed to actively assigning questions to workers. CrowdDQS [101] uses this blocking technique to further improve assignment performance. Saberi, Hussain, and Chang [150] proposed a statistical quality control framework (OSQC) for multi-label classification tasks which monitors the performance of workers and removes workers with high error estimates at the end of processing each batch. They propose a novel method to estimate the worker accuracy – the hybrid gold plurality algorithm which uses gold standard questions and plurality answer agreement mechanism. Question assignment is based on a Multi-rule Quality Control System which assigns a value (0,1) to the worker at the end of each batch based on the past error rate and the estimated current error rate. Early termination is also another similar strategy where workers can no longer provide answers to a particular question which already has an answer with sufficient certainty [119].

2.5.2.4 Question Assignment with Budget Constraints

Qiu et al. [142] investigate binary labelling tasks. Their proposed method uses previously completed gold standard questions and estimated labels from task requesters to calculate the historic error rate for workers. Then, predicts worker error rate for upcoming questions, through an auto-regressive moving average (ARMA) model. Questions are assigned by maximising the accuracy with respect to the limited budget when worker payment is not constant.

Rangi and Franceschetti [143] approach task assignment with a multi-arm-bandit setup and propose using the simplified bounded KUBE (B-KUBE) algorithm as a solution. In their method, workers indicate their interest in doing the tasks, quote their charges per task, and specify the maximum number of questions they are willing to answer. Worker accuracy is estimated using the current answer distribution.

Similarly, Singer and Mittal [157] propose a pricing framework when workers bid for tasks with their expected reward and the number of questions they wish to uptake. Their method aims to maximise the number of questions completed under a fixed-budget or minimise payments for a given number of tasks.

2.5.2.5 Assigning Gold Standard Questions

Some frameworks have investigated whether to assign a golden standard question or a regular question when a worker requests a task. Ipeirotis and Gabrilovich [88] presented 'Quizz', a gamified crowdsourcing system for answering multiple choice questions. The framework uses a Markov Decision Process to select the next action.

2.5.2.6 Other Approaches

Kang and Tay [95] introduce a game based sequential questioning strategy for question assignment in multi-class labelling questions. They convert the questions into a series of binary questions and demonstrate the reliability of their proposed approach which considers worker responses at each step.

2.5.3 Plurality Assignment

Crowd task accuracy can be improved by obtaining multiple answers from different workers for the same question. In a typical crowdsourcing platform, the number of answers required for each task is set by task requester prior to the task deployment. However, due to variations in worker capabilities and question difficulty [162], some questions may require more answers, whereas few answers would be sufficient for the others. Crowdsourcing research that addresses the plurality assignment problem [133] aim to dynamically decide how many answers are needed for each question.

For binary labelling tasks, Liu et al. [119] estimate the number of answers required for each question before conducting question assignment. They introduce two prediction models (basic model and an optimised version) that use workers' accuracy distribution. As such accuracy distributions are generally not available in crowdsourcing platforms, a sampling method is used to collect the accuracy of available workers.

Mo et al. [133] propose a dynamic programming based approach to address the plurality assignment problem while maximising the output quality under a given budget. The paper identifies two key properties in crowdsourcing tasks, monotonicity and diminishing returns that describe a question with the final answer quality increasing monotonically at a decreasing rate with its plurality. They also propose an efficient greedy algorithm that can provide near optimal solutions to plurality assignment problem when monotonicity and diminishing returns properties are satisfied.

Similarly, Siddharthan et al. [156] present an incremental Bayesian model that estimates the plurality for a classification task with a large number of categories. Results obtained through their method outperform majority voting and is comparable to a different Bayesian approach (i.e., standard multinomial naive Bayes (MNB)) that uses a larger fixed answer count.

Worker expertise and question difficulty are two key variables that impact the confidence of an answer and plurality. In a batch-processing approach, prior work by Tu, Cheng, and Chen [162] efficiently estimated these two parameters to maximise the number of questions reliably answered at the end of each batch.

Instead of determining the plurality before task deployment, we can dynamically decide and limit the number of answers that we collect for each question. Abraham et al. [2] proposed an adaptive method that considers the differences and uncertainty of the answers provided and decide on when to stop assigning another worker for the task.

2.5.4 Challenges and Limitations

We discuss general challenges and limitation in task assignment methods. There is no straightforward, low-cost and effective solution for task assignment [49]. Therefore, each method and evaluation has their merits and limitations.

Concerning worker accuracy estimation, some studies infer worker quality instead of objectively estimating them. For example, Saberi, Hussain, and Chang [150] evaluate their statistical quality control framework proposed with crowd workers on Mechanical Turk where they simulate the past error rates of workers who completed the task using a standard normal distribution. Similarly, prior work by Schmitz and Lykourentzou [153] treats the work quality assessment step as a black-box process and assumes the expertise of each worker as a known numerical value. In both cases, it is difficult to argue that findings of such studies hold in real-word crowd platforms due to broader variations in crowd worker quality.

Some studies (e.g., [7, 14, 77]) evaluate task assignment methods using synthetic data instead of using a real-time deployment or a crowdsourced dataset. Furthermore, as popular crowdsourcing platforms including Amazon Mechanical Turk do not provide sufficient means to dynamically assign tasks, all the aforementioned studies (e.g., [49, 101, 175]) have evaluated their proposed frameworks using the external question feature of these platforms. While this is the standard for crowdsourcing research, it is unclear how worker behaviour in controlled studies compares with regular task performance.

While certain assignment methods (e.g., [101]) use random or fixed values for the initial worker accuracy, other methods (e.g., [49, 88]) use gold standard questions. Gold standard questions are widely used in crowdsourcing platforms. However, as discussed in Section 2.4, there are inherent limitations that make the use of gold questions less desirable. Also, some other methods use historic records [119] and suffer from the cold-start problem. These methods do not work with new workers in a crowdsourcing platform.

2.5.4.1 Heterogeneous Task Assignment Challenges

Different worker performance estimation strategies (e.g., transfer learning from similar tasks [132], worker attributes [126], Chapter 5) are useful for task assignment. Literature only shows that they can work on specific task types. For example, real world evaluation by Mo, Zhong, and Yang [132] is limited to a single source and target task pair.

Overall, heterogeneous task assignment is a highly desirable approach that can potentially work across a broader range of tasks. However, more evidence and experiments are needed to show that they work with various tasks (e.g., Work presented in Chapter 5 uses four types of common crowdsourcing tasks) and can sustain performance over time.

2.5.4.2 Question Assignment Challenges

Question assignment methods continuously monitor worker answers and create assignments at each step, making them typically more effective than heterogeneous task assignment methods. However, key challenges in adopting question assignment are the complexity in implementation and the cost of calculating the assignments. For example, even with an efficient question assignment algorithm solution such as QASCA [175], assignment time linearly increase with the number of questions. Therefore, computational complexity is an important factor to consider when employing question assignment methods in a real world system.

The majority of question assignment methods are also limited to multi-class labelling problems [88, 101, 107, 175]. While literature argues that other types of tasks (e.g., a continuous value) can be converted to multi-class or binary labelling problems [175], there is no research that shows that question assignment methods can work in such cases.

2.5.4.3 Plurality Assignment Challenges

Plurality assignment is an important problem in crowdsourcing. Proposed methods aim to estimate plurality either upfront [119] or during the task execution [2, 162] which can help reduce the overall cost for task requesters. Similar to question assignment, estimating plurality is often investigated considering multi-class labelling questions. While it is feasible to estimate plurality for labelling questions, it is far more complicated for crowd tasks that involve complex inputs, such as audio tagging and semantic segmentation. However, plurality assignment solutions are also more valuable for such tasks as each response involves a higher work time and reward.

As plurality assignment solutions do not achieve specific worker-question match, they are less complicated than question assignment methods when we consider practicality. Plurality assignment solutions can also be more effective when implemented together with question or task assignment methods [119]. However, further research is needed to ensure their utility in a dynamic online setting.

2.6 Crowdsourcing Platforms

In this section, we briefly review existing crowdsourcing platforms and standard task assignment mechanisms available in them. At a high level, current crowdsourcing platforms do not support complex task assignment methods proposed in the literature. However, certain functionalities and limited assignment methods are available to task requesters.

In Amazon Mechanical Turk[4], requesters can use task pre-qualifications to limit the workers who are able to see and attempt their task. The platform provides a set of pre-specified qualifications such as worker historical approval rate, location and sex. In addition, task requesters can create custom qualification and include workers based on previous tasks or qualification tests. Further, by using MTurk API and other third-party libraries and tools (e.g., PsiTurk [68]), task requesters can build advanced task assignment methods on top of MTurk.

Toloka by Yandex[5] is another popular crowdsourcing platform. Toloka allows task requesters to set-up worker skills that gets automatically updated based on the rate of correct responses (with gold standard questions, majority vote, or post-verification) and behavioural features like fast responses. Requesters can also configure rules based on skills. For example, rules could automatically block workers from the task if their skill level drops below a given threshold[6]. In addition, Toloka also provides a feature called 'incremental relabeling' to facilitate dynamic plurality.

Prolific[7] is another crowdsourcing platform that is tailored for surveys and research activities. The platform provides more than 100 demographic screeners to ensure the task is assigned for a restricted worker pool.

Certain other commercial crowdsourcing platforms such as Scale[8], Appen (previously Figure Eight and CrowdFlower)[9] and Lionbridge AI[10] focus on providing an end-to-end service to task requesters. They use a combination of crowdsourced and automated approaches to complete the task. While implementation details are not available, such platforms also utilise task assignment strategies where they use automated appraoches for simpler elements of the work pipeline and get crowd workers to attempt difficult parts such as quality control, edge cases, and complex data types[11].

Further, in crowdsourcing platforms that focus on complex tasks and projects (e.g., Upwork, Freelancer, Fiverr), task assignment is explicit. Task requesters examine the candidate workers who express willingness to complete the task and assign the task to one or more worker based on their profile. This manual assignment process is only practical for complex tasks that involve specialised workers, longer task times and higher rewards.

2.7 Summary

Data quality improvement methods are employed at different stages of the crowdsourcing life cycle. In this chapter, we provide an extensive overview of online task assignment methods in crowdsourcing that are employed during task deployment. Starting with a succinct overview of data quality improvement methods in crowdsourcing, we dissect

[4]https://www.mturk.com/
[5]https://toloka.ai/
[6]https://toloka.ai/crowdscience/quality
[7]https://www.prolific.co/
[8]https://scale.com/
[9]https://appen.com/
[10]https://lionbridge.ai/
[11]https://scale.com/blog/scaling-menu-transcription-tasks-with-scale-document

online methods into heterogeneous task assignment, question assignment and plurality assignment problems.

We discuss the challenges and limitations of existing task assignment methods, particularly regarding their applicability, complexity, effectiveness, and cost. We highlight how heterogeneous task assignment methods are generally less complex and applicable to many task types than question assignment and other approaches.

We also review a wide array of worker performance estimation methods that are essential for task assignment. We noted shortcomings of using gold standard questions, current answer distribution and worker behaviour data. For example, gold data is not readily available for all tasks, while other approaches require complex implementations. However, worker attributes, such are personality and demographics have emerged as promising worker quality predictors. In particular, worker cognitive ability and context are two properties that are easy to capture and difficult for workers to manipulate. In the following chapters, we present our findings on leveraging worker cognitive ability and context-based heterogeneous task assignment as a way forward to mitigate critical problems with existing methods.

Chapter 3
Methodology

This chapter describes the methodological decisions applied during the studies described in this thesis. Specifically, we explain and motivate the study design, task selection, quantitative and qualitative data collection methods used, and our approach to data analysis. The research presented in this thesis primarily engaged crowd workers as study participants. We explain the procedure we followed to ensure the consistency and ecological validity of our crowdsourcing experiments. Further, we discuss why and how we conducted a lab study and a field deployment to study our voice-based crowdsourcing approach when a crowd deployment was not feasible. This chapter primarily focuses on methodological factors that concern the complete thesis. Readers can find specific details regarding individual studies in the respective articles included in the subsequent chapters.

3.1 User Studies

3.1.1 Crowdsourcing Tasks

This thesis investigates task assignment in crowdsourcing using common crowdsourcing tasks. In each study, we selected the specific task set informed by prior work on crowdsourcing task performance [59], crowd task taxonomies [54], task availability [37] and specific factors we investigate in the study. Similarly, we used cognitive tests that are well-established in Psychology literature [33] and also validated in crowdsourced experiments [25].

In Article I and Article II, we used five common crowdsourcing tasks (classification, counting, sentiment analysis, proofreading, and transcription) that are representative of typical tasks available in crowdsourcing platforms. All tasks contained questions of varying complexity. In Article III, we were interested in task presentation in relation to different crowdsourcing devices and worker contexts. Thus, we use three task categories based on the media-type included in the task. In each category, we have two tasks with high and low complexity. We use text-based (sentiment analysis and information finding), audio-based (audio tagging and speech transcription) and image-based (image classification and bounding box) tasks. In Article IV, we evaluate the difference between voice-compatible and voice-based tasks. Voice-compatible tasks are standard text-based crowdsourcing tasks that can be completed via voice. For example, we used sentiment analysis, comprehension, and text moderation. In contrast, audio annotation, speech transcription and emotion labelling are voice-based tasks where workers have to listen to an audio clip to complete the task.

Throughout all the studies, we used task accuracy and completion time as key performance metrics. Task accuracy for all crowdsourcing tasks was measured and scaled to a value between 0 and 1. Additional details regarding tasks, cognitive tests and their performance metrics can be found in the respective publications.

3.1.2 Crowdsourcing Studies

In this thesis, we explore the data quality of crowd worker contributions. Therefore, it was essential to engage regular crowd workers as participants of our studies whenever possible. We used Amazon Mechanical Turk as the crowdsourcing platform for all the crowd studies presented in this thesis. Mechanical Turk is a well-established platform that is widely utilised in crowdsourcing research [37, 103].

Deploying controlled user studies in crowdsourcing platforms is challenging [103]. To ensure the data consistency, we took several measures and followed best practices in deploying crowdsourced studies. First, for all crowdsourcing studies, we recruited workers from the US. Except when the study conditions specifically required not to use pre-qualifications, we also limited our tasks to workers who have completed at least 1000 questions with a 95% approval rate [141]. Second, we included detailed task instructions and examples to mitigate any potential misunderstandings. Third, we obtained responses from a higher number of individual workers. For example, Article II involved 574 crowd workers. Fourth, all our crowdsourcing studies were seamlessly integrated with Amazon Mechanical Turk platform (e.g., using PsiTurk [68]), such that crowd workers were not required to leave the platform to complete our tasks. Finally, when deploying different study conditions or iterations, we ensured the consistency in terms of the time and date of deployment window. Specifically, we deployed our tasks between 9 am - 5 pm Pacific Time during weekdays. In addition, we took other measures specific to the individual studies, that we describe in the respective publications.

3.1.3 Lab Study and Field Deployment

In Article IV, we investigate the feasibility of voice-based crowdsourcing. To this end, we built an application that runs through a digital voice assistant. First, we needed to conduct a controlled within-subject user study to compare the task performance between voice-based and regular screen-based platforms. While we wanted to involve crowd workers, it was not viable to deploy this experiment in a crowdsourcing platform. We have no visibility into the crowd user environment or the capacity to ensure a consistent voice-interaction due to variations in worker devices. Therefore, we deployed a controlled lab study with locally recruited users. In order to further evaluate the system, we complemented our lab study with a field deployment where users were given a smart speaker (google assistant) device and asked to use our crowdsourcing platform at their home over a week.

3.2 Data Analysis

The findings presented in this thesis are mainly derived through quantitative methods. We measured task performance data, such as task accuracy and completion time, across different tasks. For the crowdsourcing studies deployed on Amazon Mechanical Turk platform (Articles I, II and III) , we involved a large number of participants.

In Article I and Article II, we gathered additional qualitative data through questionnaires and semi-structured interviews. During the analysis, we used qualitative data to complement quantitative findings and provide broader insights.

3.2.1 Quantitative Analysis

In this section, we provide an overview of quantitative methods used in the different studies presented in this thesis. Following research standards in Human-Computer Interaction [112], we quantify statistical significance and reject our null hypothesis using a confidence level of 95% (p-value smaller than or equal to 0.05).

In Article I, we used fundamental machine learning techniques and statistical modelling such as Beta regression, Generalise Linear Regression and Random Forest. These methods allow us to predict an outcome variable and explain the effect of each of the input variables (predictors) on the dependent variable. We also utilised correlation analysis to understand the relationship between two variables: cognitive test scores and crowd task accuracy to identify a general relationship between the two attributes. Additionally, we used Principal Component Analysis to further visualise and interpret our data.

For Article II, we employed traditional frequentist statistics, using tests such as Wilcoxon signed-rank test and Kruskal-Wallis rank-sum test for statistical hypothesis testing. We use non-parametric tests based on our data distributions. Frequentist statistics are regularly used in HCI research [112], to compare differences in an outcome variable (e.g., task accuracy) between different study conditions (in this case, the different task assignment methods).

In Article III, we used binomial generalised linear mixed model (GLMM) with maximum likelihood (Laplace Approximation) [11]. In this study, we collected an arbitrary number of responses from each crowd worker. GLMM models allow us to explore the impact of regular dependent variables or fixed effects on a target predictor while accounting for individual variations by including random effects.

We also used traditional frequentist statistics to analyse data from the lab study and the field deployment reported in Article IV. In addition, we used descriptive statistics and visualisations from exploratory analysis to further clarify our findings.

3.2.2 Qualitative Analysis

In Article III and Article IV, we employed qualitative analysis to further strengthen and explain our quantitative findings. Specifically, we used 'in vivo coding' and formal thematic analysis. Both methods are standard qualitative analysis techniques widely used in Human-Computer Interaction research [15, 129].

After we collected the interview data, we transcribed them and applied 'open coding' to identify relevant common concepts. In open coding, we analysed and label interview data, which helped us organise similar concepts. We extracted label names directly from the interview data, which is described as 'in vivo coding' [129].

The formal thematic analysis involves six steps [15]. First, we browsed and familiarised ourselves with the interview data. Then, we formed initial codes following the in vivo coding method mentioned above. After generating the codes, we searched for, reviewed and defined the themes in three steps. Finally, we composed the results using the final themes and related interview data. Also, we involved multiple authors in the coding process to ensure that we do not neglect or miss important potential themes.

3.3 Ethical Considerations

We took multiple precautionary steps to manage risks and avoid any potential issues during our studies presented in this thesis.

During all studies, we anonymised user contributions, such that it is not possible to link them with any personal data and identify participants or crowd workers at a later stage. For data storage and processing, we use firewall-protected secure servers. Server access is managed by an authentication mechanism which requires users to have a secure password and change the password every six months. In addition, our participants had the liberty to withdraw from the study at any point in time. They were also able to request to remove their data, during or after the study. As an essential step during all studies, we obtain informed consent from all participants. For the online crowdsourcing experiments, we obtain this through an electronic consent form and a plain language statement.

Fair compensation is an important consideration in crowdsourcing studies [151, 169]. We limited the worker location to the US and calculate the payment based on the highest state minimum wage in the US and expected task time estimated through our pilot deployments for all our crowdsourcing studies. We also promptly attended to any worker queries and avoided rejecting any work contributions unreasonably.

We took additional precautions with our study that involved digital voice assistants. While digital voice assistants running on smart speakers are always-on listening devices, we could only access participant interaction data recorded after they initiated our application through a specific voice command. We only obtained text transcriptions of user utterances provided by the digital voice assistant. We did not store, record or access voice data of our participants. We also explicitly configured the digital assistant such that it does not store voice-recordings of our participants in the cloud storage. During our initial briefing, we clearly explained this to our participants and obtained their informed consent.

We are aware that considerable personal information about participants availability at home can be derived by examining application usage patterns during the field study. Also, a false initiation or an incorrect application launch may record unexpected personal conversations or information. However, there is no other reasonable alternative approach to study the feasibility of using digital voice assistants for a specific purpose in a more

ecologically valid setup. Therefore, studies involving digital voice assistants have to consider such ethical considerations and potential risks carefully and take precautionary measures as needed to ensure the users' privacy.

3.4 Limitations

Although we carefully designed and executed studies presented in this thesis, we acknowledge several limitations.

In Article I and Article II, we used brief cognitive tests with a minimum number of trials in each test. While large trial numbers can make test effects more discrete and precise, the limited number of trials was sufficient for our predictions. Additionally, we do not account for subtle variations in human cognitive ability detected throughout the day. Also, like any other supervised learning method, we need to initially train our model with data captured from a set of workers performing cognitive tests and a set of crowdsourcing tasks. In Article II, we discuss different ways to obtain this necessary training data.

Concerning the study in Article II, our evaluation is limited to a real-world deployment and does not include any simulations with synthetic data. While many online task assignment methods experiment with synthetic data to extensively test parameter variations, it is challenging to synthesise cognitive test outcomes of workers, which is essential to test our system. Also, we compare our cognitive test based assignment system with history-based and question assignment methods, but not with any prior heterogeneous task assignment methods [77, 132]. Such methods are either incompatible with our study setup or excessively complex in implementation which reduces their applicability for real-world scenarios.

In the crowdsourcing study presented in Article III, we included questions regarding crowd tasks presented through smart speakers. As no such commercial solutions are available, workers do not have any experience through such applications, impacting their decision to accept or reject the given task. Also, we investigated a limited set of contextual and task factors in our study. As we already included over 5,000 unique HITs, additional factors would have unnecessarily complicated the experimental setup. Further, only a subset of workers who completed the initial task completed the post-task survey, which provided the qualitative data analysed in this study.

While our interests lie mainly on investigating task assignment with real crowd workers, it was not feasible to recruit actual crowd workers to evaluate the voice-based crowdsourcing platform described in Article IV. Therefore, we locally recruited users, and conducted a lab study and a field deployment. We acknowledge that evaluating our system with crowd workers would yield important insights that we were unable to capture. Additionally, our field deployment duration was limited to a week as we used a finite set of questions used in the lab study for greater comparability. We note that it is possible to obtain further insights on designing voice-based crowdsourcing platforms with a longitudinal study that involves more users.

3.5 Conclusion

This chapter presents an overview of fundamental methodological approaches employed in the studies presented in this thesis. In summary, we used rigorous crowdsourcing studies with several measures to ensure a fair comparison across experimental conditions, and avoid biases introduced from the crowd workforce. Similarly, we followed a strict protocol during our lab and field deployments. The thesis aims to identify ways to match crowdsourcing tasks with workers to achieve improved data quality through these studies. The following Chapters 4, 5, 6, and 7 include scientific articles published at leading peer-reviewed HCI venues, which address the proposed research questions.

Chapter 4

Cognitive Abilities and Crowdsourcing Task Performance

4.1 Introduction

Previous work has investigated the impact of different worker attributes, such as location [36, 154], personality [99, 120] and behavioural traces [69, 149] on crowdsourcing task performance. While these attributes have been shown to be promising performance indicators, there are many shortcomings in integrating them into a crowdsourcing platform. For example, personality tests are time-consuming, and workers can easily manipulate the outcome. Furthermore, capturing and analysing behavioural traces requires complex implementations. Meanwhile, limited studies have investigated cognitive ability [59], a highly desirable worker attribute that can be objectively measured through fast-paced online tests and that can be used to achieve greater person-job compatibility [108]. This chapter investigates the relationship between cognitive tests and crowdsourcing task performance, in order to enable a more suitable assignment of crowd tasks.

Informed by literature in Psychology, we hypothesise the relationship between specific cognitive tests and crowdsourcing tasks through the executive functions of the brain [33]. To test our hypotheses, we deployed a crowdsourcing study where crowd workers complete five cognitive tests (Stroop, Flanker, Task-Switching, N-Back, Pointing), followed by five common crowdsourcing tasks (Classification, Counting, Sentiment Analysis, Proofreading, Transcription). Our results indicate a correlation between crowdsourcing task performance and worker cognitive ability. More importantly, we confirm that workers are not universally 'good' or 'bad' across all tasks. We show that workers' cognitive ability measured through online cognitive tests is an effective signal to match workers with suitable crowd tasks.

More details of the study can be found in the attached publication, Article I. The findings from this study also form the basis of our approach to dynamic task assignment through cognitive tests, which we describe in Chapter 5.

4.2 Article I

Hettiachchi D., van Berkel N., Hosio S., Kostakos V., Goncalves J. (2019) Effect of Cognitive Abilities on Crowdsourcing Task Performance. In: Human-Computer Interaction – INTERACT 2019. INTERACT 2019. Lecture Notes in Computer Science, vol 11746. Springer, Cham. https://doi.org/10.1007/978-3-030-29381-9_28

Ethics ID: 1852019, The University of Melbourne Human Ethics Advisory Group.

Effect of Cognitive Abilities on Crowdsourcing Task Performance

Danula Hettiachchi[1], Niels van Berkel[1], Simo Hosio[2], Vassilis Kostakos[1], and Jorge Goncalves[1]

[1] School of Computing and Information Systems, The University of Melbourne, Australia
{danula.hettiachchi, niels.van, vassilis.kostakos, jorge.goncalves}@unimelb.edu.au
[2] Center for Ubiquitous Computing, University of Oulu, Finland
simo.hosio@oulu.fi

Abstract. Matching crowd workers to suitable tasks is highly desirable as it can enhance task performance, reduce the cost for requesters, and increase worker satisfaction. In this paper, we propose a method that considers workers' cognitive ability to predict their suitability for a wide range of crowdsourcing tasks. We measure cognitive ability via fast-paced online cognitive tests with a combined average duration of 6.2 minutes. We then demonstrate that our proposed method can effectively assign or recommend workers to five different popular crowd tasks: Classification, Counting, Proofreading, Sentiment Analysis, and Transcription. Using our approach we demonstrate a significant improvement in the expected overall task accuracy. While previous methods require access to worker history or demographics, our work offers a quick and accurate way to determine which workers are more suitable for which tasks.

Keywords: Crowdsourcing · Cognitive ability · Task performance

1 Introduction

Although crowdsourcing is actively used for a wide variety of both academic and industry tasks, ensuring that the crowd produces data of appropriate quality remains an important challenge. As a result, a wide range of quality assurance mechanisms have been proposed, from straightforward approaches, such as the use of golden standard questions [13] to more complex approaches like monitoring worker activity on crowdsourcing markets [54]. Researchers have also explored ways to predict which workers are likely to perform a task well and facilitate appropriate task assignment [60, 22]. For instance, this can be achieved through the analysis of historical records on completed tasks over a certain period [40, 46]. However, this method is only applicable when such records exist and can be matched to individual workers, which is often not the case. Furthermore, mechanisms that do not rely on the historical performance of workers are better suited in certain scenarios, such as one-time crowdsourcing tasks/campaigns or when

considering new workers of a platform. In these cases, there is no past performance data to predict how well workers would perform on similar or relevant tasks [22].

More robust approaches entail predicting worker performance using different worker attributes, such as age [34], location [34, 56], technical skills [43], and personality [33, 41]. In this paper, we investigate a promising but understudied worker attribute to predict performance in a crowdsourcing setting – cognitive ability. Cognitive ability tests are one of the many methods used by organisations during the recruitment process to identify potential employees with the highest job compatibility. Furthermore, Psychology research has extensively shown that a person's cognitive ability is a good indicator of work performance [55]. In particular, the literature presents three core executive functions of the brain (Inhibition Control, Working Memory, and Cognitive Flexibility) as the basis to describe cognitive ability, which can be measured using appropriate tests [9]. In a crowdsourcing setting, a recent study by Goncalves *et al.* [22] reported promising results regarding the successful prediction of crowd worker performance based on their cognitive skills. However, the completion of the cognitive ability tests (visual and verbal) and crowdsourcing tasks was conducted in a lab study with a limited sample of 24 participants instead of workers from a crowdsourcing platform. Further, the researchers used the Educational Testing Service (ETS) cognitive kit [16], a collection of comprehensive yet complex and time-consuming cognitive tests that are not practical for an online setting. Goncalves *et al.* [22] report that the experiment lasted between 90 to 120 minutes per participant, which would be considered overly long in most online crowdsourcing scenarios.

In this paper we aim to establish a link between the metrics of simple and established online cognitive tests and worker task performance. This link could be used in routing tasks to enhance the efficiency and outcomes of crowd work. As a result, task requesters and crowdsourcing platforms would be able to distinguish the optimum set of workers for a particular crowd task. We conducted an online study on Amazon Mechanical Turk (MTurk)[3] with 102 workers. We asked workers to complete a set of simple and quick (*i.e.,* workers spent on average 6.2 minutes to complete five tests) online cognitive tests (Stroop [42], Flanker [17], N-back [49], Task switching [47], Pointing[50]) that capture the three core executive functions of the brain. This was followed by the completion of typical tasks available in crowdsourcing platforms (Classification, Counting, Proofreading, Sentiment Analysis, Transcription). Our results show a strong relationship between the cognitive ability of crowd workers and their performance in crowdsourcing tasks. We also identify relationships between specific cognitive tests and crowd tasks based on executive functions. Finally, we assign workers to tasks based on their cognitive test scores and demonstrate that our method can significantly improve crowd task accuracy when compared to a baseline generic task assignment.

[3] https://www.mturk.com

2 Related Work

2.1 Human Cognitive Ability and Executive Functions

Human cognitive ability has been extensively studied in Psychology and is often described using executive functions [9]. Executive functions are known to be vital for mental and physical well-being, as well as success in school [3] and at work [2]. The general consensus is that there are three core executive functions: inhibition control, working memory, and cognitive flexibility. These functions form the basis of higher order functions such as reasoning, problem-solving, and planning [9]. *Inhibition control* is the conscious or unconscious restriction of a process or behaviour, especially of impulses or desires. *Working memory* is the ability to hold information in memory and mentally work with it. *Cognitive flexibility* (also known as Switching) is the ability to adapt behaviours in response to changes in the environment and is often associated with creativity [9].

A wide variety of psychological tests such as Stroop [42], Task Switching [47], and N-Back [49] have been developed to assess executive functions. A collection of such tasks is known as a cognitive kit (*e.g.,* Cambridge Neuropsychological Test Automated Battery (CANTAB) [51], Test My Brain [21], The Addenbrooke's Cognitive Examination [45]) and is extensively used in medical and psychological research [9]. Cognitive ability measured from such tests is known to be a good indicator of performance at work, among other predictors such as personality, emotional intelligence, and job experience [55]. This is also well supported by the Person-Job fit theorem which is broadly defined as the compatibility between individuals and jobs [37]. The two aspects of the theory are the suitability of a person for the requirements of a job, and the match between the expectations of a person and the attributes of the job [37]. In theory, any organisation would benefit from optimising their employee selection processes to achieve Person-Job fit, as the literature identifies several positive outcomes such as job performance, satisfaction, and motivation [14].

In a study involving software developers, Chilton *et al.* [4] reported that a misfit between cognitive style and that of the job environment could diminish performance while increasing strain. Similar links between cognitive style and work performance have been established in a number of studies [29, 57]. Although cognitive style or the way individuals think, perceive, and remember information slightly differ from cognitive ability, it correlates with cognitive ability [19]. We also note that several studies have shown that there is no significant relationship between cognitive style and performance at work [53, 38].

In this study we aim to investigate the impact of worker cognitive ability on their task performance in crowdsourcing platforms by measuring cognitive ability using online cognitive tests that capture the three widely established executive functions of the brain.

2.2 Measuring Cognitive Ability Online

Previous work has shown that accurately measuring cognitive ability through online tests is feasible. For instance, Germine *et al.* [21] explored the validity

of using the web for timed, performance-based, and/or stimulus-controlled experiments which are critical for measuring cognitive aptitude online. They reported that web samples do not differ significantly from traditionally recruited or lab-tested samples. Furthermore, participants of their study were anonymous, uncompensated, and unsupervised.

In another example, Crump *et al.* [6] examined the viability of conducting behavioural experiments on crowdsourcing platforms. In a study conducted on MTurk, workers completed tests that are used in cognitive science and cognitive psychology (*e.g.,* Stroop, Flanker, Attentional Blink) with the results being comparable to those collected in laboratory settings. These experiments lasted up to 30 minutes and have characteristics such as multi-trial designs, stimulus presentation, complex instructions, rapid response recording, and requirement of sustained attention of participants. Given these findings and the fact that we based our online cognitive tests on the extensive literature in Psychology on this topic, we anticipate that our online cognitive tests will effectively gauge the cognitive aptitude of crowd workers by testing the three executive functions of the brain.

2.3 Cognitive Ability of Crowdworkers

Eickhoff [15] examined the effect of cognitive biases in crowdsourced relevance labelling tasks and reported that biases could significantly deteriorate the quality of output. A cognitive bias is a systematic error in thinking that affects judgements and decisions. For instance, the framing effect is one such cognitive bias where people respond to a particular option in different ways based on how it is presented. Though cognitive biases differ from cognitive aptitudes, they are closely related and the literature suggests that people with higher cognitive abilities are better at avoiding cognitive biases when making decisions [59].

Alagarai *et al.* [1] investigated different cognitive elements of crowd task design and its effect on performance. They showed that higher task accuracy could be obtained by reducing the demand for visual search and working memory within the task. Previous work by Goncalves *et al.* [22] predicted the accuracy of participants when performing crowd tasks based on cognitive skills measured. However, this experiment was conducted in a laboratory setting, with a small sample, and using the ETS cognitive kit [16], which consists of laborious and time-consuming tests. We aim to investigate this further using straightforward and quick online cognitive tests with a larger sample and explore its applicability for task assignment in crowdsourcing.

2.4 Task Assignment Based on Worker Attributes

Previous work has shown that both demographic and behavioural attributes of workers impact their work quality [33, 34, 41]. In practice, apart from more common attributes such as approval rate, the number of tasks completed, and location, crowd platforms allow requesters to narrow down the worker selection

at a premium price. For example, MTurk allows requesters to select a subset of workers based on worker gender, age, daily internet usage, job, among others.

While there is a strong relationship between crowd worker accuracy and their location in relevance labelling [24, 34] and content analysis [56], studies have confirmed that gender has no significant effect on task accuracy in crowdsourcing [34]. Beyond demographics, personality of the worker is known to affect accuracy. In a study on labelling relevance, Kazai *et al.* [33] segmented crowd workers into five categories based on personality dimensions and reported a significant correlation between personality type and the mean accuracy of the worker. In a subsequent study, Kazai *et al.* [34] also reported that certain personality traits relate to higher task accuracy. Lykourentzou *et al.* [41] examined the effect of personality on the performance of collaborative crowd work on creative tasks and reported that balanced teams containing multiple personalities produce better work in terms of the quality of outcome.

Rzeszotarski and Kittur [54] showed that it is feasible to build predictive models of task performance based on behavioural traces of the user. They introduced a method that analyses the sequence of actions (*e.g.,* mouse movements, scrolling, key-strokes) performed by the user to complete a task, which can be used to measure task accuracy and content quality. Han *et al.* [28] explored annotating the semantic structure of the web using crowdsourcing and reported that most of the behavioural factors of the worker are correlated with the annotation quality. In addition, behaviours of trained professional workers have been successfully used as golden standard to identify those with poor performance [35]. However, behaviour based task performance prediction methods can only be used as post-processing techniques to exclude subpar contributions, which differ from task routing methods. Another approach is to extract the interests of users from social media activity and serve tasks accordingly [11]. We note practical and ethical difficulties in linking worker profiles with social media data.

3 Method

In this study we measured the cognitive ability of crowd workers using five cognitive tests. We then recorded worker performance in five crowdsourcing tasks, and examined if we can utilise cognitive aptitude as an indicator of crowd task performance. We used established cognitive tests to measure the three executive functions of the brain. Table 1 describes the primary executive function measured by each test.

3.1 Cognitive Tests

A description of each cognitive test is provided below.

Stroop *Test [42].* The classic Stroop test presents two types of trials (incongruent and congruent). As shown in Figure 1, incongruent trials present names of colours (such as "green") displayed in a different colour ("red") whereas congruent trials present names in matching colour. We also included a third trial

Table 1: Cognitive tests and associated executive functions [9]

Cognitive Test	Executive Function
Stroop	Inhibition Control
Flanker	Inhibition Control
Task Switching	Cognitive Flexibility
N-Back	Working Memory
Pointing	Working Memory

type (unrelated) where non-colour words (such as "monkey") appear in either red, green, or blue colour. Participants were asked to press the key corresponding with the first letter of the colour of the word. When asked to focus on the colour of the ink and ignore the meaning of the word (*i.e.,* suppress our prepotent response to words), people are found to be slower and less accurate. This is known as the Stroop effect. Our test contained a total of 18 trials, with a total of 6 trails per type.

Eriksen ***Flanker*** *Test [17].* In each trial crowd workers were presented with a sequence of five arrow symbols (*e.g.,* >>>>>, <<><<) and were asked to pick the centre symbol and press the corresponding arrow key. This task contained 8 congruent (all arrows pointing in the same direction) and 8 incongruent (centre symbol pointing to the opposite direction from the rest) trials. The task effect is similar to the Stroop test.

Task Switching *Test [47].* This test presented a letter and a number in each trial. Depending on whether the pair appears on the upper or lower half of the display, participants were asked to indicate whether the letter is a vowel or consonant, or whether the number is even or odd. The test contained 8 repeating and 8 switching trials.

N-Back *Test [49].* In the N-Back test, crowd workers were presented with a sequence of stimuli. For each stimulus, participants were asked to decide if the current stimulus is the same as the one presented N trials ago, where N can be 1, 2, or 3. We used the 3-back version of this test with each worker completing 16 trials.

Self-ordered ***Pointing*** *Test [50].* In this task, crowd workers were shown 3 to 12 randomly distributed identical squares and were asked to click one box at a time, in any order and without repetition, making sure to click all boxes.

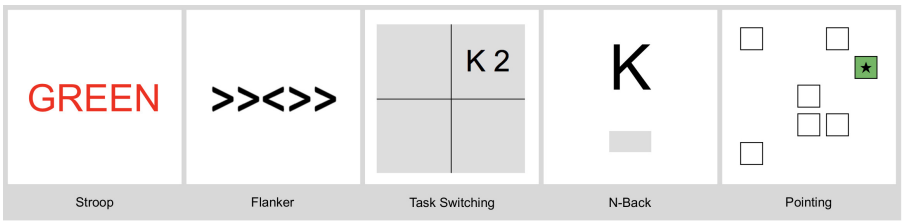

Fig. 1: Screenshots from cognitive tests

Workers received visual feedback after each choice. We tested workers' ability to remember which items they have clicked. The test contained 5 rounds with the total number of squares increasing in each round.

For each test, we specified instructions and included an example prior to the test to ensure workers fully understood the test. Except for the Pointing test, we also configured each trial within the tests to expire after 3.5 seconds. This allowed us to avoid crowd workers pausing the study in the middle of a test and get them to promptly complete each trial. For the Stroop, Flanker, Task Switching, and N-Back tests we recorded accuracy, response time, and trial type (if applicable) for each trial. Based on the trial type, for the Stroop, Flanker, Task Switching tests, test effect was calculated (*e.g.,* Stroop effect in terms of accuracy is the difference in accuracy between congruent and incongruent trials).

3.2 Crowdsourcing Tasks

We used crowdsourcing tasks that are representative of typical tasks available in popular crowdsourcing platforms. Crowd task taxonomy [20] and task availability [10] reported in the literature were also considered. The sentiment analysis and proofreading tasks were adopted from previous work by Goncalves *et al.* [22], and the counting task from Rogstadius *et al.* [52] and Goncalves *et al.* [23, 25]. The transcription and item classification tasks were created specifically for this study. Screenshots from the crowdsourcing tasks are shown in Figure 2 and a description of each task is given below. All tasks had varying complexity as shown in Figure 3 and were presented to participants in random order.

Sentiment Analysis. Crowd workers were asked to identify the sentiment of a sentence (*i.e.,* point of view, opinion). A sentence's sentiment was classified as either 'negative', 'neutral', or 'positive'. The task contained a total of 16 unique sentences. Half of the sentences were straightforward (*e.g.,* "The weather is great today"), while the other half were more challenging due to sentiment ambiguity, context, or sarcasm (*e.g.,*"I'm so pleased road construction woke me up with a bang").

Counting. In this task, workers were presented with an image of a petri dish and asked to count malaria-infected blood cells. Workers were provided with specific instructions on how to differentiate an infected blood cell from an ordinary blood cell. The task contained 8 images that were generated algorithmically con-

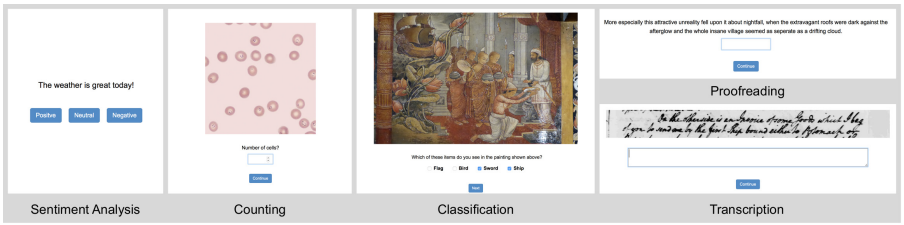

Fig. 2: Screenshots from crowdsourcing tasks

taining varying numbers of infected and ordinary blood cells. Accuracy for each image was determined by $max(0, 1 - \frac{|response - ground\_truth|}{ground\_truth})$.

Item ***Classification***. In this task, crowd workers were presented with 16 paintings (primarily from The Metropolitan Museum of Art[4] and the remaining from Flickr[5], all images licensed for public use) and were asked to identify and mark the items appearing in each painting from a given list of four items. Images represent different painting styles from different countries and contain one or more of the listed items. Certain items could be easily spotted, whereas others were more challenging (*e.g.,* the classification image shown in Figure 2 contains both objects 'Ship' and 'Sword', where the latter is more challenging to locate).

Proofreading. In this task, crowd workers were asked to proofread 12 sentences. Two sentences contained no errors. The remaining sentences contained a single error such as a misspelled word, a grammatical error, or an incorrect word. Workers were asked to type the correct word which should replace the identified erroneous word.

Transcription. Crowdworkers were required to type out a piece of text from a given image. We included 12 images extracted from The George Washington Papers at the Library of Congress [58] in the task. As shown in Figure 3, manuscripts had varying complexity based on the writing style, date, and content. We calculated Levenshtein distance (LD) [7] between the response string and the ground truth and measured accuracy using $max(0, 1 - \frac{2 \times LD}{length(ground\_truth)})$.

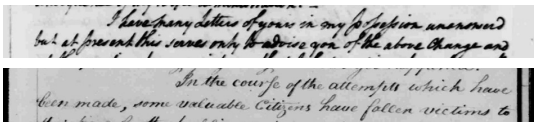

Fig. 3: Transcription tasks of high (top) and low (bottom) complexity.

The cognitive tests were implemented using *jsPsych*, a JavaScript library for online behavioural experiments [39]. Our experiment was integrated with MTurk using *psiTurk* [26], which let us host the experiment on our own server without the need of redirecting users and asking them to submit a completion code.

All tests were encapsulated to a single *Human Intelligent Task* (HIT) and posted to MTurk. When participants accepted the HIT, they were required to electronically sign an informed consent form to start the study. Workers first completed the five cognitive tests, followed by the five crowdsourcing tasks. Both the order of the tests and tasks was randomised. In the last step of the study, participants were requested to provide demographic information (age, gender, and education level). From a pilot study, we estimated that workers would spend around 40 minutes to complete the study. Based on the prevailing federal minimum wage of the United States of $7.25, we payed $5.00 (USD) for each worker

[4] https://www.metmuseum.org/art/collection
[5] https://www.flickr.com

who completed all the tests and tasks. The amount we payed for a worker is comfortably above the average pay one would receive for regular tasks in MTurk [10].

We considered the executive functions associated with each crowdsourcing task during task selection in order to be able to relate them to the different cognitive tests. For example, our counting and classification tasks require sustained attention (Inhibition Control), and demands Working Memory skills while going through the different elements [9]. For the Proofreading task, it is critical to relate to and apply different grammar rules and language patterns (Working Memory and Cognitive Flexibility) [5]. Initially, three of the paper's authors individually identified executive functions linked to each crowdsourcing task based on the literature and their own judgement. The authors then discussed the results, which led to the mapping shown in Table 2.

Table 2: Crowdsourcing tasks and related executive functions

Task	Executive Functions
Classification	Inhibition Control & Working Memory
Counting	Inhibition Control & Working Memory
Proofreading	Working Memory & Cognitive Flexibility
Sentiment Analysis	Cognitive Flexibility & Inhibition Control
Transcription	Cognitive Flexibility & Working Memory

4 Results

A total of 102 workers completed the study (Female 48, Male 54). On average, workers spent 43.6 minutes to complete the study, with 37.0 minutes spent on the crowdsourcing tasks ($SD = 10.7$) and 6.2 minutes on the cognitive tests ($SD = 2.1$). Based on a Pearson Correlation test, we found a significant correlation between the worker scores for the cognitive tests and the mean accuracy for the crowdsourcing tasks ($r = 0.47, p < 0.01$), as shown in Figure 4.

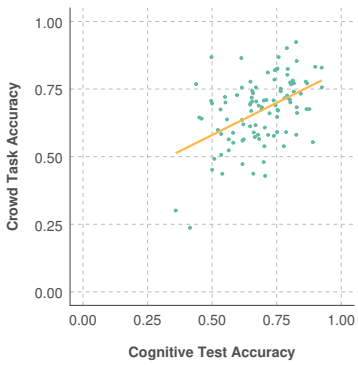

Fig. 4: Accuracy of crowdsourcing tasks vs accuracy of cognitive tests.

4.1 Cognitive Tests

Figure 5 shows worker performance across the five cognitive tests. Workers found the Stroop test to be relatively easier than the rest. In contrast, the mean accuracy of the N-back task is consistently low. Workers are slightly faster in responding to the two tests that measure inhibition control, Stroop and Flanker.

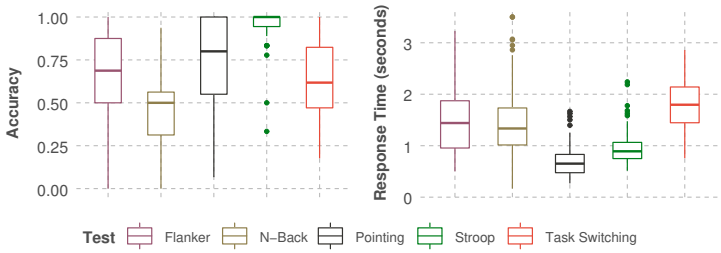

Fig. 5: Accuracy and response time for cognitive tests.

Figure 6 summarises the observed Stroop, Flanker, and Task Switching effects in terms of response time and error rate. As indicated by ANOVA results, for both Stroop and Flanker tests, workers were less error prone ($F(1, 202) = 26.88, p < 0.01$, $F(1, 202) = 8.80, p < 0.01$) and faster ($F(1, 202) = 16.16, p < 0.01$, $F(1, 202) = 5.22, p < 0.05$) when presented with congruent tasks. In the Task Switching test workers were generally faster ($F(1, 202) = 6.78, p < 0.01$) when the same type of task was repeated as opposed to switching from one type to another. This confirms that the effect of the tests was in the expected direction [42, 17, 47].

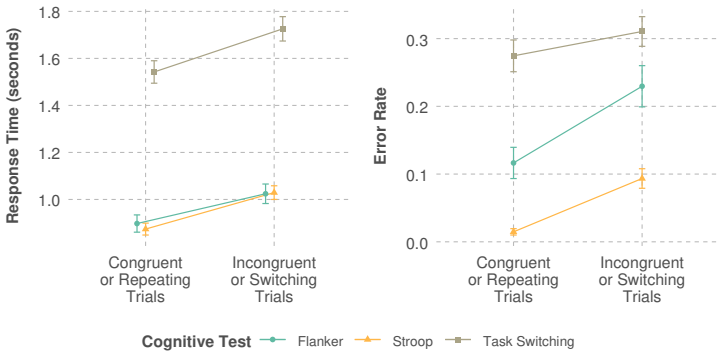

Fig. 6: Stroop, Flanker, and Task Switching effects.

4.2 Crowdsourcing Tasks

Figure 7 shows that workers were generally faster and more accurate in the Sentiment Analysis task as compared to other tasks. Worker accuracy was lowest for the Proofreading task. Figure 8 visualises the accuracy of workers for each sub task of the crowdsourcing tasks (*e.g.,* an individual sentence in the sentiment task). This demonstrates that there is a varying level of complexity within each of our crowdsourcing tasks, an aspect we aimed for in the initial study design. Finally, we do not observe a significant impact of gender, age, or education level on crowd task performance.

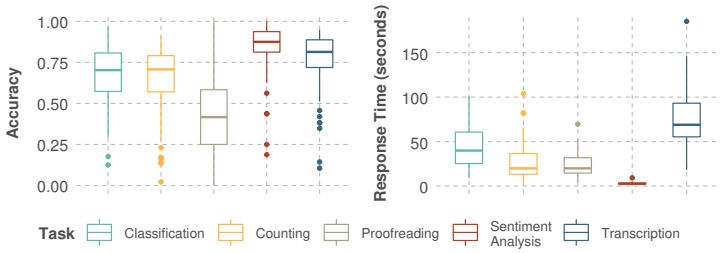

Fig. 7: Accuracy and response time for crowd tasks.

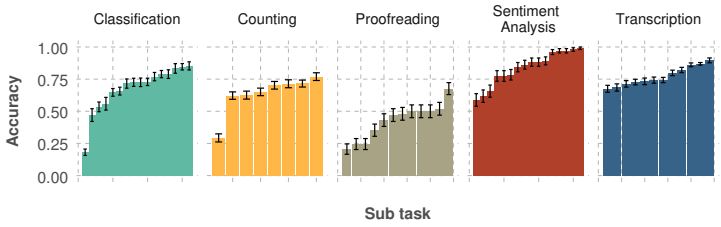

Fig. 8: Accuracy of sub tasks for each crowdsourcing task (Sub tasks are ordered in ascending order of mean accuracy).

4.3 Predicting Crowd Task Accuracy

We used the outcomes of the cognitive tests (*e.g.,* accuracy, response time, Stroop effect) as features to predict the overall accuracy of each worker. Other features include mean response time of instructions and demographic information (age, gender, and education level).

We used Generalised Linear Models, Random Forest, and Beta Regression to predict the overall task accuracy. Mean Absolute Error (MAE), Root Mean

Square Error (RMSE), and R-Squared values for the models with 5-fold cross validation with 10 repeats are shown in Table 3. Inter-correlations were checked prior to constructing the models and the variance inflation factors values of our predictors were below the often-used threshold of 5 to detect multicollinearity [27]. As Beta Regression is optimised for datasets where the output value is in the range (0,1), we had to slightly modify the accuracy values (y) using the equation, $(y * (n - 1) + 0.5)/n$ where n is the number of observations.

Table 3: Results of predictive models (5-fold cross validation with 10 repeats)

Method	MAE	RMSE	R^2
Generalised Linear Model	0.085	0.105	0.320
Random Forest	0.085	0.105	0.303
Beta Regression	0.083	0.105	0.290

We also predicted the accuracy for individual crowdsourcing tasks using the same procedure. Based on the results (MAE, RMSE, and R-Squared values), we selected Random Forest for further investigation and prediction as it produces slightly better results over the other two models in this analysis. Table 4 presents the features that were shown to be the most important based on feature importance scores of Random Forest models and the respective executive functions that those features relate to, as well as the executive functions we hypothesised each crowdsourcing task covers (Table 2).

Table 4: Significant features and related executive functions

Crowd Task	Hypothesis	Significant Features	Imp. Score	Related Executive Functions
Classification	In. Control W. Memory	Pointing (Accuracy) Flanker (Response Time) Stroop (Accuracy)	4.95 3.07 2.45	In. Control W. Memory
Counting	In. Control W. Memory	Flanker (Effect Accuracy) Pointing (Response Time) Stroop (Accuracy)	5.57 3.72 3.37	In. Control W. Memory
Proofreading	W. Memory Cog. Flexibility	Task Switching (Accuracy) Pointing (Accuracy) Instructions (Response Time)	7.93 5.60 4.08	W. Memory Cog. Flexibility
Sen. Analysis	Cog. Flexibility In. Control	Stroop (Response Time) Instructions (Response Time) Flanker (Effect Accuracy)	9.68 6.90 5.74	In. Control
Transcription	Cog. Flexibility W. Memory	Task Switching (Accuracy) Task Switching (Effect Accuracy)	3.03 2.98	Cog. Flexibility

In addition, we applied Principal Component Analysis (PCA) separately for both the cognitive test and crowdsourcing task results. PCA can be used to show the distance and relatedness among a population. We visualise this analysis in Figures 9 & 10. These figures, known as variable correlation plots, visualise the relationship between all variables. In Figure 9, we observe that the N-back and Pointing tests are grouped together, implying they are highly correlated. Both tests measure Working Memory. Similarly, Stroop and Flanker tests, which both measure Inhibition Control, are positively correlated as shown in Figure 9. More importantly, Figure 9 confirms that our cognitive test results are in agreement with the literature regarding the measured executive functions (as presented in Table 1). The yellow circle indicates a 100% representation of a variable in the given space. The length of the arrows (close to the edge of the circle) indicates that all variables are well represented in both plots.

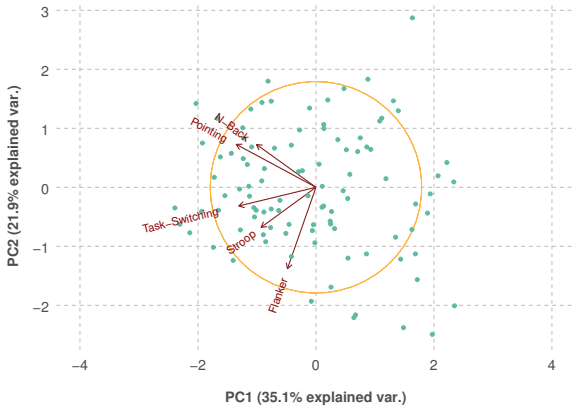

Fig. 9: Principal Component Analysis (PCA) of cognitive tests.

We make two important observations in Figure 10. First, workers are spread throughout the space, which shows the diversity in terms of worker expertise. For example, worker marked as 'W1' in Figure 10 did not perform well on Proofreading and Transcription tasks, but performs above average on Sentiment Analysis and Counting tasks. Our aim is to capture these differences via cognitive tests to facilitate effective task assignment. Second, we identify strong positive correlations among Proofreading and Transcription task pair, and Sentiment Analysis and Counting task pair. This suggests a similarity between tasks in terms of underlying executive functions. According to our findings (Table 4), Cognitive Flexibility is important for both Proofreading and Transcription tasks while Inhibition Control is significant for Sentiment Analysis and Counting tasks.

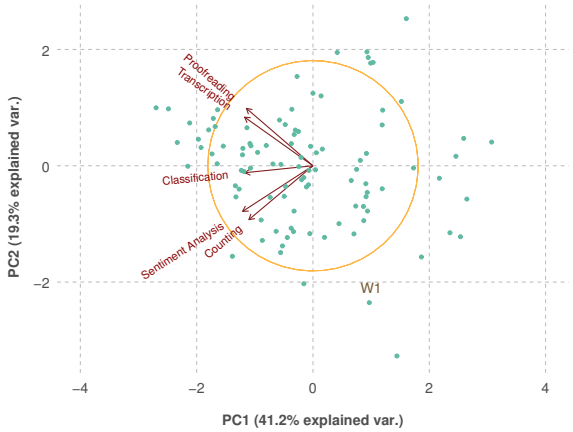

Fig. 10: Principal Component Analysis (PCA) of crowdsourcing tasks.

4.4 Task Assignment Based on Cognitive Skills

Next we developed a strategy to exemplify how cognitive tests can be used for task assignment. To evaluate our strategy, we first select workers for tasks solely based on cognitive test scores, and then compare their task performance as recorded in the study. Here, we transform our prediction from a regression problem to a binary classification problem and focus on predicting if a particular worker should be assigned to a particular crowdsourcing task or not.

For any specific task, we can select a subset of workers from a worker pool in order to maximise the predicted accuracy. For each task, we trained a Random Forest model with 5-fold cross validation using measures from cognitive tests and demographic information as features. Using the models, we predicted the expected accuracy for each worker for each task. Then for each task, based on predicted worker accuracy, we categorised workers into two classes ('Selected' or 'Not Selected'). We used a variable 'Worker Qualification Limit' (L) to determine which portion of workers to consider for assignment. For instance when $L = 40$ for the Classification task, the top 40% of workers in terms of their predicted accuracy in this task are labelled as 'Selected' and the remaining 60% are labelled as 'Not Selected'.

The observed accuracy for workers based on prediction outputs for all five tasks with different L values is shown in Figure 11. For instance, for the Sentiment Analysis task, if we select the top 51 workers out of 102 ($L = 50$) in terms of our predictive model, we observe that those 51 selected workers actually achieve a mean accuracy of 0.88 whereas the 51 unselected workers achieve a mean accuracy of 0.80. The overall mean accuracy for the Sentiment Analysis task for all 102 workers is 0.84 (shown in black horizontal line in Figure 11). Also, we note that for any L value, our assignment method selects a subset of workers whose mean accuracy for the task is better than the mean accuracy of the remaining workers or the mean accuracy of the entire worker pool.

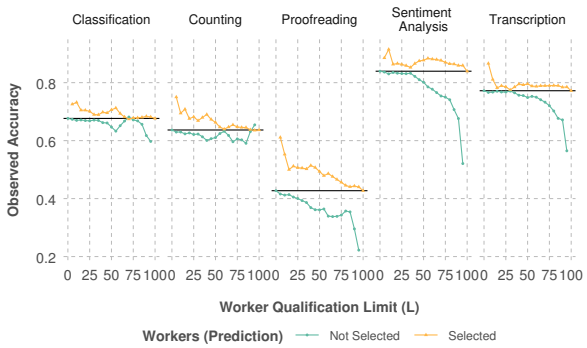

Fig. 11: Accuracy of workers for each task based on output of prediction.

Next we investigated to what extent our method leads to worker discrimination. In other words, does it always favour a handful of skillful workers? We calculate the total number of tasks each worker would be assigned to once we select workers for all five tasks based on our approach. Figure 12 summarises the outcome distribution. If task assignment is carried out based on our model with L as 50, we observer that 11 (10.8%) workers are selected for all five tasks, and 18 (17.6%) workers are not assigned any task. A higher L value (*e.g.,* $L = 75$) assigns more workers to all five tasks. For lower values (*e.g.,* $L = 25$), which represent a more "exclusive" model, we observe that no worker is assigned to all 5 tasks. In other words, at low L values, task routing is so exclusive that there is no single worker in our sample that would meet the expectations for all 5 tasks.

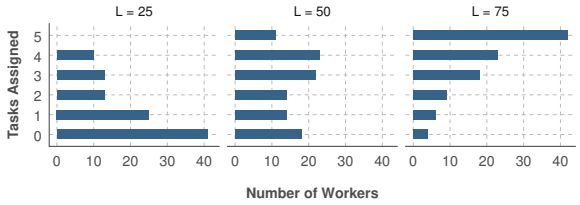

Fig. 12: Number of workers against the total number of tasks assigned to each worker.

5 Discussion

5.1 Using Cognitive Tests to Predict Performance

Apart from cognitive skills, previous work has explored the relationship between crowd task performance and a number of worker attributes, such as age [34], location [34, 56], skills [43], and personality [33, 41]. However, we note a common pitfall in these studies: evaluation is based on a single type of task. For example, Kazai *et al.* [33, 34] only used relevance labelling tasks; Shaw *et al.* [56] used

content analysis questions; Mavridis *et al.* [43] used a set of multiple choice questions on the topic of 'Computer Science'; and Lykourentzou *et al.* [41] explored collaborative advertisement creation. To ensure the applicability of our findings to generic crowd work, our study included five different crowdsourcing tasks.

Previous work by Goncalves *et al.* [22] demonstrated that it is possible to predict the accuracy of crowd workers based on their cognitive skills. While their study used 8 different crowdsourcing tasks to validate their findings, we note three major deficiencies. First, aptitude tests (visual and verbal) as well as the crowdsourcing tasks were conducted in a lab study, using a limited sample of 24 participants that are not representative of the crowd worker population. In contrast, we deployed our entire study on MTurk where 102 actual crowd workers completed the study. Second, compared to the ETS cognitive kit [16] used by Goncalves *et al.* [22], the tests we used to assess the cognitive ability of participants contained fewer trials which were mostly fast-paced. According to the specifications of ETS kit, it takes 44 minutes in total to complete the first part of each cognitive test employed in [22]. In contrast, workers spent on average 6.2 minutes to complete all five of our tests which indicates a significant reduction in required time. Third, unlike ETS tests which are not practical for an online setting (*e.g.,* one task requires paper folding), our online tests can be readily utilised by crowd platforms or task requesters with low effort. Thus, we eliminate any uncertainty associated with the previous study and establish that it is viable to use online cognitive tests to predict crowd task performance.

Furthermore, our prediction model can also be used along with other task routing frameworks. For example, Zheng *et al.* [60] proposed a task assignment system that uses expectation maximisation to populate an *estimated distribution matrix* containing estimated task accuracies. They select optimum tasks to be assigned to a worker based on this matrix. One could easily apply our model based on cognitive skills to predict task accuracy and then generate the estimated distribution matrix.

5.2 Conducting Cognitive Tests Online

We observed Stroop, Flanker, and Task Switching effects that replicate the results of classic Psychology experiments [42, 17, 47]. More importantly, our findings are in line with previous work by Crump *et al.* [6], that demonstrated that these effects could be effectively observed in online experiments. However, the effects we observed indicate a smaller effect size when compared to previous work. One reason for this could be the fact that we used a lower number of trials. For instance, we used 18 trials per worker with 102 workers for Stroop task, whereas the previous work by Crump *et al.* [6] is based on a total of 40 workers, each completing 96 trials.

Furthermore, our study identifies a strong relationship between each crowd task and several cognitive tests, validating our assumption that corresponding executive functions have an extensive impact on the crowd task (see Table 4). Based on this finding, a task requester could either select cognitive tests based on our results or pick executive functions that best explain the nature of the

work and choose matching tests. Alternatively, the requester could implement multiple tests covering all executive functions and then figure out which tests to be used by piloting with a small set of workers. From a crowdsourcing platform perspective, it is more viable to implement a collection of cognitive tests similar to the tests applied in our study, so that the outcomes of such tests can be used to route or recommend a wide variety of tasks to workers.

5.3 Task Assignment

Assigning tasks based on historical performance of workers in crowdsourcing platforms may be impractical for many reasons including anonymity, fluctuations in worker availability [31], or the lack of ground truth data to assess the historical accuracy of workers. On the other hand, using post-processing techniques to reject work could have consequences like workers avoiding the requester in future [44]. Here, we attempt to address these issues by using cognitive tests as predictors of crowd task performance. Our approach for assigning or recommending users, whereby we select a subset of workers who would possibly perform better at each task, can also be seen as a top-N recommendation task [8]. As shown in Figure 12 (for $L = 50$), we observe that tasks are well-distributed amongst workers – despite selecting the best workers for each task. Only 18 workers out of 102 end up not assigned to any task, while 11 workers are selected for all five tasks. This indicates that our proposed model is able to capture different expertise of workers and assign tasks accordingly. Fair task distribution is extremely important when we consider the task assignment problem from the perspective of the crowd workers. In contrast to widely used methods such as approval rate [32], our method does not aim to reward a superior set of workers who are capable in all tasks. Instead, our method focuses on finding the best suited task or tasks for each worker. This will allow workers to complete tasks that are more compatible with their skill set, which has been shown to improve worker satisfaction and reduce the likelihood of task abandonment [36, 30].

Due to budget constraints, crowd task requesters often have to either limit the number of answers expected for each question or reduce the payment for each answer. Both of these actions can reduce output quality [10]. We show that it is possible to obtain higher accuracy by selecting a subset of workers based on cognitive skills (Figure 11), therefore reducing the total number of answers and task cost. In a situation where the requester opts to use cognitive tests as a qualification test, an additional cost would incur for running the cognitive tests. However, typically the number of questions in each task is large enough [31, 10] to recover this initial investment.

5.4 Limitations

We acknowledge several limitations in our study. First, as we wanted to ensure that the cognitive tests took as little time as possible to complete, the total number of trials for each test was kept to a minimum. While measures of cognitive tests would become more accurate and distinct when increasing the number of

trials, the limited number of trials was sufficient for our predictions. Second, human cognitive ability is known to demonstrate subtle variations during the day [12], this is an aspect that we do not account for in our study. Third, similar to any supervised learning method, in the initial stages, our model needs to be trained using data captured from a set of workers performing cognitive tests followed by a set of crowdsourcing tasks similar to those presented in this study.

6 Conclusion and Future Work

In this paper we demonstrate the possibility of using brief online cognitive tests to predict the performance of crowd workers across a range of tasks. We present a study conducted on Amazon Mechanical Turk with 102 workers, where each worker completed a set of cognitive tests followed by a series of crowdsourcing tasks. Through our analysis we highlight the relationships between particular cognitive tests that measure one or more specific executive functions and crowdsourcing task performance.

We show that our proposed method can effectively assign or recommend workers to 5 distinct crowd tasks from a pool of 102 workers with significant improvements to task accuracy while also utilising the majority of the worker pool. Our results also suggest that suitability of a worker for a specific crowdsourcing task could be predicted using the outcome of two or three cognitive tests. Given that each of our cognitive tests could be completed within less than 2 minutes and can be seamlessly integrated with online crowdsourcing platforms, our findings could be readily adopted by researchers, general task requesters, and crowdsourcing platforms.

Further research on the longitudinal impact of the process of measuring cognitive ability would allow us to decide on the optimum frequency with which these tests should be repeated. Cognitive tests should not be repeated too often as it could lead to workers being familiarised with tests. It is known that training obtained in cognitive tests could contribute towards an improvement in metrics of those particular tests but has no impact on other tests or general performance of other tasks [48]. As there are a number of different tests that measure the same executive function [9], one alternative would be to randomly select tests from a pool of tests instead of using identical dedicated tests. In addition, a future study that dynamically routes tasks based on worker cognitive ability and compares the results with other routing methods can further establish the effectiveness of the proposed method in practice.

In our evaluation, we consider assigning workers to tasks one after the other, which will result in repeatedly selecting some workers for multiple tasks. In future work, we intend to explore how we could assign or recommend tasks to workers based on cognitive skills when we have multiple tasks at hand. For this we could either adopt task routing frameworks presented in the literature [60, 40, 18] or propose a novel approach considering additional parameters such as the number of unique questions in each task, the number of answers required for each question, and payment.

References

1. Alagarai Sampath, H., Rajeshuni, R., Indurkhya, B.: Cognitively inspired task design to improve user performance on crowdsourcing platforms. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 3665–3674. CHI '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2556288.2557155
2. Bailey, C.E.: Cognitive accuracy and intelligent executive function in the brain and in business. Annals of the New York Academy of Sciences **1118**, 122–141 (2007). https://doi.org/10.1196/annals.1412.011
3. Borella, E., Carretti, B., Pelegrina, S.: The specific role of inhibition in reading comprehension in good and poor comprehenders. Journal of Learning Disabilities **43**(6), 541–552 (2010). https://doi.org/10.1177/0022219410371676
4. Chilton, M.A., Hardgrave, B.C., Armstrong, D.J.: Person-job cognitive style fit for software developers: The effect on strain and performance. Journal of Management Information Systems **22**(2), 193–226 (2005). https://doi.org/10.1080/07421222.2005.11045849
5. Clair-Thompson, H.L.S., Gathercole, S.E.: Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. The Quarterly Journal of Experimental Psychology **59**(4), 745–759 (2006). https://doi.org/10.1080/17470210500162854
6. Crump, M.J.C., McDonnell, J.V., Gureckis, T.M.: Evaluating amazon's mechanical turk as a tool for experimental behavioral research. PLoS ONE **8**(3), 1–18 (2013). https://doi.org/10.1371/journal.pone.0057410
7. Damerau, F.J.: A technique for computer detection and correction of spelling errors. Communications of the ACM **7**(3), 171–176 (1964)
8. Deshpande, M., Karypis, G.: Item-based top-n recommendation algorithms. ACM Trans. Inf. Syst. **22**(1), 143–177 (Jan 2004). https://doi.org/10.1145/963770.963776
9. Diamond, A.: Executive functions. Annual Review of Psychology **64**(1), 135–168 (2013). https://doi.org/10.1146/annurev-psych-113011-143750
10. Difallah, D.E., Catasta, M., Demartini, G., Ipeirotis, P.G., Cudré-Mauroux, P.: The dynamics of micro-task crowdsourcing: The case of amazon mturk. In: Proceedings of the 24th International Conference on World Wide Web. pp. 238–247. WWW '15, IW3C2, Switzerland (2015). https://doi.org/10.1145/2736277.2741685
11. Difallah, D.E., Demartini, G., Cudré-Mauroux, P.: Pick-a-crowd: Tell me what you like, and i'll tell you what to do. In: Proceedings of the 22nd International Conference on World Wide Web. pp. 367–374. WWW '13, ACM, New York, NY, USA (2013). https://doi.org/10.1145/2488388.2488421
12. Dingler, T., Schmidt, A., Machulla, T.: Building cognition-aware systems: A mobile toolkit for extracting time-of-day fluctuations of cognitive performance. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **1**(3) (Sep 2017). https://doi.org/10.1145/3132025
13. Downs, J.S., Holbrook, M.B., Sheng, S., Cranor, L.F.: Are your participants gaming the system?: Screening mechanical turk workers. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 2399–2402. CHI '10, ACM, New York, NY, USA (2010). https://doi.org/10.1145/1753326.1753688
14. Edwards, J.R.: Person-job fit: A conceptual integration, literature review, and methodological critique. John Wiley & Sons, England (1991)

15. Eickhoff, C.: Cognitive biases in crowdsourcing. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. pp. 162–170. WSDM '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3159652.3159654
16. Ekstrom, R.B., Dermen, D., Harman, H.H.: Manual for kit of factor-referenced cognitive tests, vol. 102. Educational Testing Service, Princeton, NJ, USA (1976)
17. Eriksen, B.A., Eriksen, C.W.: Effects of noise letters upon the identification of a target letter in a nonsearch task. Perception & Psychophysics **16**(1), 143–149 (1974)
18. Fan, J., Li, G., Ooi, B.C., Tan, K.l., Feng, J.: icrowd: An adaptive crowdsourcing framework. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. pp. 1015–1030. SIGMOD '15, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2723372.2750550
19. Federico, P.A., Landis, D.B.: Cognitive styles, abilities, and aptitudes: Are they dependent or independent? Contemporary Educational Psychology **9**(2), 146–161 (1984). https://doi.org/10.1016/0361-476X(84)90016-X
20. Gadiraju, U., Kawase, R., Dietze, S.: A taxonomy of microtasks on the web. In: Proceedings of the 25th ACM Conference on Hypertext and Social Media. pp. 218–223. HT '14, ACM, New York, NY, USA (2014)
21. Germine, L., Nakayama, K., Duchaine, B.C., Chabris, C.F., Chatterjee, G., Wilmer, J.B.: Is the web as good as the lab? comparable performance from web and lab in cognitive/perceptual experiments. Psychonomic Bulletin & Review **19**(5), 847–857 (2012). https://doi.org/10.3758/s13423-012-0296-9
22. Goncalves, J., Feldman, M., Hu, S., Kostakos, V., Bernstein, A.: Task routing and assignment in crowdsourcing based on cognitive abilities. In: Proceedings of the 26th International Conference on World Wide Web. pp. 1023–1031. WWW '17, IW3C2, Switzerland (2017). https://doi.org/10.1145/3041021.3055128
23. Goncalves, J., Ferreira, D., Hosio, S., Liu, Y., Rogstadius, J., Kukka, H., Kostakos, V.: Crowdsourcing on the spot: Altruistic use of public displays, feasibility, performance, and behaviours. In: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing. pp. 753–762. UbiComp '13 (2013). https://doi.org/10.1145/2493432.2493481
24. Goncalves, J., Hosio, S., van Berkel, N., Ahmed, F., Kostakos, V.: Crowdpickup: Crowdsourcing task pickup in the wild. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **1**(3), 51:1–51:22 (Sep 2017). https://doi.org/10.1145/3130916
25. Goncalves, J., Hosio, S., Rogstadius, J., Karapanos, E., Kostakos, V.: Motivating participation and improving quality of contribution in ubiquitous crowdsourcing. Comput. Netw. **90**(C), 34–48 (Oct 2015). https://doi.org/10.1016/j.comnet.2015.07.002
26. Gureckis, T.M., Martin, J., McDonnell, J., Rich, A.S., Markant, D., Coenen, A., Halpern, D., Hamrick, J.B., Chan, P.: psiturk: An open-source framework for conducting replicable behavioral experiments online. Behavior Research Methods **48**(3), 829–842 (Sep 2016). https://doi.org/10.3758/s13428-015-0642-8
27. Hair, J., Black, W., Babin, B., Anderson, R.: Multivariate Data Analysis. Prentice-Hall (2010)
28. Han, S., Dai, P., Paritosh, P., Huynh, D.: Crowdsourcing human annotation on web page structure: Infrastructure design and behavior-based quality control. ACM Trans. Intell. Syst. Technol. **7**(4), 56:1–56:25 (Apr 2016)
29. Hoffman, B.J., Woehr, D.J.: A quantitative review of the relationship between person–organization fit and behavioral outcomes. Journal of Vocational Behavior **68**(3), 389–399 (2006). https://doi.org/10.1016/j.jvb.2005.08.003

30. Hosio, S., Goncalves, J., Lehdonvirta, V., Ferreira, D., Kostakos, V.: Situated crowdsourcing using a market model. In: Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology. pp. 55–64. UIST '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2642918.2647362
31. Jain, A., Sarma, A.D., Parameswaran, A., Widom, J.: Understanding workers, developing effective tasks, and enhancing marketplace dynamics: A study of a large crowdsourcing marketplace. Proc. VLDB Endow. **10**(7), 829–840 (2017). https://doi.org/10.14778/3067421.3067431
32. Kazai, G.: In search of quality in crowdsourcing for search engine evaluation. In: Advances in Information Retrieval. pp. 165–176. Springer, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20161-5_17
33. Kazai, G., Kamps, J., Milic-Frayling, N.: Worker types and personality traits in crowdsourcing relevance labels. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. pp. 1941–1944. CIKM '11, ACM, New York, NY, USA (2011). https://doi.org/10.1145/2063576.2063860
34. Kazai, G., Kamps, J., Milic-Frayling, N.: The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. pp. 2583–2586. CIKM '12, ACM, New York, NY, USA (2012). https://doi.org/10.1145/2396761.2398697
35. Kazai, G., Zitouni, I.: Quality management in crowdsourcing using gold judges behavior. In: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. pp. 267–276. WSDM '16, ACM, New York, NY, USA (2016). https://doi.org/10.1145/2835776.2835835
36. Kittur, A., Nickerson, J.V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., Horton, J.: The future of crowd work. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work. pp. 1301–1318. CSCW '13, ACM, New York, NY, USA (2013). https://doi.org/10.1145/2441776.2441923
37. Kristof, A.L.: Person-organization fit: an integrative review of its conceptualizations, measurement, and implications. Personnel Psychology **49**(1), 1–49 (1996)
38. Kristof-Brown, A.L., Zimmerman, R.D., Johnson, E.C.: Consequences of individuals' fit at work: A meta-analysis of person-job, person-organization, person-group, and person-supervisor fit. Personnel Psychology **58**(2), 281–342 (2005)
39. de Leeuw, J.R.: jspsych: A javascript library for creating behavioral experiments in a web browser. Behavior Research Methods **47**(1), 1–12 (2015)
40. Liu, X., Lu, M., Ooi, B.C., Shen, Y., Wu, S., Zhang, M.: Cdas: A crowdsourcing data analytics system. Proc. VLDB Endow. **5**(10), 1040–1051 (Jun 2012)
41. Lykourentzou, I., Antoniou, A., Naudet, Y., Dow, S.P.: Personality matters: Balancing for personality types leads to better outcomes for crowd teams. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. pp. 260–273. CSCW '16, ACM, New York, NY, USA (2016)
42. MacLeod, C.M.: Half a century of research on the stroop effect: An integrative review. Psychological Bulletin **109**(2), 163 (1991)
43. Mavridis, P., Gross-Amblard, D., Miklós, Z.: Using hierarchical skills for optimized task assignment in knowledge-intensive crowdsourcing. In: Proceedings of the 25th International Conference on World Wide Web. pp. 843–853. WWW '16, IW3C2, Switzerland (2016). https://doi.org/10.1145/2872427.2883070
44. McInnis, B., Cosley, D., Nam, C., Leshed, G.: Taking a hit: Designing around rejection, mistrust, risk, and workers' experiences in amazon mechanical turk. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. pp. 2271–2282. CHI '16, ACM, New York, NY, USA (2016)

45. Mioshi, E., Dawson, K., Mitchell, J., Arnold, R., Hodges, J.R.: The addenbrooke's cognitive examination revised (ace-r): a brief cognitive test battery for dementia screening. International Journal of Geriatric Psychiatry **21**(11), 1078–1085 (2006)
46. Mo, K., Zhong, E., Yang, Q.: Cross-task crowdsourcing. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 677–685. KDD '13, ACM, New York, NY, USA (2013)
47. Monsell, S.: Task switching. Trends in Cognitive Sciences **7**(3), 134–140 (2003)
48. Owen, A.M., Hampshire, A., Grahn, J.A., Stenton, R., Dajani, S., Burns, A.S., Howard, R.J., Ballard, C.G.: Putting brain training to the test. Nature **465**, 775 (2010). https://doi.org/10.0.4.14/nature09042
49. Owen, A.M., McMillan, K.M., Laird, A.R., Bullmore, E.: N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. Human Brain Mapping **25**(1), 46–59 (2005). https://doi.org/10.1002/hbm.20131
50. Petrides, M., Alivisatos, B., Evans, A.C., Meyer, E.: Dissociation of human mid-dorsolateral from posterior dorsolateral frontal cortex in memory processing. Proceedings of the National Academy of Sciences **90**(3), 873–877 (1993)
51. Robbins, T.W., James, M., Owen, A.M., Sahakian, B.J., McInnes, L., Rabbitt, P.: Cambridge Neuropsychological Test Automated Battery (CANTAB): A Factor Analytic Study of a Large Sample of Normal Elderly Volunteers. Dementia and Geriatric Cognitive Disorders **5**(5), 266–281 (1994). https://doi.org/10.1159/000106735
52. Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J., Vukovic, M.: An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In: Proceedings of the Fifth International AAAI Conference on Web and Social Media. ICWSM, vol. 11, pp. 17–21. AAAI, California, USA (2011)
53. Ruble, T.L., Cosier, R.A.: Effects of cognitive styles and decision setting on performance. Organizational Behavior and Human Decision Processes **46**(2), 283–295 (1990). https://doi.org/10.1016/0749-5978(90)90033-6
54. Rzeszotarski, J.M., Kittur, A.: Instrumenting the crowd: Using implicit behavioral measures to predict task performance. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology. pp. 13–22. UIST '11, ACM, New York, NY, USA (2011). https://doi.org/10.1145/2047196.2047199
55. Schmidt, F.L., Hunter, J.: General mental ability in the world of work: occupational attainment and job performance. Journal of personality and social psychology **86**(1), 162 (2004)
56. Shaw, A.D., Horton, J.J., Chen, D.L.: Designing incentives for inexpert human raters. In: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work. pp. 275–284. CSCW '11, ACM, New York, NY, USA (2011)
57. Verquer, M.L., Beehr, T.A., Wagner, S.H.: A meta-analysis of relations between person–organization fit and work attitudes. Journal of vocational behavior **63**(3), 473–489 (2003). https://doi.org/10.1016/S0001-8791(02)00036-2
58. Washington, G.: George washington papers, series 5, financial papers: Copybook of invoices and letters, 1754-1766 (1766), https://www.loc.gov/item/mgw500003
59. West, R.F., Toplak, M.E., Stanovich, K.E.: Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. Journal of Educational Psychology **100**(4), 930 (2008)
60. Zheng, Y., Wang, J., Li, G., Cheng, R., Feng, J.: Qasca: A quality-aware task assignment system for crowdsourcing applications. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. pp. 1031–1046. SIGMOD '15, ACM, New York, NY, USA (2015)

Chapter 5

Dynamic Task Assignment and Recommendation using Cognitive Abilities

5.1 Introduction

Stemming from our work [72] presented in Chapter 4, we built 'CrowdCog', a cognitive skill based online dynamic task assignment and recommendation system. This chapter presents our proposed method and a crowdsourcing study that evaluates the performance of CrowdCog in a real-time environment. Particularly, we detail how 'CrowdCog' issues cognitive tests and crowdsourcing tasks when workers request crowdsourcing tasks in a platform.

We deployed CrowdCog on Amazon Mechanical Turk, with five cognitive tests and four crowdsourcing tasks. Through this deployment, where worker-task engagement was highly similar to a standard crowdsourcing workflow, we aimed to ensure the ecological validity of using cognitive tests for crowdsourcing task assignment. Our results indicate that both task assignment and recommendation can significantly improve the data quality when compared to a baseline condition where workers select the tasks. We also compared our task assignment method with prior work and show that the task accuracy is on-par with a history-based method, and a state-of-the-art question assignment method that selects individual questions based on current responses [175]. Our method also has many practical advantages over other methods as we do not need to evaluate current worker answers in real-time or access historical work records of individual workers.

Finally, we discuss key considerations for implementing CrowdCog in a crowdsourcing platform. We elaborate possible ways to extend the assignment model to new types of tasks, repeatedly test users using a pool of cognitive tests, and adjust the threshold values in CrowdCog to obtain more tailored outcomes. We also discuss why task recommendation can be a valuable alternative to assignment in certain scenarios. More details on the CrowdCog system and the crowdsourcing study can be found in the attached publication, Article II.

5.2 Article II

Hettiachchi, D., van Berkel, N., Kostakos, V., Goncalves, J. (2020). CrowdCog: A Cognitive Skill based System for Heterogeneous Task Assignment and Recommendation in Crowdsourcing. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2), 1–22. https://doi.org/10.1145/3415181

Ethics ID: 1853314, The University of Melbourne Human Ethics Advisory Group.

CrowdCog: A Cognitive Skill based System for Heterogeneous Task Assignment and Recommendation in Crowdsourcing

DANULA HETTIACHCHI, The University of Melbourne, Australia
NIELS VAN BERKEL, Aalborg University, Denmark
VASSILIS KOSTAKOS, The University of Melbourne, Australia
JORGE GONCALVES, The University of Melbourne, Australia

While crowd workers typically complete a variety of tasks in crowdsourcing platforms, there is no widely accepted method to successfully match workers to different types of tasks. Researchers have considered using worker demographics, behavioural traces, and prior task completion records to optimise task assignment. However, optimum task assignment remains a challenging research problem due to limitations of proposed approaches, which in turn can have a significant impact on the future of crowdsourcing. We present 'CrowdCog', an online dynamic system that performs both task assignment and task recommendations, by relying on fast-paced online cognitive tests to estimate worker performance across a variety of tasks. Our work extends prior work that highlights the effect of workers' cognitive ability on crowdsourcing task performance. Our study, deployed on Amazon Mechanical Turk, involved 574 workers and 983 HITs that span across four typical crowd tasks (Classification, Counting, Transcription, and Sentiment Analysis). Our results show that both our assignment method and recommendation method result in a significant performance increase (5% to 20%) as compared to a generic or random task assignment. Our findings pave the way for the use of quick cognitive tests to provide robust recommendations and assignments to crowd workers.

CCS Concepts: • **Human-centered computing** → *Computer supported cooperative work*; • **Information systems** → **Crowdsourcing**.

Additional Key Words and Phrases: crowdsourcing, dynamic task assignment, cognitive abilities

ACM Reference Format:
Danula Hettiachchi, Niels van Berkel, Vassilis Kostakos, and Jorge Goncalves. 2020. CrowdCog: A Cognitive Skill based System for Heterogeneous Task Assignment and Recommendation in Crowdsourcing. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 110 (October 2020), 22 pages. https://doi.org/10.1145/3415181

1 INTRODUCTION

The availability of an extensive pool of workers willing and able to complete a high number of tasks has led to crowdsourcing platforms being widely used for data collection efforts by researchers from many different scientific disciplines (*e.g.,* Psychology, Astronomy, Computer Science, Medicine) and by organisations. With the increased use of crowdsourced data in critical applications, researchers have extensively explored approaches to improve the quality of gathered data [12]. While basic approaches such as gold standard questions and qualification tests [18] are commonly used, they

Authors' addresses: Danula Hettiachchi, danula.hettiachchi@unimelb.edu.au, The University of Melbourne, Melbourne, VIC, Australia; Niels van Berkel, nielsvanberkel@cs.aau.dk, Aalborg University, Aalborg, Denmark; Vassilis Kostakos, vassilis.kostakos@unimelb.edu.au, The University of Melbourne, Melbourne, VIC, Australia; Jorge Goncalves, jorge.goncalves@unimelb.edu.au, The University of Melbourne, Melbourne, VIC, Australia.

2573-0142/2020/10-ART110 $15.00
https://doi.org/10.1145/3415181

have inherent limitations which hinder their applicability. For instance, both of the aforementioned methods are task specific and a requester needs to curate the questions or the tests for each task. In addition, crowd tasks often lack ground truth information which makes the creation of gold standard questions challenging. More robust approaches include observing the historic or recent performance of the worker and subsequently estimating the worker's task performance prior to assigning the task [16, 24, 44, 60, 67]. However, in practice, these approaches are ineffective when there is limited or no prior task completion records available. This is particularly problematic when considering the influx of new crowd workers on a platform or one-time crowdsourcing tasks/campaigns. Hence, in our work, we seek to develop a task assignment method which is not based on a worker's prior records and which can be applied across a variety of crowdsourcing task types.

As a single crowdsourcing task is often organised as a collection of sub-tasks or questions of the same task (*e.g.,* translating 50 sentences), task assignment can be further dismantled into two steps: initial task assignment and subsequent question assignment. While task assignment aims to match workers with different types of tasks, question assignment focuses on selecting questions for the worker. The literature shows that crowdsourcing tasks vastly differ in terms of their complexity, required skills, expected time commitment, and allocated payment [15, 26]. Thus, task selection becomes increasingly relevant as the number of tasks and workers available on a crowdsourcing marketplace increases. On the other hand, prior work shows workers often struggle to find compatible or desirable tasks on marketplaces [9]. To match workers with suitable tasks, we investigate the task assignment problem for heterogeneous tasks and use the cognitive skills of crowd workers to predict task performance. Apart from cognitive skills, researchers have also studied the effect of many different worker characteristics like location [43, 63], age [43], personality [42, 49], mood [68] and technical skills [51] on crowd task performance. However, using a worker's cognitive ability for task assignment has a number of benefits over these approaches, such as being straightforward to measure [28], difficult to fabricate [14], and applicable to many task types [29, 36].

Goncalves et al. [29] first showed that cognitive ability can be a good predictor of crowd tasks performance. In addition, recent work by Hettiachchi et al. [36] on Amazon Mechanical Turk (MTurk) shows that online cognitive test performance is correlated with crowdsourcing task performance. Further Hettiachchi et al. [36] propose a model that uses the executive functions of the brain [14] to explain the relationship between cognitive tests and crowdsourcing tasks. However, in their study, workers completed all the cognitive tests and crowdsourcing tasks in a single task unit (*i.e.,* a HIT or Human Intelligent Task in MTurk) which lasts for more than 40 minutes on average. The study does not involve dynamic task routing, but instead conducts an offline analysis of the results. Further, they do not present a system or a framework that demonstrates how cognitive tests can be used to assign tasks or compare the results with any existing task assignment methods. In contrast to the prior work, our experiment replicates typical crowd work conditions where workers have the flexibility to decide on the number of questions they wish to complete and questions are organised into HITs that can be completed in a short time period.

In this paper, we present 'CrowdCog', a real-time online task routing framework that uses cognitive test scores to assign or recommend crowdsourcing tasks to workers. We deploy our study on MTurk and match four different crowdsourcing tasks to workers using the results of five cognitive tests. We show that using our proposed task assignment method, workers are significantly more accurate when compared to a baseline generic task assignment strategy. We also show that workers perform better when following our recommendations instead of selecting the tasks to complete on their own. Further, we compare the performance of our method to a state-of-the-art question assignment method [67] and a standard qualification that uses workers' task completion records. We achieve either similar or better task accuracy through our task routing method that does not use historical data or involve any question selection within the task.

2 RELATED WORK

2.1 Task Assignment

The literature presents a number of task assignment algorithms or frameworks that can be integrated with or used in place of existing crowdsourcing platforms. They consider a variety of different quality metrics (*e.g.,* accuracy, task completion time) and implement one or more quality improvement techniques (*e.g.,* gold standard questions [18], removing erroneous workers [44]) to enhance these quality metrics. The primary motivation behind each assignment method can also be divergent. For example, some methods aim to maximise the quality of the output (*e.g.,* [23, 61, 67]) while other methods attempt to reduce the cost by achieving a reasonable accuracy with a minimum number of workers (*e.g.,* [44]). Task assignment can be further classified into either *heterogeneous task assignment* and *question assignment.*

2.1.1 Task Assignment with Heterogeneous Tasks. As crowdsourcing platforms contain a variety of tasks (*e.g.,* sentiment analysis, classification, transcription), heterogeneous task assignment focuses on matching different task types with workers. Heterogeneous task assignment can be particularly useful in cases where 'expert' workers must be allocated for more difficult tasks [38].

There is limited prior work on heterogeneous task assignment in crowdsourcing. Ho and Vaughan [38] propose a method based on the online primal-dual framework, which has been utilised for different online optimisation problems. In the study, researchers use three types of ellipse classification tasks to account for different expertise levels and use a translation task to simulate different skills. However, their approach assumes that the requester can immediately evaluate the quality of completed work. This vastly limits the applicability of their approach in a real-world crowdsourcing problem. Ho et al. [37] further investigate heterogeneous task assignment in classification tasks with binary labels. However, for the assignment, they use gold standard questions of each task type to estimate the accuracy of the workers.

Assadi et al. [4] studied the task assignment from the requester perspective. They propose an online algorithm that can be used by a requester to maximise the number of tasks allocated with a fixed budget. In a different approach for task assignment, Mo et al. [52] apply a hierarchical Bayesian transfer learning model. They use the historical performance of workers in similar or different type of tasks to estimate the accuracy for the new tasks. Their experiment with a real-world dataset shows the effectiveness of the proposed approach when transferring knowledge from related but different crowd tasks (*e.g.,* questions on sports vs makeup and cooking). However, their real-world evaluation is limited to a single scenario with one source task and one target task. Difallah et al. [16] also propose a system where tasks are allocated based on worker profile data such as interested topics captured from a social media network. The general applicability of this method raises numerous practical and ethical considerations.

While a number of studies have investigated the online task assignment problem, many of them have evaluated only using synthetic data (*e.g.,* [4, 37]). Our study involves a large number of crowd workers and replicates the conditions of typical crowdsourcing platforms.

2.1.2 Question Assignment. Unlike heterogeneous task assignment, the general online task assignment problem has been widely studied in the context of 'question assignment'. In question assignment, which is also often referred to as 'task assignment' (*e.g.,* [44, 67]), the aim is to find the most suitable set of questions from the same task for a given worker.

Zheng et al. [67] propose a task assignment framework, 'QASCA' that uses expectation maximisation on either accuracy or F-score. They experimented on MTurk with five task types including three variants of sentiment labelling of tweets, entity resolution using product descriptions, and selecting which was published earlier from two given films. The method is primarily proposed for multiple

choice questions with a single correct label. QASCA is shown to outperform several other methods including CDAS [48], AskIt! [7], MaxMargin (selecting questions with the highest expected marginal improvement) and ExpLoss (selecting questions based on the expected loss).

'CrowdDQS' is a dynamic task routing mechanism which examines voting patterns and selectively assigns gold standard questions (explicitly verifiable questions) to workers with the aim of identifying and removing workers with poor performance in real-time [44]. The proposed system, which integrates seamlessly with Mechanical Turk, was shown to reduce the number of votes required to accurately answer questions when compared to a round-robin assignment with majority voting. According to the study results, even though CrowdDQS is better than the Expectation Maximisation-based QASCA at worker accuracy estimation, the task accuracy gain is similar.

Fan et al. [23] introduced another dynamic framework named 'iCrowd' that uses a graph-based estimation model to assign tasks to workers with a higher chance of accurately completing the task. They also consider the task similarity when estimating worker accuracy. In another example, Saberi et al. [61] propose a statistical quality control framework (OSQC) for multi-label classification tasks which monitors the performance of workers and removes the workers with high error estimates at the end of processing each batch of tasks. They propose a novel method to estimate the worker accuracy which uses gold standard questions and a plurality answer agreement mechanism. We note that in their evaluation with crowd workers on Amazon's Mechanical Turk, they simulate the past error rates of workers who completed the task, by using a standard normal distribution.

While the literature suggests that these frameworks can produce positive results, their applications are limited for several reasons, such as the fact that these methods have been developed for specific types of crowd work (*e.g.,* [61]) and implemented or tested with a specific crowdsourcing platform (*e.g.,* [44, 61, 67]). One other limitation with regard to benchmarking different methods is the lack of an established crowdsourcing task dataset that spans into different types of crowd tasks.

2.2 Effect of Worker Attributes

When looking at task or question assignment from the workers' perspective, many other worker attributes have been shown to have an impact on crowd task performance. For instance, personality type of the worker is known to be related to the accuracy in relevance labelling tasks [42, 43] and when working in groups [49]. Location of the worker has a significant impact on the task accuracy in content analysis [63] and in relevance labelling [31, 32, 43]. While these studies do not attempt to match workers to tasks based on the said attributes, the results imply that using these approaches is feasible. However, there are inherent difficulties in integrating worker attributes into a task assignment system. Certain attributes like demographics are self-reported by workers. Comprehensive personality tests are time-consuming and there is a possibility for workers to manipulate the outcome. Also, less competent crowd workers tend to overestimate their performance in self-assessments [25].

Previous work has also shown that it is possible predict task performance based on worker behaviour for worker pre-selection [24, 34, 60]. In content creation and information finding tasks, Gadiraju et al. [24] classify workers into five categories using behavioural traces from completed HITs. The study demonstrates that significant accuracy improvements could be achieved by selecting workers to tasks based on given categories. Rzeszotarski and Kittur [60] examined the way workers complete HITs by extracting user activity like mouse movements, scrolling activities, and key-strokes. Their model can successfully predict output quality in content generation, classification and comprehension tasks. Han et al. [34] reported a similar relationship between worker behaviour and content quality in annotating tasks. However, analysing worker behaviour observed over a considerable time period does not provide the utility we aim to achieve through brief cognitive tests.

2.3 Cognitive Ability and Tests

The compatibility between job requirements and the respective worker, and the agreement between job expectations of the worker and the job specifics are two key aspects of the Person-Job fit theorem [46]. This person-job match is known to result in numerous benefits in different work environments such as enhanced job performance, and satisfaction and motivation [19]. Therefore, organisations often seek to achieve a high person-job compatibility for their positions and use a wide variety of performance measures like cognitive ability, personality, general knowledge, emotional intelligence, and work experience [62].

Human cognitive ability has been long identified as an indicator of performance in education [8] and at work [5]. Psychological tests like Stroop [50], Simon [39] and Corsi Block [47] are often used to capture and measure the cognitive ability and are widely used in medical and psychological research [14]. Many of such tests have also been implemented online as test kits or collections like Test My Brain [28] and Cambridge Neuropsychological Test Automated Battery (CANTAB) [59]. In a study that uses Test My Brain, Germine et al. [28] show that it is viable to conduct cognitive tests on the web. Further, Crump et al. [10] conducted a study where crowd workers in MTurk platform were asked to complete cognitive tests such as Stroop, Flanker and Attention Blink. They show that results do not differ from lab-based studies and that it is feasible to use crowdsourcing platforms for such behavioural experiments.

In this study, we aim to use short online cognitive tests to capture the cognitive skills of crowd workers, and use the test outcome to predict their crowd task performance.

2.4 Impact of Cognitive Ability on Crowd Task Performance

Previous work by Eickhoff [20] and Alagarai Sampath et al. [1] indicate the possibility of using cognitive tests for crowdsourcing task assignment. The study by Eickhoff [20] investigates cognitive biases, a closely related trait to cognitive skills. Cognitive biases are known as systematic errors in thinking and can impact peoples everyday judgements and decisions. Literature also reports that cognitive skills can help people avoid cognitive biases [65]. The study shows that cognitive biases negatively impact crowd task performance in relevance labelling. Furthermore, Alagarai Sampath et al. [1] examined the cognitive elements in crowd task design. The study shows that reducing the demand for cognitive work, such as tasks involving visual search and working memory, could lead to higher overall task accuracy.

Goncalves et al. [29] first examined the possibility of using cognitive tests to predict the crowdsourcing task performance using a lab study. While the study reports promising results, it uses a set of time-consuming and paper-based cognitive tests from ETS cognitive kit [21] that are not practical for an online setting. A recent study by Hettiachchi et al. [36] investigates the effect of cognitive abilities on crowdsourcing task performance in an online setting. The work leverages the three executive functions of the brain (inhibition control, cognitive flexibility and working memory) [14] to describe and model the relationship between cognitive tests and crowdsourcing tasks. The study conducted on MTurk with the participation of 102 workers shows that there is a significant correlation between the cognitive test and crowdsourcing task performance. Further, they use multiple models to predict the task performance and show that a worker selection based on predicted scores could lead to better task accuracy.

Our work builds on this prior work as we aim to present an online dynamic task assignment framework that uses cognitive test results to estimate the worker performance, and assign the workers to suitable tasks.

3 STUDY

Next, we detail our experimental design starting with a description of the cognitive tests and crowdsourcing tasks used in the study. Then, we describe the proposed task assignment method, followed by details of the system architecture and study deployment.

3.1 Cognitive Tests

We use five cognitive tests similar to those used in a previous study by Hettiachchi et al. [36]. A description of each cognitive test is provided below, followed by Figure 1 which shows an example of each test. Results from these cognitive tests are used to inform worker task assignment.

3.1.1 ***Stroop*** *Test [50].* Stroop test is one of the classical cognitive tests that evaluate the human ability to overpower the prepotent response to words. In this test, participants encounter three types of trials (incongruent, congruent and unrelated). In incongruent trials, participants see the name of a colour displayed in another colour (*e.g.*, the word "blue" written in a "green" colour as shown in Figure 1). For congruent trials, the name of the colour matches the display colour. In unrelated trials, words displayed are non-colour words. In each trial, the participant needs to ignore the meaning of the word and respond to the colour of the word by pressing a key. Stroop effect states that people are less accurate and slower in incongruent trials when compared with congruent trials.

3.1.2 Eriksen ***Flanker*** *Test [22].* Similar to Stroop test, Flanker test also measures inhibition control but uses a different element. Here, we present 16 trials with two types of images that show five arrow symbols. Congruent trials show all arrows in same direction (*e.g.*, >>>>>) whereas incongruent trials show arrow in the middle in opposite direction (*e.g.*, <<><<). We ask participants to focus on and respond to the symbol in the centre. For the Flanker test, literature reports an effect similar to the Stroop test.

3.1.3 ***Task Switching*** *Test [53].* As shown in Figure 1, in the task switching test, participants see a letter and a number in one of the four squares in a 2×2 layout. In each trial, participants need to respond to one of the two questions; 'is the letter a vowel?' or 'is the number even?' depending on the position the stimuli appearing on the grid. Two types of trials are present in this test. Repeating trials let the participant answer the same question as the previous trial whereas switching trials force participants to change the question from the previous trial. There are 16 trials with 8 of each type.

3.1.4 ***N-Back*** *Test [55].* N-Back test measures the working memory of individuals by asking them to keep track of a series of stimuli. We use the 3-Back version of this test with 16 trials and letters appearing at each trail as shown in Figure 1. Participants are asked to decide if the current letter matches with the one they saw 3 trials ago.

3.1.5 Self-ordered ***Pointing*** *Test [58].* Pointing task tests participants ability to remember a sequence of recent actions. Here, we present 5 trials. In each trial, participants see 3 to 12 squares randomly distributed but identical in size. At any given time, a single square contains a reward. Participants are required to click one square at a time, without repeating until the reward is found. At each click, visual feedback indicates if the reward is found. The reward switches to a different square each time its found and the trial ends when the reward has shifted to all the squares in the trial.

Each cognitive test measures one of the three core executive functions of the brain as detailed in Table 1. *Inhibition control* is the conscious or unconscious restriction of a process or behaviour, particularly of impulses or desires. *Working memory* is the ability to hold information in memory and mentally work with it. *Cognitive flexibility* or Switching is the ability to adapt behaviours in response to changes in the environment and is often associated with creativity [14, 54].

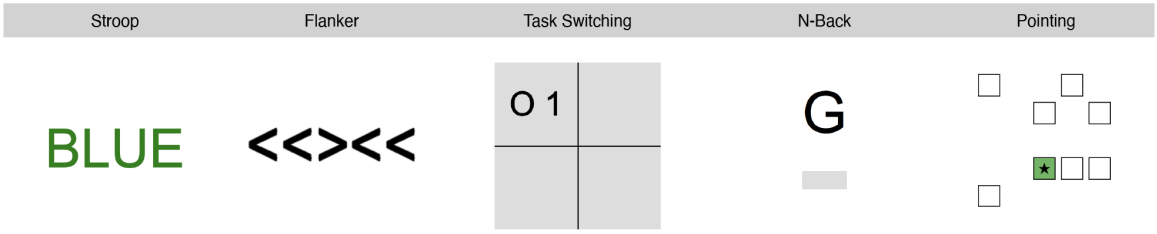

Fig. 1. Examples of each Cognitive Test

Table 1. Cognitive tests and primary executive function they measure [14]

Executive Function	Cognitive Test(s)
Inhibition Control	Stroop and Flanker
Cognitive Flexibility	Task Switching
Working Memory	N-back and Pointing

We provide instructions as well as an example before each test to aid workers. Apart from the pointing test, each trial in all tests is set to expire in 3.5 seconds. This important measure ensures that workers do not pause the test and prevents them from getting distracted while completing the test. We record accuracy and response time for each trial in the Stroop, Flanker, Task Switching, and N-Back tests. For the Pointing test, we gather the number of errors and the mean response time for trials in each round. Additionally, we record and use the trial type to calculate the test effect for Stroop, Flanker and Task Switching tests (*e.g.*, in Stroop test, the difference in accuracy between congruent and incongruent trials is called the Stroop effect related to accuracy).

3.2 Crowdsourcing Tasks

We chose four different crowdsourcing tasks for our experiment. These tasks have been carefully curated based on a crowd task taxonomy [26] and task availability [16] from prior work to be representative of typical tasks available in crowdsourcing platforms. Counting and sentiment analysis tasks were originally utilised by Goncalves et al. [30] and Goncalves et al. [29]. Each crowdsourcing task has multiple unique questions with varying complexity. Both sentiment analysis and counting tasks have 12 questions each while classification and transcription tasks have 9 questions. We also note that these tasks represent different answer types like multiple choice and text input. The tasks can be seen in Figure 2.

*3.2.1 Item **Classification**.* This is a multiple choice question with one or more possible correct answers. Each question contains a painting sourced from The Metropolitan Museum of Art[1] or Flickr[2] where all images are licensed for public use. Workers are given a list of four items and are asked to verify if the items are visible in the painting. Paintings depict a variety of styles that span into different continents. We use the following equation to calculate the accuracy for each question q with a set of A answers provided by a worker and a set of C correct answers. $Accuracy(q,A,C) = max\left[0, \sum_{a \in A} \frac{1}{|C|} \times \begin{cases} 1, & \text{if } a \in C \\ -1, & \text{otherwise} \end{cases}\right]$

*3.2.2 **Counting**.* The counting task presents workers the challenge of counting malaria-infected blood cells in a petri dish which also contain regular blood cells. Images we use in the task were generated using an algorithm to contain varying numbers of infected and regular blood cells. When

[1]https://www.metmuseum.org/art/collection
[2]https://www.flickr.com

workers provide a response a accuracy for each question q of this task with single correct answer c is calculated from $Accuracy(q,a,c) = max\left[0, 1 - \frac{|a-c|}{c}\right]$.

3.2.3 ***Sentiment Analysis***. In this labelling task, workers determine the sentiment of a given sentence witch could be either 'positive', 'negative' or 'neutral'. Task contains two types of sentences. The sentiment of straightforward sentences like "My friends think the price is too expensive" can be easily classified. Other sentences like "Absolutely adore it when my bus is late." are more challenging due to context or language specifics like sarcasm. When a worker provides an answer a to a question q with a correct answer c, we use $Accuracy(q,a,c) = \begin{cases} 1, & \text{if } a=c \\ 0, & \text{otherwise} \end{cases}$ to calculate the accuracy.

3.2.4 ***Transcription***. The transcription task presents workers with an image that contain several text elements. Workers require to recognise and type the text content in a provided text box. We use image segments from The George Washington Papers at the Library of Congress [64] [3]. Due to the time and individual variations in handwriting, selected images have varying complexity. To obtain the accuracy for each question q with a correct answer c and response a, we calculated Levenshtein distance (LD) [11] between the response string and the ground truth and used the equation $Accuracy(q,a,c) = max\left[0, 1 - \frac{2 \times LD(a,c)}{string\_length(c)}\right]$.

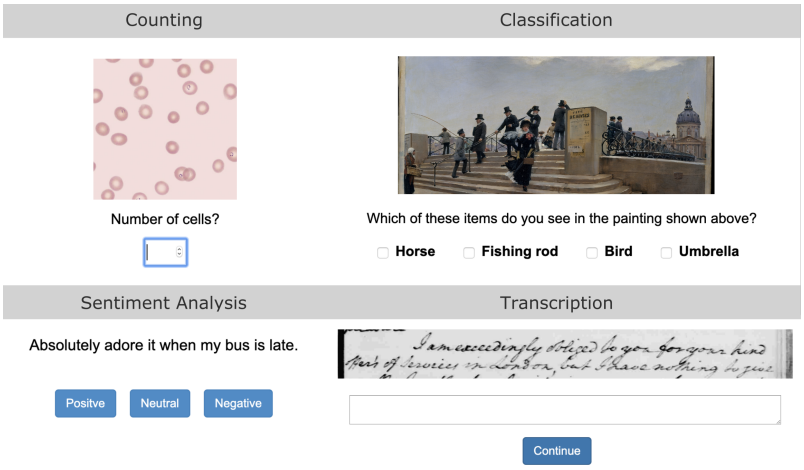

Fig. 2. Examples of Each Crowdsourcing Task

3.3 Task Assignment

3.3.1 Problem. Here, we define the task assignment problem we attempt to solve in this work. Assume that we have a set of tasks $T = \{t_1, .., t_k\}$ and a set of workers $W = \{w_1, .., w_m\}$ where $|T| = k$ and $|W| = m$. Each task t may contain an arbitrary number of questions. In order to maximise the overall quality of the data we gather, for each worker, we aim to assign the task t' where the worker is more likely to produce results of better quality.

The problem we attempt to address in this work is slightly different from question assignment in crowdsourcing, which is also often referred to as 'task assignment' (*e.g.,* [44, 67]). Crowd tasks usually contain several sub-tasks or questions in each task. For example, consider the case shown in Figure 3. There are three tasks (*e.g.,* triangle, square and circle) with each task having four questions. When a worker requests a task, the aim of the task assignment is to select the most suitable task (*e.g.,* circle). Once a task is selected, the question assignment determines the specific question(s) that should be allocated.

[3] https://www.loc.gov/collections/george-washington-papers/about-this-collection

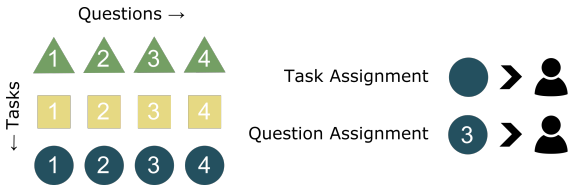

Fig. 3. Task assignment vs Question assignment

We propose two task assignment methods based on cognitive skills of crowd workers. In our first approach "CrowdCog-Assign" we aim to select and assign the optimum task for each worker as determined by our method. Our second approach "CrowdCog-Recommend" is a more relaxed approach where we provide workers with our task recommendation and let them select the task they want to work on. To help readers understand our proposed methods, an overview of the two proposed methods is provided in Figure 4. Here, green coloured elements in dashed line are exclusive to the CrowdCog-Recommend method while blue coloured elements in dotted line solely represent the CrowdCog-Assign method.

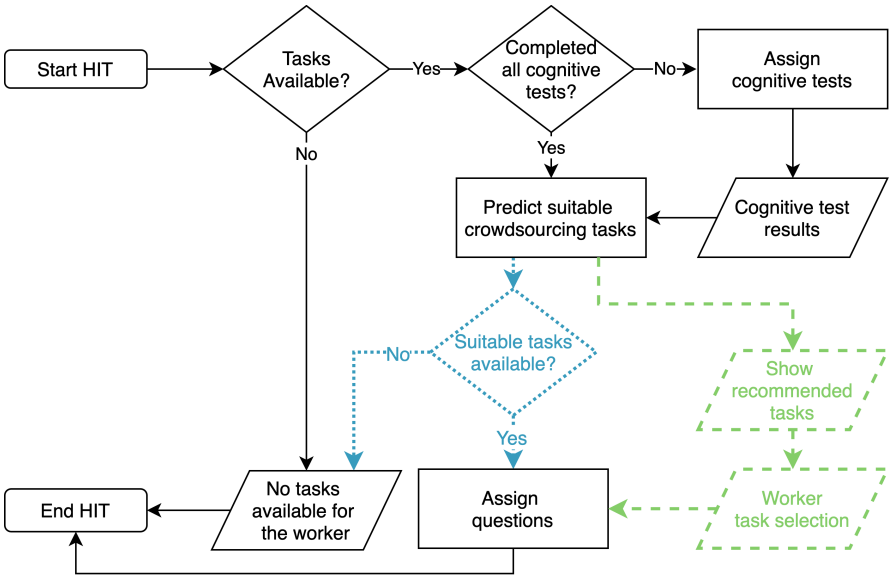

Fig. 4. Flow chart of CrowdCog-Assign in blue dotted line and CrowdCog-Recommend in green dashed line.

3.3.2 CrowdCog-Assign. We introduce a set of cognitive tests, $C = \{c_1,..,c_l\}$ where $|C| = l$ with each test measuring one of the three executive functions (inhibition control, cognitive flexibility, and working memory). We also define two parameters that determine the size of each task unit (*i.e.,* HIT in MTurk). The maximum number of cognitive tests to be included in each HIT, C_{HIT_MAX} and the maximum number of questions to be included in each HIT, Q_{HIT_MAX}.

For each task $t \in T$, we have a set of questions $Q_t = \{q_{t,1},..,q_{t,p}\}$. For each of these questions, we need to obtain an arbitrary number of votes or answers. For simplicity, here we assume all questions in all tasks require a Z number of votes. We also need to keep track of the number of votes or answers provided at a given moment. Lets define $V_t = \{v_{t,1},..,v_{t,p}\}$ where $v_{t,q}$ is the current number of votes or answers received for the question q in task t.

When a worker starts a HIT, we check if there are tasks available for the worker. This check is based on two steps. First, we obtain a list of questions that still need to be answered. For any task t, an available question $q_{t,j}$ is where $v_{t,j} < Z$. Second, for each question in the available question list, we remove questions the worker has already attempted based on the worker task completion history.

Then, we select the tasks that correspond to the remaining questions. At the end of this filtering step, we have a list of tasks T' that could be potentially assigned to the worker. If there are no tasks in the list, we end the HIT with a message to the worker.

Then we assign cognitive tests for the worker. Here, we keep track of the cognitive tests the worker has already completed C^w and first obtain a list of tests the worker has not completed yet ($C - C^w$). Then, we randomly assign a C_{HIT_MAX} number of tests or the total number of tests if $|C - C^w| < C_{HIT\_MAX}$. If the worker has already completed all the cognitive tests, we skip this step and directly move to task prediction. Following the cognitive test assignment, the worker will complete all the assigned tests and upon completion, the system will receive the results. Once we receive the cognitive test results we attempt to assign a task to the worker. Task prediction is based on the model and the relationship between cognitive tests and crowdsourcing tasks proposed by Hettiachchi et al. [36]. We use individual random forest models for each task with parameters number of trees set to 1000 and features selected at each split to 3.

Based on Table 1 and Table 2 and from prior work, we already know the set of cognitive tests (C^t) that a worker needs to complete in order for us predict the accuracy of that worker for a particular task t. For instance, if a worker has completed all the cognitive tests related to Cognitive Flexibility (*e.g.,* Task Switching), we can predict the accuracy for Transcription task using our model. Likewise, we predict the accuracy of all the available tasks for the current worker T^w. Then for each task t, if the predicted accuracy, $accuracy_{w,t}$ for worker w is higher than the pre-determined threshold $accuracy_t^0$, we consider that task as a candidate for the assignment. Finally, we select a task from the possible assignments. In our study, we select a random task from the candidate list to best replicate a real-life crowd-market scenario where workers would be allowed to attempt many tasks that they qualify for as based on the results of a common set of cognitive tests. Therefore using our model, we can find the task that should be assigned to the worker as detailed in the Algorithm 1.

Table 2. Relationship between Crowdsourcing Tasks and Cognitive Tests [36]

Crowd Task	Significant Features	Related Executive Functions
Classification	Pointing (Accuracy) Flanker (Resp. Time) Stroop (Accuracy)	Inhibition Control Working Memory
Counting	Flanker (Effect Accuracy) Pointing (Resp. Time) Stroop (Accuracy)	Inhibition Control Working Memory
Sentiment Analysis	Stroop (Resp. Time) Instructions (Resp. Time) Flanker (Effect Accuracy)	Inhibition Control
Transcription	Task Switching (Accuracy) Task Switching (Effect Accuracy)	Cognitive Flexibility

Following the task assignment, we select the questions to be assigned to the crowd workers. For this purpose, we also keep track of the number of answers still required for each question to avoid redundant labels. The worker then completes the assigned questions. As the final step, we collect the responses for questions, record them and mark the HIT as submitted in the Amazon Mechanical Turk platform.

$C^w \leftarrow$ Set of cognitive tests completed by worker w
$T^w \leftarrow$ Set of available tasks for worker w
assignment $\leftarrow$ The task assigned for the worker w
input : C^w,T^w
output: *assignment*

$possible\_assignments \leftarrow \emptyset$; $assignment \leftarrow \varnothing$;
foreach $t \in T^w$ **do**
  if $\forall c(c \in C^t \cap C^w)$ **then**
    $accuracy_{w,t} \leftarrow Predict(c)$;
    if $accuracy_{w,t} > accuracy_t^0$ **then**
      $possible\_assignments \leftarrow possible\_assignments \cup \{t\}$;
    end
  end
end
if *possible_assignments is not* $\emptyset$ **then**
  $assignment \leftarrow RandomSample(possible\_assignments)$
end

Algorithm 1: Task assignment based on cognitive test results

3.3.3 CrowdCog-Recommend. A key restriction in the proposed CrowdCog-Assign assignment strategy is that it does not let workers select the tasks they want to work on. While this may have a positive impact on the performance, in certain cases, a crowdsourcing platform might still prefer to provide workers with the flexibility of selecting their own tasks. To allow for this, we propose the CrowdCog-Recommend method.

For this approach, we follow a similar process as the CrowdCog-Assign method until tasks are predicted from cognitive test results. As we finish iterating over tasks in T^w, we return *possible_assignments* without selecting a single task (See Algorithm 1). Instead of assigning the task, here we present workers with our task recommendation and ask them to select the task they want to work on. After task selection, the rest of the process is identical to the CrowdCog-Assign method.

3.4 Study Conditions

The study was conducted under five conditions as described below.

- *Baseline*: In the baseline, workers select the task they want to work on and the questions are randomly assigned by the system. The baseline is comparable to the task assignment in a generic crowdsourcing platform like MTurk.
- *CrowdCog-Assign*: The proposed method where tasks are directly assigned based on the cognitive test performance and questions are assigned randomly.
- *CrowdCog-Recommend*: The proposed method with tasks recommended using cognitive test results. Workers see the recommendation but still have the liberty to choose any task. Questions are assigned randomly.
- *QASCA*: We compare with QASCA proposed by Zheng et al. [67]. Under QASCA, workers select the task but questions are assigned based on Expectation Maximisation. We chose QASCA as it has been shown to perform better when compared to four other methods CDAS [48], AskIt! [7], MaxMargin and ExpLoss.
- *History-based*: Under this method which uses historical data, task and question selection is similar to the baseline. However, workers are allowed to attempt tasks only if they have

previously completed 1000 HITs in the platform with an approval rate of 95% or above. This worker selection criteria is widely utilised by researchers and the literature reports a significant increase in data quality when selecting workers with a high approval rate and a high number of HITs completed [57].

We deployed all four tasks under these conditions. As QASCA is originally proposed for multi-label questions with a single correct answer, we were not able test transcription task which gathers text input. Also under QASCA, we had to transform the answers for counting tasks into three labels using a bracketing method as suggested in the prior work [56]. For classification task which contains multiple correct labels, we only considered a single option when evaluating with QASCA.

For CrowdCog-Assign and CrowdCog-Recommend conditions, each HIT contained a maximum of 2 cognitive tests ($C_{HIT\_MAX} = 2$) as we need results from at least two cognitive tests to make a task assignment or a recommendation (See Table 2). Each HIT also included a maximum of 3 questions ($Q_{HIT\_MAX} = 3$) to be consistent with the study design of prior work [67] and to ensure we can equally distribute all questions within a task (our tasks contain either 12 or 9 questions). For the evaluation, we set the threshold for task assignment ($accuracy_t^0$) at a modest 50% accuracy of each task. This threshold can be adjusted by the requester depending on the urgency of the data collection and available funds.

Each condition was deployed in MTurk at independent iterations. Each iteration was deployed during the same time window on weekdays. Using a qualification, we prevented any worker from attempting tasks in more than one condition. We only allowed workers from the United States and workers were compensated at the rate of $0.4 (USD) for each HIT. The payment was decided based on the time estimations gathered from our pilot study and the highest state minimum wage of the United States $13.25. Workers were compensated with a bonus payment of $ 0.2 (USD) for each cognitive test they completed in addition to the tasks. We ensured the bonus payment is issued for cognitive tests even when no tasks were assigned to the workers. For all conditions except history-based, we did not employ any additional worker selection criteria like approval rate. The research is approved by the ethics committee of our university. When participants accepted their first HIT from our study, they were also required to accept an informed consent form in order to continue the study.

We built our system primarily using Python (Django Framework). The system was hosted in a standalone server and workers accessed tasks through the external task function in MTurk. The experiment was presented to the worker through a popup window that automatically submits the HIT at the end. Several elements that allow for this seamless integration with the MTurk platform were extended from *PsiTurk*, an open platform for building experiments on MTurk [33]. For the creation of cognitive tests, we also used *jsPsych*, a JavaScript library for running behavioural experiments in a web browser [13].

4 RESULTS

In our study, a total of 574 workers completed 983 task units (HITs) across five conditions. Completed HITs accounted for 838 cognitive tests and 1,703 answers for crowdsourcing tasks. On average workers spent 2.95 minutes on HITs that contained crowd tasks and 2.98 minutes on HITs that contained both cognitive tests and crowd tasks. For the analysis, we use task accuracy as the primary evaluation metric which is calculated as described under the crowdsourcing tasks section (Section 3.2).

4.1 Cognitive Test Validation

Participant responses collected for the three cognitive tests can be validated using the difference in trial accuracy and response time between different types of trials. For example, a one-sample Wilcoxon signed rank test shows that the difference in accuracy ($M = 0.13, SD = 0.22$) between congruent and incongruent trials in Stroop test is significantly higher than 0 ($V = 3377.5, p < 0.001$) whereas

a one-sample t-test shows the difference in response time ($M = -196.68$, $SD = 258.67$) is significantly lower than 0 ($t(183) = -10.31$, $p < 0.001$). Similarly, the difference in accuracy ($M = 0.25$, $SD = 0.39$) was significantly higher than 0 ($V = 4761.5$, $p < 0.001$) and response time ($M = -97.20$, $SD = 236.93$) was significantly lower than 0 ($t(171) = -5.38$,$p < 0.001$) for the Flanker test. In the Task Switching test, difference in accuracy ($M = 0.01$, $SD = 0.20$) and response time ($M = -17.61$, $SD = 378.65$) between switching and repeating trials are not significantly different from 0 as opposed to the Stroop and Flanker tests. The difference in direction follows the findings from prior literature [22, 50, 53].

4.2 Task Recommendation

For tasks completed under the CrowdCog-Recommend condition, we analyse the difference in accuracy between two cases. First, under *No Recommendation*, workers attempt a task when there is no task recommendation given from the system. Second, under *Attempt Recommended*, workers attempt a task that was recommended by the system. Figure 5 shows that workers performed better when attempting recommended tasks when compared to other tasks. A Wilcoxon rank sum test shows that task accuracy for Attempt Recommended case is significantly higher when compared to the No Recommendation case ($W = 21034$, $p < 0.01$). We also note that workers were more likely to accept a recommendation. In our CrowdCog-Recommend setting, workers were presented with a task recommendation in 89 HITs. Workers opted to work on a recommended task in 61 HITs (68.53%).

Fig. 5. Accuracy and Standard Error for each task for the task recommendation conditions

4.3 Task Assignment

Under the CrowdCog-Assign setting, 239 unique workers initiated our HIT and 63 (35.80%) of them were assigned to one or more tasks. Out of 176 workers who were not assigned to tasks, 156 (88.60%) workers did not attempt more than a single HIT which includes only 2 cognitive tests. In Figure 6 we observe that as workers completed more cognitive tests, they were more likely to be assigned to a task. We validate this observation through a Chi-squared test ($\chi^2 = 85.39$,$p < 0.001$). Further, when considering workers who completed all five tests, 72% of them were assigned to at least one task.

Fig. 6. Variation in task assignment against the number of cognitive tests completed

4.4 Comparing CrowdCog to Other Methods

We analyse and compare the performance of proposed CrowdCog methods with three other conditions: baseline, QASCA and history-based method. For CogCrowd-Assign we also included the answers obtained under attempt-recommended of CogCrowd-Recommend.

We report a significant improvement in the accuracy of the workers compared to the baseline. As the study comprises of tasks accounting for both discrete and continuous accuracy values, our data does not pass the Levene's test for homogeneity of variance and Shapiro-Wilk normality test. Hence, we use Kruskal-Wallis rank sum test and report a significant difference in accuracy ($\chi^2 = 32.37$, $p < 0.01$, df = 4) among five conditions. Further, we conduct a post-hoc analysis via Dunn Test with p-values adjusted with the Benjamini-Hochberg method. Results show that when compared to the baseline, the accuracy is significantly higher in the CrowdCog-Assign method ($Z = -4.17$, $p < 0.01$) as well as in the CrowdCog-Recommend method ($Z = -2.51$, $p = 0.02$). While accuracy of CrowdCog-Assign method is significantly higher when compared to QASCA ($Z = -2.64$, $p = 0.02$), there is no significant difference in accuracy between history-based method and CogCrowd-Assign ($Z = 1.11$, $p = 0.27$). Figure 7 visualises the mean accuracy and standard error values for all the tasks across the baseline and proposed methods. Accuracy values are also summarised in Table 3.

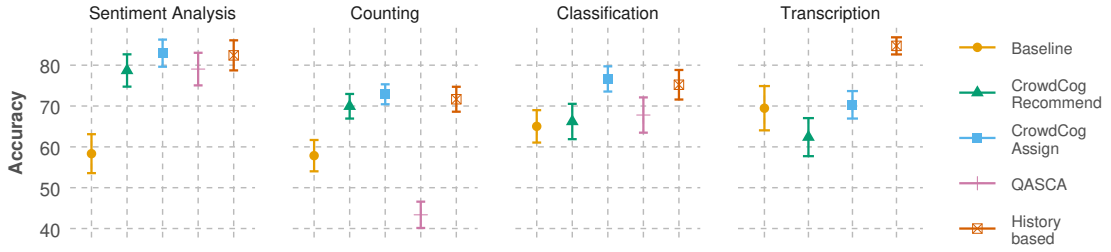

Fig. 7. Accuracy and Standard Error of tasks

Table 3. Task Accuracy across conditions

Condition	Baseline	CrowdCog Rec.	CrowdCog Assign	QASCA	History based
Sentiment Analysis	58.3	78.7	**82.9**	79.0	82.4
Counting	57.8	69.9	**72.9**	43.4	71.7
Classification	65.0	66.2	**76.6**	67.8	75.2
Classification [a]	64.2	78.0	**85.6**	71.6	80.0
Transcription	69.5	62.4	70.3	-	**84.7**

[a] Accuracy calculated considering only a single option to be comparable with QASCA

Figure 8 shows the mean response time in seconds for each task across three conditions. Although workers appear to be generally faster in our CrowdCog-Assign condition for most tasks, we do not observe any statistically significant difference in terms of response time across conditions.

To examine whether we have collected a sufficient number of responses for the tasks, we observe the variation in accuracy as we gather participant answers. Figure 9 shows that the accuracy is relatively stable after we accumulate 50% of the answers for sentiment analysis, counting, and classification tasks.

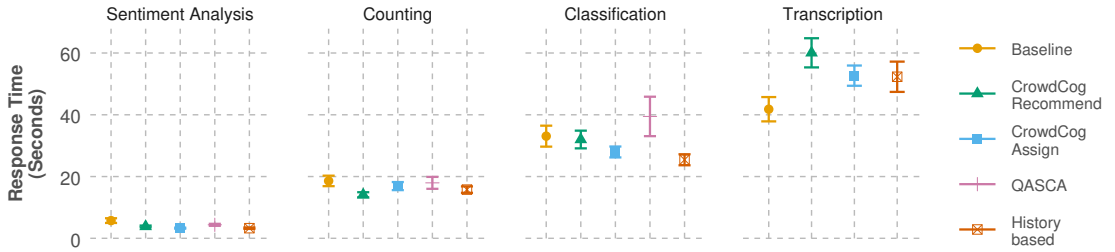

Fig. 8. Response Time and Standard Error of tasks

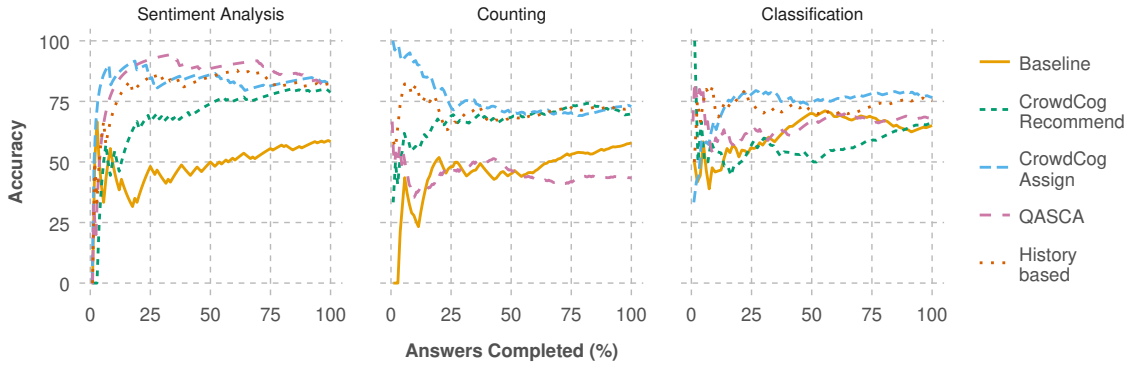

Fig. 9. Task accuracy against the answers completed

4.5 Cost Analysis

Our study included 42 questions across four tasks (Counting - 12, Classification - 9, Sentiment Analysis - 12, Transcription - 9) and we collected 9 answers for each question under different conditions. Here, in order to analyse the costs, we consider the order in which we received these answers and calculate the task accuracy by aggregating a varying number of answers. Figure 10 shows that fewer answers with CrowdCog-Assign method is sufficient to outperform the baseline with a larger number of answers. Next, we present a cost analysis where we only consider the first 3 answers for CrowdCog-Assign method and compare it against the baseline with 9 answers.

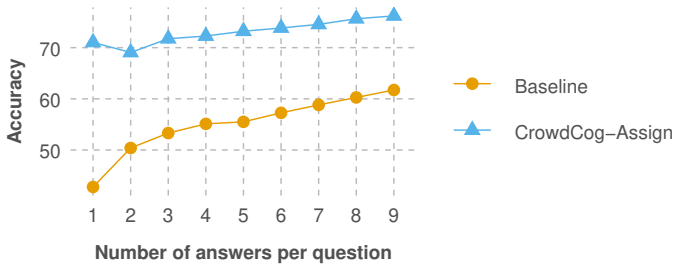

Fig. 10. Variation in task accuracy against the total number of answers aggregated

We show in Figure 11 that for all the tasks, the accuracy obtained from 3 answers per question under CrowdCog-Assign method is still higher than the accuracy from baseline with 9 answers per question. We calculate the total cost for 42 questions under the two conditions. First, under baseline, the cost is straightforward. As each answer costs \$0.13 (workers were payed \$0.4 for a HIT containing 3 questions), the total cost for obtaining 9 answers each for all the questions is $\$0.13 \times 9 \times 42 = \$49.14$.

Second, under CrowdCog-Assign method, the cost for all the answers would be $\$0.13 \times 3 \times 42 = \$16.38$. The additional cost for cognitive tests depend on the number of workers required for the task. We estimate the number of workers needed to obtain 3 answers, based on the number of

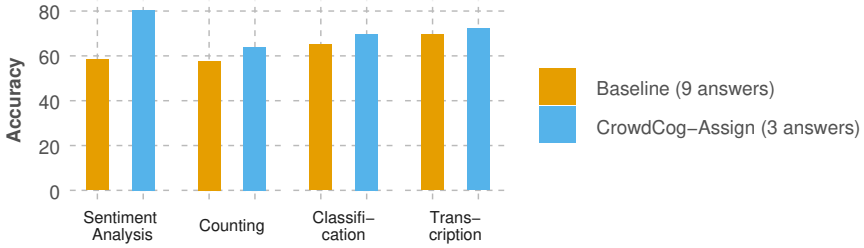

Fig. 11. Task accuracy with first 3 answers of each question from CrowdCog-Assign vs 9 from baseline

workers completed the study under this condition providing 9 answers for each question (174) and their cognitive test completion rates. The results show that 72.8% of workers completed only a single HIT (2 cognitive) tests, 6.3% completed two HITs (4 cognitive tests) and 20.9% completed three or more HITs (all 5 cognitive tests). Therefore, we determine the cost for cognitive tests $174 \times \frac{3}{9} \times (0.728 \times 2 + 0.063 \times 4 + 0.209 \times 5) \times \$0.2 = \$31.93$. Hence, the total cost for CrowdCog-Assign method adds up to \$48.31 in total. From Figure 11 and the calculated costs (Baseline \$49.14 and CrowdCog-Assign \$48.31), we show that the proposed CrowdCog-Assign method is capable of producing better results than the baseline at the same cost. While QASCA and history based methods do not result in additional costs, unlike CrowdCog, history based methods are not applicable for new workers and QASCA requires task specific calculations at each HIT submission.

5 DISCUSSION

5.1 CrowdCog Task Assignment

Crowdsourcing literature identifies task assignment in crowd platforms as one of the research foci [45]. Appropriate task assignment has many positive outcomes. From the perspective of a task requester, data quality can be increased while reducing the number of required labels, maximising cost-benefit. In the absence of task assignment, workers can find it challenging to locate appropriate tasks and tend to prioritise recently posted or new tasks, as well as tasks with the most number of HITs [9]. This also leads to requesters repeatedly posting the same task and flooding the platforms to attract workers [6]. If a platform is able to assign workers with compatible tasks, it will benefit workers by reducing the time and effort needed for task search and increasing worker satisfaction by achieving better person-job fit [19].

While numerous task assignment methods have been proposed, we note several shortcomings such as the inability to cater for a wide range of tasks (*e.g.,* [37, 44, 67]), and reliance on prior task records or external data (*e.g.,* [16, 24, 52, 60]). Concerning the validation of these previously introduced assignment methods, many evaluations are limited to synthetic data (*e.g.,* [4, 37]), one or two tasks (*e.g.,* [52]), or an offline analysis as opposed to online dynamic task assignment (*e.g.,* [24, 29, 36, 60]).

Our results indicate that when compared to the baseline (workers select the task without any recommendations), the proposed CrowdCog-Assign method (tasks assigned based on worker's cognitive test performance) produces significantly more accurate results. This increase in worker accuracy ranges from 5% to 20% across a variety of different task types. We also show that our method which works with new workers can achieve similar results compared to a widely used worker qualification that relies on historical data. In addition, the history-based method aims to restrict the available worker pool to a limited subset of workers who generally perform well across tasks. In contrast, we show that our method can successfully match workers to different types of tasks. Under CrowdCog-Assign, 72% of the workers who completed all five cognitive tests were assigned to at least one task.

We highlight that the proposed method is straightforward to implement and can be practised by both task requesters and platforms. However, a platform-level implementation could yield greater benefits. Once worker cognitive test results are captured, they could be utilised to assign many tasks. We highlight several factors that should be considered. First, as the cognitive ability of a worker could vary over time [17], cognitive tests should be repeated at a reasonable frequency. When repeating tests, the pool of tests would ideally consist of multiple tests for each executive function (*e.g.,* Stroop, go/no-go, Simon and many other tests for Inhibition Control [14]) as well as variants of the same test (*e.g.,* Stroop Test [50]) to ensure workers do not get familiarised with tests. Nevertheless, as cognitive tests include fast-paced time-restricted trials, workers would find it difficult to manipulate the outcomes [10] when compared to other task-independent approaches that could work without historical records such as demographics [63], personality tests [42] and self-assessments [25]. Second, when finding the relevant cognitive test for a particular task that does not relate to any of the tasks examined in our work, future researchers will have to identify the related executive functions of the task. They can replicate the approach detailed by Hettiachchi et al. [36] to build a hypothesis based on broad literature on human psychology. Alternatively (or in addition), a pilot implementation that includes three cognitive tests representing three executive functions can be used to determine what executive functions relate well to specific tasks. Once the relevant executive functions are identified, it is straightforward to determine the relevant cognitive test [14]. Third, an accuracy threshold needs to be set (see Algorithm 1) for each task before assigning tasks. This could be achieved via a pilot task set or using values based on our work. The threshold could also vary depending on the urgency of data collection. A lower threshold will result in an increased data collection rate but a lower accuracy increment as compared to the baseline.

Naturally, crowd task requesters are cautious of the additional costs that can be associated with more complex task assignment methods or quality control mechanisms [2, 35]. For the majority of common methods, such as gold standard questions and qualification tests, this additional cost is repeated for every new task. We supplement our study with a cost analysis to emphasise that cognitive tests could be incorporated in a crowdsourcing marketplace without increasing the potential costs. As shown in Figure 10, a reduction in the total number of answers required when applying our method compensates for the additional expenses required for cognitive tests. Further, when compared to the number of questions we have in our tasks (12 or 9), a typical crowd task has a sufficient number of questions [15, 41] to account for the additional amount requesters need to invest on cognitive tests.

5.2 Task vs Question Assignment

As the end goal of data quality improvement in crowdsourcing could be achieved through both task and question assignment, we argue that our comparison with question assignment methods is important. Question assignment methods also represent a large portion of rigorous frameworks proposed in the literature [12]. Based on the results of our study, we establish that the performance of our method is better or similar to the state-of-the-art question assignment methods. When considering the performance of the counting task, we observe that the task accuracy for QASCA is not significantly different from the baseline. Each question in the counting task has a single numeric input which we transformed into three groups using bracketing to apply expectation maximisation. This is the probable reason for the sub-par performance. Although prior work on QASCA suggests bracketing for handling questions with numeric input, they only experiment with multiple choice questions with a single correct label [67].

Another important consideration when using a real-time task assignment method is the impact on performance. If we deploy a sophisticated question assignment method such as QASCA, we need to carry out certain calculations at the end of each HIT which typically contains one or a few questions. This accumulates to a high demand for computational power when we consider the task completion rate in a standard crowdsourcing platform [15]. Therefore, unless the requester maintains

a third party resource that can calculate real-time scores, it can be quite challenging to implement a question assignment method like QASCA within a crowdsourcing platform. Our method provides a less computationally costly solution by reusing the worker cognitive test results for estimating performance for a variety of tasks.

Further, we note that our method could be used along with any question assignment method. For instance, a platform could implement our proposed method for task selection and use any of the question selection methods for question selection. While such a fine-grained task assignment implementation would be complex and computationally intensive, it could potentially increase the accuracy even further.

5.3 Task Recommendation

While task assignment aims to maximise the overall performance, it is important to consider potential negative consequences for the workers in terms of agency. In crowdsourcing, 'self-identification of contributors' [40] or workers' liberty to attempt a task they prefer is deemed important. Thus, task recommendation is often considered a more flexible alternative to task assignment [27]. Our work shows that the use of task recommendation based on cognitive skills still achieves significantly higher task performance when compared to the baseline. Prior attempts on task recommendation in crowdsourcing mainly rely on user-provided profile data, feedback collected from previous tasks [3], or worker task browsing history [66]. Also, Geiger and Schader [27] in a review of crowdsourcing task recommendation systems, identify the lack of an online analysis as a major drawback of the previous studies. In our study, we apply an online empirical analysis which shows that task recommendation based on workers' cognitive ability can lead to higher data quality when compared to a baseline of worker task selection.

In addition to the positive task recommendations applied in this paper, future work could potentially indicate negative task recommendations for tasks that are not recommended for a worker. This will allow workers to distinguish between tasks that are not recommended for them based on cognitive tests and tasks for which we are unable to make a prediction.

5.4 Limitations

We acknowledge several limitations of our study. First, many online task assignment frameworks often experiment with synthetic data to validate the proposed methods (*e.g.,* [4, 37, 61]). A handful of these studies have complemented the synthetic study results with a small scale real-time deployment on a platform like MTurk (*e.g.,* [44, 67]). However, because our results are based on cognitive tests, we only validate our method using a real-world deployment albeit with a high number of crowd workers. Unlike synthetic studies, real-world deployment limits our ability to extensively explore different parameter configurations. Second, we do not compare with any of the heterogeneous task assignment methods [4, 52]. This is mainly due to the incompatibility with our study setting and complexity in implementation of such proposed methods. However, we do compare with state-of-the-art question assignment methods.

6 CONCLUSION AND FUTURE WORK

In this paper, we study the heterogeneous task assignment problem through a novel and online assignment and recommendation method. We propose the use of short online cognitive tests for dynamic task assignment in a crowdsourcing platform across a variety of tasks. We built the CrowdCog system by integrating our novel task assignment and recommendation framework with MTurk. We evaluate the system using a real world study involving 574 crowd workers and 983 HITs across four tasks. Our study compares the proposed task assignment and task recommendation methods with a baseline generic task assignment and reports significantly higher task accuracy in both cases. We also show

that the proposed methods are comparable in improving worker's task accuracy when compared to state-of-the-art question assignment methods as well as a standard history-based qualification. At the same time, our method has a number of additional advantages, such as applicability to a variety of different tasks, not relying on historical performance data, and a better person-job fit which has been shown to lead to higher worker satisfaction [19].

Future work could explore a selection mechanism that takes into account the current task availability and cognitive test completion of the worker to further enhance the efficiency and productivity of the proposed method. Furthermore, once we have a list of eligible tasks for a worker, we randomly select a task from the list as opposed to the use of an optimised selection method. While this selection is less likely to impact the accuracy, an informed selection at this stage could further improve the efficiency of the data collection process. However, as both these enhancements dependent on various factors, future work in this domain will require a carefully crafted study design to account for the added complexity. In addition, a longitudinal study which investigates the frequency with which the cognitive tests should be repeated and the strategies for reusing cognitive tests will further strengthen the applicability of our findings.

REFERENCES

[1] Harini Alagarai Sampath, Rajeev Rajeshuni, and Bipin Indurkhya. 2014. Cognitively Inspired Task Design to Improve User Performance on Crowdsourcing Platforms. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. ACM, New York, NY, USA, 3665–3674. https://doi.org/10.1145/2556288.2557155

[2] Mohammad Allahbakhsh, Boualem Benatallah, Aleksandar Ignjatovic, Hamid Reza Motahari-Nezhad, Elisa Bertino, and Schahram Dustdar. 2013. Quality Control in Crowdsourcing Systems: Issues and Directions. *IEEE Internet Computing* 17, 2 (2013), 76–81. https://doi.org/10.1109/MIC.2013.20

[3] Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2011. Towards Task Recommendation in Micro-task Markets. In *Human Computation Workshop at the Twenty-Fifth AAAI Conference on Artificial Intelligence*. AAAI Press, Palo Alto, California, USA.

[4] Sepehr Assadi, Justin Hsu, and Shahin Jabbari. 2015. Online assignment of heterogeneous tasks in crowdsourcing markets. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing (HCOMP'15)*. AAAI Press, Palo Alto, California, USA.

[5] Charles E. Bailey. 2007. Cognitive Accuracy and Intelligent Executive Function in the Brain and in Business. *Annals of the New York Academy of Sciences* 1118, 1 (2007), 122–141. https://doi.org/10.1196/annals.1412.011

[6] Michael S. Bernstein, David R. Karger, Robert C. Miller, and Joel Brandt. 2012. Analytic methods for optimizing realtime crowdsourcing. In *Proceedings of Collective Intelligence 2012*.

[7] R. Boim, O. Greenshpan, T. Milo, S. Novgorodov, N. Polyzotis, and W. Tan. 2012. Asking the Right Questions in Crowd Data Sourcing. In *2012 IEEE 28th International Conference on Data Engineering*. 1261–1264. https://doi.org/10.1109/ICDE.2012.122

[8] Erika Borella, Barbara Carretti, and Santiago Pelegrina. 2010. The Specific Role of Inhibition in Reading Comprehension in Good and Poor Comprehenders. *Journal of Learning Disabilities* 43, 6 (2010), 541–552. https://doi.org/10.1177/0022219410371676

[9] Lydia B. Chilton, John J. Horton, Robert C. Miller, and Shiri Azenkot. 2010. Task search in a human computation market. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*. ACM, 1–9.

[10] Matthew J. C. Crump, John V. McDonnell, and Todd M. Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE* 8, 3 (2013), 1–18. https://doi.org/10.1371/journal.pone.0057410

[11] Fred J. Damerau. 1964. A Technique for Computer Detection and Correction of Spelling Errors. *Commun. ACM* 7, 3 (1964), 171–176. https://doi.org/10.1145/363958.363994

[12] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Comput. Surv.* 51, 1, Article 7 (Jan. 2018), 40 pages. https://doi.org/10.1145/3148148

[13] Joshua R. de Leeuw. 2015. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods* 47, 1 (2015), 1–12. https://doi.org/10.3758/s13428-014-0458-y

[14] Adele Diamond. 2013. Executive Functions. *Annual Review of Psychology* 64, 1 (2013), 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

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

[16] Djellel E. Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2013. Pick-a-crowd: Tell Me What You Like, and I'll Tell You What to Do. In *Proceedings of the 22nd International Conference on World Wide Web* (Rio de Janeiro, Brazil) *(WWW '13)*. ACM, New York, NY, USA, 367–374. https://doi.org/10.1145/2488388.2488421

[17] Tilman Dingler, Albrecht Schmidt, and Tonja Machulla. 2017. Building Cognition-Aware Systems: A Mobile Toolkit for Extracting Time-of-Day Fluctuations of Cognitive Performance. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 47 (Sept. 2017), 15 pages. https://doi.org/10.1145/3132025

[18] Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. Are Your Participants Gaming the System?: Screening Mechanical Turk Workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. ACM, New York, NY, USA, 2399–2402. https://doi.org/10.1145/1753326.1753688

[19] Jeffrey R. Edwards. 1991. *Person-job fit: A conceptual integration, literature review, and methodological critique.* John Wiley & Sons.

[20] Carsten Eickhoff. 2018. Cognitive Biases in Crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA) *(WSDM '18)*. ACM, New York, NY, USA, 162–170. https://doi.org/10.1145/3159652.3159654

[21] Ruth B. Ekstrom, Diran Dermen, and Harry Horace Harman. 1976. *Manual for kit of factor-referenced cognitive tests.* Vol. 102. Educational Testing Service, Princeton, NJ, USA.

[22] Barbara A. Eriksen and Charles W. Eriksen. 1974. Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics* 16, 1 (1974), 143–149. https://doi.org/10.3758/BF03203267

[23] Ju Fan, Guoliang Li, Beng Chin Ooi, Kian-lee Tan, and Jianhua Feng. 2015. iCrowd: An Adaptive Crowdsourcing Framework. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (Melbourne, Victoria, Australia) *(SIGMOD '15)*. ACM, New York, NY, USA, 1015–1030. https://doi.org/10.1145/2723372.2750550

[24] Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. 2018. Crowd Anatomy Beyond the Good and Bad: Behavioral Traces for Crowd Worker Modeling and Pre-selection. *Computer Supported Cooperative Work (CSCW)* (Jun 2018). https://doi.org/10.1007/s10606-018-9336-y

[25] Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. 2017. Using Worker Self-Assessments for Competence-Based Pre-Selection in Crowdsourcing Microtasks. *ACM Trans. Comput.-Hum. Interact.* 24, 4, Article 30 (Aug 2017), 26 pages. https://doi.org/10.1145/3119930

[26] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. 2014. A Taxonomy of Microtasks on the Web. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media* (Santiago, Chile) *(HT '14)*. ACM, New York, NY, USA, 218–223. https://doi.org/10.1145/2631775.2631819

[27] David Geiger and Martin Schader. 2014. Personalized task recommendation in crowdsourcing information systems – Current state of the art. *Decision Support Systems* 65 (2014), 3–16. https://doi.org/10.1016/j.dss.2014.05.007

[28] Laura Germine, Ken Nakayama, Bradley C. Duchaine, Christopher F. Chabris, Garga Chatterjee, and Jeremy B. Wilmer. 2012. Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review* 19, 5 (2012), 847–857. https://doi.org/10.3758/s13423-012-0296-9

[29] Jorge Goncalves, Michael Feldman, Subingqian Hu, Vassilis Kostakos, and Abraham Bernstein. 2017. Task Routing and Assignment in Crowdsourcing Based on Cognitive Abilities. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1023–1031. https://doi.org/10.1145/3041021.3055128

[30] 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. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Zurich, Switzerland) *(UbiComp '13)*. ACM, New York, NY, USA, 753–762. https://doi.org/10.1145/2493432.2493481

[31] Jorge Goncalves, Simo Hosio, Denzil Ferreira, and Vassilis Kostakos. 2014. Game of Words: Tagging Places through Crowdsourcing on Public Displays. In *Proceedings of the 2014 Conference on Designing Interactive Systems* (Vancouver, BC, Canada) *(DIS '14)*. Association for Computing Machinery, New York, NY, USA, 705–714. https://doi.org/10.1145/2598510.2598514

[32] Jorge Goncalves, Simo Hosio, Niels van Berkel, Furqan Ahmed, and Vassilis Kostakos. 2017. CrowdPickUp: Crowdsourcing Task Pickup in the Wild. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 51 (Sept. 2017), 22 pages. https://doi.org/10.1145/3130916

[33] Todd M. Gureckis, Jay Martin, John McDonnell, Alexander S. Rich, Doug Markant, Anna Coenen, David Halpern, Jessica B. Hamrick, and Patricia Chan. 2016. psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods* 48, 3 (01 Sep 2016), 829–842. https://doi.org/10.3758/s13428-015-0642-8

[34] Shuguang Han, Peng Dai, Praveen Paritosh, and David Huynh. 2016. Crowdsourcing Human Annotation on Web Page Structure: Infrastructure Design and Behavior-Based Quality Control. *ACM Trans. Intell. Syst. Technol.* 7, 4, Article 56 (April 2016), 25 pages. https://doi.org/10.1145/2870649

[35] Danula Hettiachchi, Zhanna Sarsenbayeva, Fraser Allison, Niels van Berkel, Tilman Dingler, Gabriele Marini, Vassilis Kostakos, and Jorge Goncalves. 2020. "Hi! I Am the Crowd Tasker" Crowdsourcing through Digital Voice Assistants. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376320

[36] Danula Hettiachchi, Niels van Berkel, Simo Hosio, Vassilis Kostakos, and Jorge Goncalves. 2019. Effect of Cognitive Abilities on Crowdsourcing Task Performance. In *Human-Computer Interaction – INTERACT 2019*. Springer International Publishing, Cham, 442–464. https://doi.org/10.1007/978-3-030-29381-9_28

[37] Chien-Ju Ho, Shahin Jabbari, and Jennifer Wortman Vaughan. 2013. Adaptive Task Assignment for Crowdsourced Classification. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 28)*. PMLR, Atlanta, Georgia, USA, 534–542. http://proceedings.mlr.press/v28/ho13.html

[38] Chien-Ju Ho and Jennifer Wortman Vaughan. 2012. Online Task Assignment in Crowdsourcing Markets. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* (Toronto, Ontario, Canada) *(AAAI'12)*. AAAI Press, Palo Alto, California, USA, 45–51.

[39] Bernhard Hommel. 2011. The Simon effect as tool and heuristic. *Acta Psychologica* 136, 2 (2011), 189–202. https://doi.org/10.1016/j.actpsy.2010.04.011 Responding to the Source of Stimulation: J. Richard Simon and the Simon Effect.

[40] Jeff Howe. 2008. *Crowdsourcing : why the power of the crowd is driving the future of business* (1st ed.). Crown Business, New York, NY, USA.

[41] Ayush Jain, Akash Das Sarma, Aditya Parameswaran, and Jennifer Widom. 2017. Understanding Workers, Developing Effective Tasks, and Enhancing Marketplace Dynamics: A Study of a Large Crowdsourcing Marketplace. *Proc. VLDB Endow.* 10, 7 (2017), 829–840. https://doi.org/10.14778/3067421.3067431

[42] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2011. Worker Types and Personality Traits in Crowdsourcing Relevance Labels. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (Glasgow, Scotland, UK) *(CIKM '11)*. ACM, New York, NY, USA, 1941–1944. https://doi.org/10.1145/2063576.2063860

[43] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2012. The Face of Quality in Crowdsourcing Relevance Labels: Demographics, Personality and Labeling Accuracy. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (Maui, Hawaii, USA) *(CIKM '12)*. ACM, New York, NY, USA, 2583–2586. https://doi.org/10.1145/2396761.2398697

[44] Asif R. Khan and Hector Garcia-Molina. 2017. CrowdDQS: Dynamic Question Selection in Crowdsourcing Systems. In *Proceedings of the 2017 ACM International Conference on Management of Data* (Chicago, Illinois, USA) *(SIGMOD '17)*. ACM, New York, NY, USA, 1447–1462. https://doi.org/10.1145/3035918.3064055

[45] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, Texas, USA) *(CSCW '13)*. ACM, New York, NY, USA, 1301–1318. https://doi.org/10.1145/2441776.2441923

[46] Amy L. Kristof. 1996. Person-organization fit: an integrative review of its conceptualizations, measurement, and implications. *Personnel Psychology* 49, 1 (1996), 1–49. https://doi.org/10.1111/j.1744-6570.1996.tb01790.x

[47] Muriel D. Lezak, Diane B. Howieson, David W. Loring, and Jill S. Fischer. 2004. *Neuropsychological assessment*. Oxford University Press, New York, NY, USA.

[48] Xuan Liu, Meiyu Lu, Beng Chin Ooi, Yanyan Shen, Sai Wu, and Meihui Zhang. 2012. CDAS: A Crowdsourcing Data Analytics System. *Proc. VLDB Endow.* 5, 10 (June 2012), 1040–1051. https://doi.org/10.14778/2336664.2336676

[49] Ioanna Lykourentzou, Angeliki Antoniou, Yannick Naudet, and Steven P. Dow. 2016. Personality Matters: Balancing for Personality Types Leads to Better Outcomes for Crowd Teams. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. ACM, New York, NY, USA, 260–273. https://doi.org/10.1145/2818048.2819979

[50] Colin M. MacLeod. 1991. Half a Century of Research on the Stroop Effect: An Integrative Review. *Psychological Bulletin* 109, 2 (1991), 163. https://doi.org/10.1037/0033-2909.109.2.163

[51] Panagiotis Mavridis, David Gross-Amblard, and Zoltán Miklós. 2016. Using Hierarchical Skills for Optimized Task Assignment in Knowledge-Intensive Crowdsourcing. In *Proceedings of the 25th International Conference on World Wide Web* (Montreal, Québec, Canada) *(WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 843–853. https://doi.org/10.1145/2872427.2883070

[52] Kaixiang Mo, Erheng Zhong, and Qiang Yang. 2013. Cross-task Crowdsourcing. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Chicago, Illinois, USA) *(KDD '13)*. ACM, New York, NY, USA, 677–685. https://doi.org/10.1145/2487575.2487593

[53] Stephen Monsell. 2003. Task switching. *Trends in Cognitive Sciences* 7, 3 (2003), 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7

[54] Jonas Oppenlaender, Elina Kuosmanen, Jorge Goncalves, and Simo Hosio. 2019. Search Support for Exploratory Writing. In *Human-Computer Interaction – INTERACT 2019*. Springer International Publishing, Cham, 314–336.

https://doi.org/10.1007/978-3-030-29387-1_18

[55] Adrian M. Owen, Kathryn M. McMillan, Angela R. Laird, and Ed Bullmore. 2005. N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping* 25, 1 (2005), 46–59. https://doi.org/10.1002/hbm.20131

[56] Aditya G. Parameswaran, Hector Garcia-Molina, Hyunjung Park, Neoklis Polyzotis, Aditya Ramesh, and Jennifer Widom. 2012. CrowdScreen: Algorithms for Filtering Data with Humans. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (Scottsdale, Arizona, USA) *(SIGMOD '12)*. ACM, New York, NY, USA, 361–372. https://doi.org/10.1145/2213836.2213878

[57] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods* 46, 4 (01 Dec 2014), 1023–1031. https://doi.org/10.3758/s13428-013-0434-y

[58] Michael Petrides, Bessie Alivisatos, Alan C. Evans, and Ernst Meyer. 1993. Dissociation of human mid-dorsolateral from posterior dorsolateral frontal cortex in memory processing. *Proceedings of the National Academy of Sciences* 90, 3 (1993), 873–877. https://doi.org/10.1073/pnas.90.3.873

[59] T. W. Robbins, M. James, A. M. Owen, B. J. Sahakian, L. McInnes, and P. Rabbitt. 1994. Cambridge Neuropsychological Test Automated Battery (CANTAB): A Factor Analytic Study of a Large Sample of Normal Elderly Volunteers. *Dementia and Geriatric Cognitive Disorders* 5, 5 (1994), 266–281. https://doi.org/10.1159/000106735

[60] Jeffrey M. Rzeszotarski and Aniket Kittur. 2011. Instrumenting the Crowd: Using Implicit Behavioral Measures to Predict Task Performance. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) *(UIST '11)*. ACM, New York, NY, USA, 13–22. https://doi.org/10.1145/2047196.2047199

[61] Morteza Saberi, Omar K. Hussain, and Elizabeth Chang. 2017. An Online Statistical Quality Control Framework for Performance Management in Crowdsourcing. In *Proceedings of the International Conference on Web Intelligence* (Leipzig, Germany) *(WI '17)*. ACM, New York, NY, USA, 476–482. https://doi.org/10.1145/3106426.3106436

[62] Frank L. Schmidt and John Hunter. 2004. General mental ability in the world of work: occupational attainment and job performance. *Journal of personality and social psychology* 86, 1 (2004), 162. https://doi.org/10.1037/a0012842

[63] Aaron D. Shaw, John J. Horton, and Daniel L. Chen. 2011. Designing Incentives for Inexpert Human Raters. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work* (Hangzhou, China) *(CSCW '11)*. ACM, New York, NY, USA, 275–284. https://doi.org/10.1145/1958824.1958865

[64] George Washington. 1766. George Washington Papers, Series 5, Financial Papers: Copybook of Invoices and Letters, 1754-1766. https://www.loc.gov/item/mgw500003

[65] Richard F. West, Maggie E. Toplak, and Keith E. Stanovich. 2008. Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology* 100, 4 (2008), 930. https://doi.org/10.1037/a0012842

[66] Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. 2015. Taskrec: A task recommendation framework in crowdsourcing systems. *Neural Processing Letters* 41, 2 (2015), 223–238. https://doi.org/10.1007/s11063-014-9343-z

[67] Yudian Zheng, Jiannan Wang, Guoliang Li, Reynold Cheng, and Jianhua Feng. 2015. QASCA: A Quality-Aware Task Assignment System for Crowdsourcing Applications. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (Melbourne, Victoria, Australia) *(SIGMOD '15)*. ACM, New York, NY, USA, 1031–1046. https://doi.org/10.1145/2723372.2749430

[68] Mengdie Zhuang and Ujwal Gadiraju. 2019. In What Mood Are You Today? An Analysis of Crowd Workers' Mood, Performance and Engagement. In *Proceedings of the 10th ACM Conference on Web Science* (Boston, Massachusetts, USA) *(WebSci '19)*. Association for Computing Machinery, New York, NY, USA, 373–382. https://doi.org/10.1145/3292522.3326010

Received January 2020; revised June 2020; accepted July 2020

Chapter 6

Crowd Worker Context and Cross-Device Task Acceptance

6.1 Introduction

Another promising crowd worker attribute for worker performance estimation is worker context. Like cognitive ability, we can objectively infer worker context through background information collected by their devices, and then use this information for task assignment. For instance, a worker might prefer and be able to more accurately complete a classification task when crowdsourcing through a mobile device, whereas a bounding box task is more challenging to be completed on such a device. As a first step towards cross-device crowd task assignment, in this study we aim to understand whether workers are willing to accept crowdsourcing tasks presented on different devices when their work context changes.

To this end, we ran a crowdsourcing study where we presented workers with a hypothetical scenario that describes a crowdsourcing task, including several different task and contextual attributes. Contextual attributes described the scenario using the time of the day, approximate location (e.g., at their primary workstation at home, commuting), social context and type of device. Task attributes included task type, the total number of questions available, reward, and allocated time.

Our results show that different contextual factors influence workers decision to accept or reject tasks when given the option to complete tasks on varying work devices. Further, our qualitative findings highlight that while workers mainly prefer to work at their primary workstation, they are willing to accept tasks on alternative devices in certain scenarios. Our findings pave the way to develop effective worker context-based task assignment methods of cross-device crowdsourcing. The attached publication Article III provides more details regarding this study. Our findings led us to explore the feasibility of using voice interaction for conducting crowd work, which we discuss in Chapter 8.

6.2 Article III

Hettiachchi, D., Wijenayake, S., Hosio, S., Kostakos, V., Goncalves, J. (2020). How Context Influences Cross-Device Task Acceptance in Crowd Work. In Proceedings of the Eighth AAAI Conference on Human Computation and Crowdsourcing (pp. 53–62). AAAI Press. https://ojs.aaai.org/index.php/HCOMP/article/view/7463

Ethics ID: 2056409, The University of Melbourne Human Ethics Advisory Group.

Proceedings of the Eighth AAAI Conference on Human Computation and Crowdsourcing (HCOMP-20)

How Context Influences Cross-Device Task Acceptance in Crowd Work

Danula Hettiachchi,[1] **Senuri Wijenayake,**[1] **Simo Hosio,**[2]
Vassilis Kostakos,[1] **Jorge Goncalves**[1]
[1]The University of Melbourne, [2]University of Oulu
first.last@unimelb.edu.au, simo.hosio@oulu.fi

Abstract

Although crowd work is typically completed through desktop or laptop computers by workers at their home, literature has shown that crowdsourcing is feasible through a wide array of computing devices, including smartphones and digital voice assistants. An integrated crowdsourcing platform that operates across multiple devices could provide greater flexibility to workers, but there is little understanding of crowd workers' perceptions on uptaking crowd tasks across multiple contexts through such devices. Using a crowdsourcing survey task, we investigate workers' willingness to accept different types of crowd tasks presented on three device types in different scenarios of varying location, time and social context. Through analysis of over 25,000 responses received from 329 crowd workers on Amazon Mechanical Turk, we show that when tasks are presented on different devices, the task acceptance rate is 80.5% on personal computers, 77.3% on smartphones and 70.7% on digital voice assistants. Our results also show how different contextual factors such as location, social context and time influence workers decision to accept a task on a given device. Our findings provide important insights towards the development of effective task assignment mechanisms for cross-device crowd platforms.

Introduction

Information workers have used stationary desktop or laptop computers as their primary work tool for decades. A similar trend can be seen in crowd work, with workers typically completing tasks from home and mainly using a desktop workstation (Williams et al. 2019). However, with the advancements in wireless internet technologies and widespread availability of more sophisticated and powerful mobile computing devices (e.g., digital voice assistants, smartphones), digital workers now have more flexibility than ever before to work in different contexts.

Research has shown that crowdsourcing is increasingly conducted via non-traditional devices, such as voice-interaction through smartphones (Vashistha, Sethi, and Anderson 2017), digital voice assistants including smart speakers (Hettiachchi et al. 2020a), situated touch-screen displays (Hosio et al. 2014; Goncalves et al. 2013), as well as low-cost phones (Vashistha, Garg, and Anderson 2019). Given the wide range of crowdsourcing interfaces, workers have the flexibility to complete crowd tasks in a variety of different contexts (Hettiachchi et al. 2020a). These platforms engender additional benefits, such as improved accessibility of crowdsourcing marketplaces for workers with visual impairments (Vashistha, Sethi, and Anderson 2018) (e.g., via voice interaction) and low-income (Vashistha, Garg, and Anderson 2019) (e.g., via the use of low-cost phones or situated touch-screens) to engage in crowd work.

Although crowd work is feasible through many devices, current commercial platforms are primarily built for desktop/laptop access. An integrated crowdsourcing platform that is accessible via different devices, like smartphones and digital voice assistants, has potential for offering various benefits to workers. However, it remains unclear whether – given a choice – crowd workers would be willing to complete different types of tasks on devices other than desktop or laptop computers, particularly when considering different contexts.

Thus, in this study we aim to better understand how workers decide which type of device to use, and particularly how context affects this decision. Through a Human Intelligent Task (HIT) deployed in Amazon Mechanical Turk (MTurk)[1], we collected 25,920 responses from 329 unique crowd workers. Our results indicate that task parameters (*e.g.,* HIT time estimation, available HIT count) and contextual factors (*e.g.,* approximate location, social context) play an important role on workers' decisions to accept or reject tasks. Our findings inform the creation of integrated crowdsourcing platforms and effective cross-device task assignment mechanisms that can increase overall crowd data quality and worker satisfaction.

Related Work

Impact of Worker Context

Data quality in crowdsourcing is an important research avenue that has been critical to the widespread adoption of crowdsourcing in both academic and commercial applications. While there are many different data quality enhance-

[1]https://www.mturk.com

ment techniques, the majority of them are centred around matching tasks with workers, improving task design and workflow, or aggregating answers from the crowd (Daniel et al. 2018; Kittur et al. 2013).

Previous work has proposed many different task matching or assignment strategies that capitalise on different factors, such as worker characteristics (e.g., personality (Kazai, Kamps, and Milic-Frayling 2012), skills (Mavridis, Gross-Amblard, and Miklós 2016), cognitive ability (Hettiachchi et al. 2019; Goncalves et al. 2017; Hettiachchi et al. 2020b)) and behavioural traces (Gadiraju et al. 2019; Goyal et al. 2018). However, there is far less research investigating the impact of contextual factors related to the crowd worker's environment. Such contextual factors are of particular importance when the goal is to achieve task assignment or recommendation in a crowdsourcing platform that can be accessed through different types of devices.

For example, Ikeda and Hoashi (2017) show that worker busyness and presence of a companion can impact task acceptance in mobile crowdsourcing. On a related note, as tasks in spatial crowdsourcing are directly related to a specific location, they are typically accessed through smartphones and contextual information plays an important role in task assignment (Gummidi, Xie, and Pedersen 2019). Similarly, contextual factors such as worker location, device sensing capabilities and battery level are critical in crowd sensing applications (Hassani, Haghighi, and Jayaraman 2015).

Devices for Crowd Work

Several recent studies have explored the characteristics of worker devices and their impact on task performance. Gadiraju et al. (2017) investigated the effect of the work environment on micro-task performance in CrowdFlower. The study which involves workers from the US and India shows that factors like screen resolution and device speed can have an impact on the task completion time. In a study investigating the work-life of crowd workers of MTurk, Williams et al. (2019) report that the number of monitors of the primary work computer is the most productivity defining attribute related to the workspace.

Although micro-task crowdsourcing has been traditionally limited to web interfaces accessed via desktop/laptop computers, crowd workers increasingly use smartphones to complete tasks (Chi, Batra, and Hsu 2018; Chatzimilioudis et al. 2012). Also, recent work has shown the possibility of using a wide variety of devices for crowdsourcing. Crowd work is possible through digital voice assistants through smart speakers (Hettiachchi et al. 2020a), basic phones (Vashistha, Garg, and Anderson 2019), situated touch-screen displays (Hosio et al. 2014; Goncalves et al. 2013) as well as wearable devices like smartwatches (Acer et al. 2019). Hettiachchi et al. (2020a) present a voice-based crowdsourcing platform that works through a digital voice assistant. Results of their lab study show that task accuracy for native English speakers in voice-interaction is similar to the screen-interfaces across five different common crowdsourcing tasks. Vashistha, Garg, and Anderson (2019) use interactive-voice-response (IVR) in basic phones to crowdsource speech transcription tasks. Their application is targeted at economically disadvantaged crowd workers and provides means to engage in crowd work with minimum resources.

While connected crowd platforms that can operate through many devices can be beneficial to crowd workers, there is no work that sheds light on worker perceptions of when tasks are presented and possible to complete on multiple devices.

Task Search and Acceptance

Crowdsourcing marketplaces typically expose a list of crowd tasks to workers from which they have to choose and accept to work on. While the aim is to provide greater autonomy and agency to workers, searching for a suitable task has become increasingly difficult and time consuming for workers (Chilton et al. 2010). Also, searching for optimal tasks is perceived as unpaid work for crowd workers (Hara et al. 2018).

There are many tools that can help workers find suitable tasks (Kaplan et al. 2018; Williams et al. 2019). For example, Turkopticon is one of the most widely adopted browser extensions that is used to evaluate and review requesters and HITs (Irani and Silberman 2013). Similar tools have been proposed to estimate the time that is needed to complete a task (Saito et al. 2019). However, such tools are limited to web-interfaces and are not always available in other devices, such as smartphones or smart speakers.

On the other hand, task search times can be much longer when interacting with devices such as smartphones and smart speakers when compared to desktop or laptop computers (Hettiachchi et al. 2020a). In smartphones, the amount of information that a worker can obtain at any given time is limited in smartphones due to the screen size. Similarly, voice-interaction limits the amount of information presented on smart speakers (Hettiachchi et al. 2020a).

Therefore, appropriate task assignment and recommendation is quite important for a cross-device crowd platform, especially when workers request tasks through smartphones or smart speakers. In this study, we take the initial steps to understand cross-device task acceptance, which is essential to create an effective task assignment model that can increase the overall data quality and worker satisfaction.

Study

Our study consists of two main components deployed on MTurk. First, we deployed the main task, where workers marked their stated preference in accepting tasks. Second, we invited workers to complete two different surveys, depending on the number of completed HITs.

Main Task

To understand workers' preferences in accepting tasks on different devices in various scenarios, we constructed a simple task. As shown in Figure 1, in each HIT, we presented workers with a list of parameters related to a hypothetical task (HIT). These parameters include task characteristics, such as Task Name, number of HITs available as well as

contextual parameters, such as workers' approximate location, device and time of the day. Workers were asked to carefully examine the parameters and decide if they would accept and start working on the task. We clarified that they would not be asked to actually complete the given hypothetical task. In addition to the binary response of either accepting or rejecting the given task, workers were asked to indicate through a series of range sliders the extent to which certain factors influenced their decision. Following the design guidelines proposed in the literature, the range sliders had no tick marks in the axis and dynamically displayed the value to users as they move the marker (Matejka et al. 2016; Hosio et al. 2018). As shown in Figure 1, we listed five factors: Location, Device, Time, Social Context and Task details.

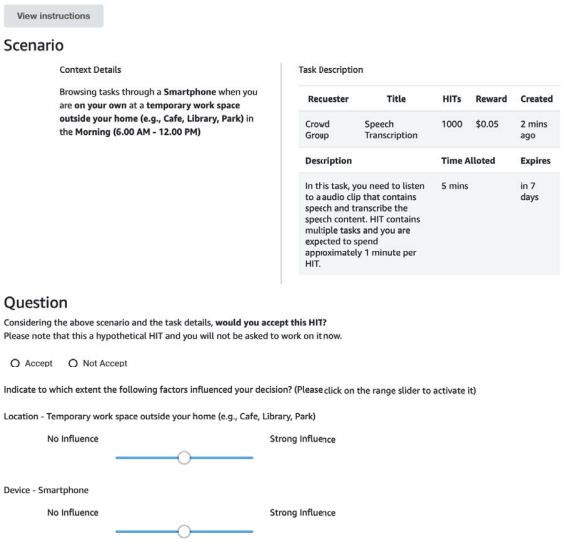

Figure 1: A portion of the HIT which shows the presentation of task parameters and the questions.

Tasks were selected based on typical tasks that are available on crowdsourcing platforms (Difallah et al. 2015) and task taxonomies purposed in the literature (Gadiraju, Kawase, and Dietze 2014). We also ensured that there is an equal number of text-based, audio-based and image-based tasks. As workers consider reward a key factor for accepting tasks (Hara et al. 2018), we kept the reward proportional to the expected time to complete the task. We provided workers with the maximum task time through the 'Time Allotted' parameter. A more realistic estimate of the actual time to complete the HIT was given in the task description. HIT count was set at 1, 10 and 1000 based on common values prevalent in typical marketplaces (Difallah et al. 2015). Requester name, HIT created time, and HIT expiration time were consistent across all HITs.

We created 5,184 HITs by using all possible combinations of the parameters listed in Table 1 and collected five responses per HIT. We set out three pre-qualification requirements for workers. Based on the qualifications, all our workers were from the US and have completed more than 1000 tasks with an approval rate of 95% or higher.

Table 1: Task and Context parameters

Parameter	**Values**
Task Type	Sentiment Analysis Information Finding Audio Tagging Speech Transcription Image Classification Bounding Box
HITs	1, 10 or 1000
Reward	$ 0.01,$ 0.05, or $ 0.50
Created	2 mins ago
Time Allotted	1 min, 5 mins, or 10 mins
Expires	in 7 days
Time of the day	Morning (6.00 AM - 12.00 PM) Afternoon (12.00 PM - 6.00 PM) Evening (6.00 PM - 12.00 AM) Night (12.00 AM - 6.00 AM)
Approximate Location	at home (at your primary workstation) at home (other space different to your primary workstation) at a temporary work space (e.g., Cafe, Library, Park) commuting (e.g, in a Car or Train)
Social Context	on your own with family/friends
Device	Desktop/Laptop Smartphone Smart speaker

Surveys

All workers who completed at least one HIT in our main task were asked to complete a demographic survey. The survey included questions about workers' preferred time to conduct crowd work, the average time they spend on crowd work, crowd work income (as a percentage of total income) and whether they have used voice assistants in general. We also captured basic demographic information such as age, gender, and primary internet device.

Furthermore, we invited workers who completed more than 20 HITs in our main task to complete an additional follow-up survey. We asked workers to provide detailed answers with examples of how different task characteristics and contextual factors impact their task acceptance based on previous crowd work experience. We also queried which factors they would consider if crowdsourcing platforms are available through multiple devices. We further inquired on their preference for task assignment and task recommendation on standard crowd market places as well as on different devices. Workers received USD $1.00 for the completion of this survey.

Parameter	Estimate	Std. Error	Z value	
Intercept	2.00	0.20	9.92	***
Task - Information Finding	−0.33	0.07	−4.75	***
Task - Audio Tagging	−0.54	0.07	−7.86	***
Task - Speech Transcription	−0.73	0.07	−10.58	***
Task - Image Classification	0.06	0.07	0.91	
Task - Image Bounding box	−0.58	0.07	−8.45	***
Time Allotted	0.04	0.01	8.11	***
HIT Count	0.04	0.02	2.63	**
Device - Smartphone	−0.86	0.14	−6.04	***
Device - Smart speaker	−1.46	0.14	−10.32	***
Time - Afternoon (12.00 PM - 6.00 PM)	−0.09	0.10	−0.88	
Time - Evening (6.00 PM - 12.00AM)	−0.08	0.10	−0.73	
Time - Night (12.00 AM - 6.00 AM)	−0.48	0.10	−4.81	***
Social Context - with your family/friends	−0.73	0.07	−10.41	***
Location - at home (other space different to your primary workstation)	−0.17	0.10	−1.67	
Location - at a temporary work space (e.g., Cafe, Library, Park)	−0.64	0.10	−6.25	***
Location - commuting (e.g, in a Car or Train)	−0.99	0.10	−9.92	***

Table 2: Fixed effects of Generalised Linear Mixed Model. Significance '***' $p < 0.001$, '**' $p < 0.01$, '*' $p < 0.05$

Parameter	Estimate	Std. Error	Z value	
Device - Smartphone: Time - Afternoon (12.00 PM - 6.00 PM)	0.13	0.14	0.94	
Device - Smart speaker: Time - Afternoon (12.00 PM - 6.00 PM)	0.07	0.14	0.55	
Device - Smartphone: Time - Evening (6.00 PM - 12.00AM)	0.17	0.14	1.21	
Device - Smart speaker: Time - Evening (6.00 PM - 12.00AM)	−0.08	0.14	−0.55	
Device - Smartphone: Time - Night (12.00 AM - 6.00 AM)	0.06	0.14	0.46	
Device - Smart speaker: Time - Night (12.00 AM - 6.00 AM)	0.13	0.14	0.96	
Device - Smartphone: Social Context - with your family/friends	0.41	0.10	4.27	***
Device - Smart speaker: Social Context - with your family/friends	0.46	0.10	4.81	***
Device - Smartphone: Location - at home (other space)	0.18	0.14	1.28	
Device - Smart speaker: Location - at home (other space)	0.28	0.14	2.02	*
Device - Smartphone: Location - at a temporary work space	0.44	0.14	3.13	**
Device - Smart speaker: Location - at a temporary work space	0.16	0.14	1.14	
Device - Smartphone: Location - commuting (e.g, in a Car or Train)	0.40	0.14	2.91	**
Device - Smart speaker: Location - commuting (e.g, in a Car or Train)	0.44	0.14	3.23	**

Table 3: Interactions of Generalised Linear Mixed Model

Results

We collected 25,920 responses for our main task with a total of 329 workers contributing to the task. Each worker completed 78.8 tasks on average and spent 51.3 seconds on each single response on average.

Worker Demographics

60 (18.2%) out of 329 workers completed the demographics survey. The number of answers provided by this subset of workers accounts for 22.0% of the total responses in the main task.

We present an estimation of the worker demographics based on the collected survey responses. Based on self-reported gender, 33 women and 27 men answered the survey with an average age of 38.6 (SD = 10.9) years. Workers reported spending an average of 22.3 hours per week on the Mechanical Turk platform with a majority (86.7%) of them working on Mechanical Turk during both weekdays and weekends. Workers stated earning on average 41.3% of their monthly income from crowd work. Furthermore, 15 workers stated that 90% or more of their monthly income comes from crowd work.

The majority of workers (98.3%) reported that they use a desktop computer or a laptop computer as their primary internet device to complete crowd tasks. Only one worker stated that they use an iPad as their primary device for crowd work. 58.8% of the workers reported to have previously used the mobile version of Mechanical Turk, whereas 62.7% of the workers have used a digital voice assistant. When inquired about the locations from where they complete crowd work, workers mainly mentioned workstation at home, bedroom at home, and living room at home as their primary work location.

Task Acceptance

In response to our primary question given in the main task, workers decided to accept the given HIT in 19,759 (76.2%) of the cases. To investigate the impact of task and contex-

tual parameters on task acceptance, we fitted a binomial generalised linear mixed model with maximum likelihood (Laplace Approximation) using the R-package lme4 (Bates et al. 2015). We included all the parameters listed in Table 1 and interactions between the device and contextual factors (Time, Location and Social Context). Worker ID, which is unique for each worker, was included as a random effect. Our results indicate significant fixed effects both in terms of the task parameters and contextual factors, and are detailed in the Table 2.

Impact of Task Parameters

The results indicate strong fixed effects in terms of the task type, time estimation and number of HITs available. In Figure 2, we observe that workers prefer tasks that have an estimated completion time of 1 minute as opposed to very short (10 second) or long (10 minute) HITs. This preference is evident across all devices. However, we note that workers are more reluctant to accept long (10 minute) HITs in smart speakers when compared to other devices.

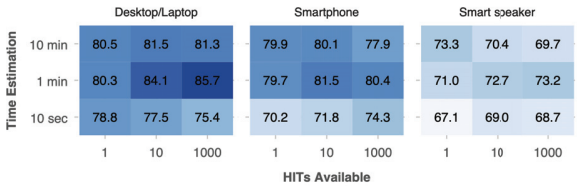

Figure 2: Task Acceptance rate on different devices across varying time estimations and number of HITs available.

As shown in Figure 3, the task acceptance rate also varied by task type, but did not exhibit major variations across devices.

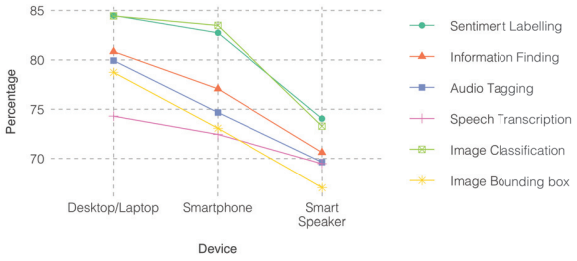

Figure 3: Task Acceptance rate on different devices across task types.

Impact of Contextual Factors

Our results suggest that approximate location, social context, and device, influenced workers' decision to either accept or reject a given task.

Approximate Location When considering tasks presented on Desktop or Laptop computers, results indicate the highest acceptance rate at their primary workstation at home. However, when we examine task acceptance rates on smartphones and smart speakers, the acceptance rate is higher when the workers are at a space within their home different to their primary workstation as compared to the rest of the locations.

Social Context With regard to social context, workers are more likely to accept a task when they are on their own (78.3%) as compared to a situation where they are accompanied by family or friends (74.1%). As seen in Figure 4 (middle), this effect is consistent across devices.

Time Time of the day did not have a significant impact on workers choice except that workers preferred Morning (78.0%), Afternoon (76.5%) and Evening (77.4%) when compared to Night (73.0%). Similarly, as shown in Figure 4 (bottom), we did not find any meaningful interaction effect between Time and the Type of Device.

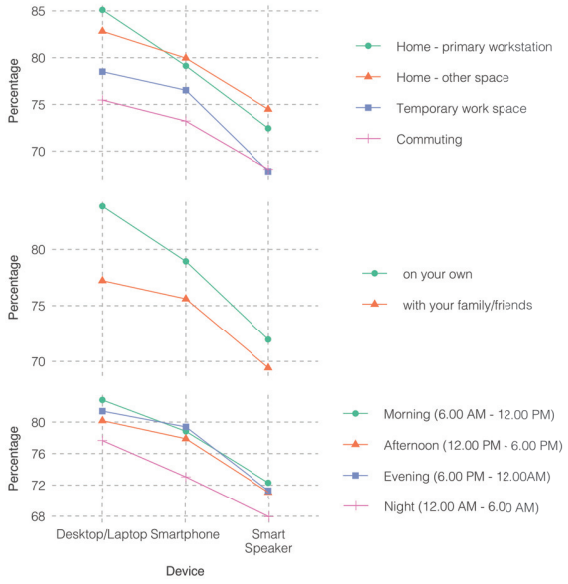

Figure 4: Task acceptance rate across approximate locations (top), social contexts (middle), and time intervals (bottom) on different devices.

Worker Responses on Contextual Factors In addition to the binary decision to accept the task, we asked workers to indicate which factors influenced their decision. Figure 5 summarises worker responses. From the response mean values and distributions, we observe that all task and contextual factors influenced the decision when accepting or rejecting given tasks. Task parameters were identified as being slightly more important than contextual factors when rejecting tasks. Also, response distributions (bi-modal distributions in Not Accept and normal distributions in Accept) indicate that workers were more decisive on factors when they did not accept tasks when compared to the cases where they accepted tasks.

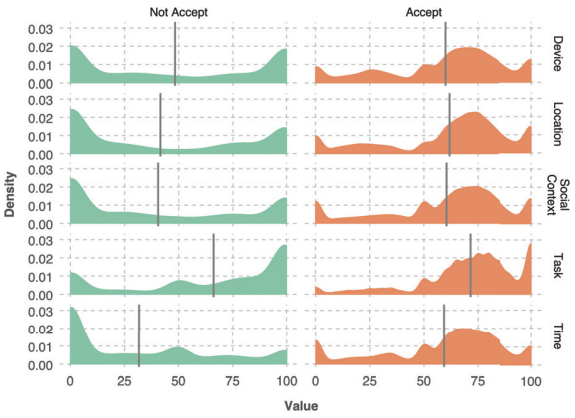

Figure 5: Reported importance of different parameters when deciding whether to accept the given task. The vertical line indicates the mean in each group.

Follow-up Survey

From 94 invited workers who completed more than 20 HITs in the main task, we received a total of 30 responses to the follow-up survey. Two of the paper's authors individually applied a deductive thematic analysis (Braun and Clarke 2006) to the eligible responses based on the paper's research objectives. Following this, the authors met to discuss their outcomes. Next, we present the main findings of this analysis.

Impact of Contextual Factors on Task Acceptance We set out to investigate how contextual factors such as the social context, approximate location, device type, and time of the day impact task acceptance among workers.

First, the social context of workers emerged as a crucial determinant of task acceptance. Participants highlighted how they prefer to work on tasks when they are alone and can adequately concentrate on the task, and would not accept tasks that are audio-based or require higher concentration, when with family or friends;

> "*If I'm alone I will attempt just about any task. However if there are people around, I typically tend to stick to less involved tasks that don't require much concentration, especially those tasks that require listening to audio because it often becomes hard to hear the audio.*" (P8)

Moreover, participants were most likely to work on HITs from their primary workstation at home due to its comfortable and stable setup. However, some participants also note how their location may influence their device preference and the type of tasks they would attempt;

> "*When I have long waits e.g., at doctor's office I will do quick surveys or batch HITS on my phone. So I guess where I am determines the device.*" (P2)

The majority of our participants stressed their preference to complete tasks using a desktop or a laptop computer, over other devices (*e.g., smartphones*), as they offer larger displays and other controls (*e.g., keyboard and mouse*) requiring lesser effort to complete tasks;

> "*I would not do a task if it was not offered on a laptop. The laptop is the best device to use because of the decent size screen and the use of the keyboard and mouse. With a laptop, I have easy access with the click of a mouse and I can use my keyboard to complete some tasks. All the other devices would tire me out faster.*" (P10)

Moreover, if the task was available across multiple devices, participants would consider the compatibility of the device that they are using at the time and the task at hand, when deciding whether to accept a task or not;

> "*[Task acceptance] would depend on what device I'm using currently and how easy or difficult it would be complete that job on that device or if it would be better to switch. Some jobs require a larger screen so if I'm using my phone (rarely) I would want to do that job on a tablet or my Chromebook for example.*" (P7)

Impact of Task Characteristics on Task Acceptance We also investigated the impact of task characteristics such as HIT count, reward and requester profile on how workers determine whether or not to accept a task.

We note a significant preference among workers for tasks with a substantial HIT count (in thousands), each HIT requiring a small time period to complete;

> "*I love to complete hits with large hit counts that are fast and easy to complete, allowing me to sit and focus on them for extended periods of time.*" (P1)

Participants explained that completing simple and repetitive tasks allow them to stay focused for a long period of time at once, thereby maximising their earnings per hour;

> "*I love to complete hits with large hit counts that are fast and easy to complete, allowing me to sit and focus on them for extended periods of time.*" (P1)

Furthermore, participants also emphasised how the reward allocated for a task could impact task acceptance. In general, participants were keen on maintaining an acceptable hourly earning and therefore would calculate reward/estimated task time prior to accepting a task.

Moreover, despite general reluctance from participants to accept tasks requiring a device other than a desktop or a laptop, we note that a substantial reward could encourage them to do so;

> "*I love to complete hits with large hit counts that are fast and easy to complete, allowing me to sit and focus on them for extended periods of time.*" (P1)

Additionally, participants were seen to consider the requester profile - especially their approval rate and average pay time - when considering tasks for acceptance;

> "*I steer clear of requesters whose approval percentage is below 90 or whose average pay time is more than 3-4 days. These are signals that my HITs probably won't be approved or paid out, which negatively impacts my worker profile.*" (P8)

Participants also explained how they tend to "only do a few HITs from a requester that has an approval rate under 95% to see if they are approved first" (P12) as a precaution in such cases.

Preference for Task Recommendation and Assignment We note both supportive and critical opinions from participants regarding task recommendation and assignment in standard crowd market places. Participants supportive of this notion explained how task recommendations could connect each worker with tasks compatible with their personal skill set, reducing the amount of time they otherwise spend on searching compatible tasks;

> "*It would be nice just to have that option. Viewing tasks that are compatible with me would streamline the amount of tasks I complete because I spend a lot of time searching for HITs to complete.*" (P10)

They also suggested that task recommendations could be based on workers' ratings on completed tasks in addition to tasks they have completed successfully so that recommendations would include a mix of tasks they enjoy as well as tasks they are competent at;

> "*I do think it's nice to be able to sent certain tasks if there was a way to be sure that the recommendation system worked off of something like ratings from the users as opposed to just the history of tasks worked. If it were truly able to work that way then it would make it much easier to be able to jump on a task I enjoy and not have it taken by someone else who may or may not enjoy doing it so that would be a bonus for both the requester and the worker.*" (P14)

Moreover, participants highlighted how task recommendations could be especially important when tasks are offered across multiple devices. For instance, they note how recommending tasks that are compatible with the device currently in use could be valuable;

> "*I would prefer task assignment. I like doing tasks that are compatible to the device at use. There's no point in trying to complete a task that isn't presented on your device in a way that makes it easy for you to complete it.*" (P10)

P3 also commented on how device-based task recommendations could "ensure the task is administered in the most efficient manner, using the most compatible device", resulting in higher quality responses. Participants also emphasised that the opportunity to specify devices that they prefer not to use could "help the right workers get the right HITs and stop so many of the HITs from being picked up and returned" (P14), which would also prevent workers getting frustrated due to device – task incompatibilities.

Alternatively, some participants were critical of task assignment and recommendation as they were sceptic of the platform's ability to account for their personal preferences and other contextual factors;

> "*I definitely would not think [task recommendation] is helpful and in fact it would annoy me. I like to search and scroll through the tasks so I can evaluate them on my own judgement. A platform is simply an AI and it doesn't know any of my other factors, like how much time I have left in a day to complete HITs, what kind of HITs I want to complete today, and what my financial quota is for those HITs.*" (P5)

Some others also emphasised that while task recommendation could assist them find work faster, task assignment would be detrimental to their sense of agency;

> "*I would not like to be assigned work because the whole point of doing MTurk, for me, is functionally being my own boss.*" (P7)

Participants were also concerned that task recommendation may limit them to only certain types of HITs (based on their working history), restricting their opportunities to attempt new and interesting tasks in future.

Discussion

Crowd Work Devices

Our qualitative results indicate that crowd workers prefer workstations with desktop or laptop computers, mainly due to usability factors, such as large screen area and familiarity with keyboard-mouse setup. This preference is also evident in Figure 4 (top) through the high rate (85.1%) of task acceptance in desktop/laptop devices when workers are at their primary workstation. This is also in line with the findings of literature that investigate crowd worker preferences (Williams et al. 2019).

We also note that workers were willing to accept 77.3% of the given tasks on smartphones and 70.7% of the tasks on smart speakers as compared to 80.5% on desktop/laptop computers. This receptiveness towards alternative devices in the proposed scenarios shows promising signs with regard to the feasibility of cross-device crowd platforms that involve voice-interaction (Vashistha, Sethi, and Anderson 2017; Hettiachchi et al. 2020a).

Absence of different work tools (*e.g.,* browser extensions that filter tasks (Kaplan et al. 2018)) can make other devices less desirable for crowd workers. Similarly, as observed in our results, complex image related tasks such as image bounding box are less desired on smartphones and smart speakers, due to limited screen-size and restricted interaction options. This is also evident in our results, where workers stated that they are concerned about how easy it would be to complete the task on a device of interest. Therefore, when making tasks available through different devices, it is important to validate if the interaction style (*i.e.,* touch interaction in smartphones and voice interaction in smart speakers) and device capabilities are compatible with the task.

Worker Context and Task Acceptance

In this work, we explore how contextual factors impact task acceptance in a cross-device scenario. Approximate location and social context appear to be particularly important for workers. When closely examining the task acceptance rates (Figure 4 (top)), for both smartphones and smart speakers, the acceptance rate is higher when at other spaces at home than when at the primary workstation. While extracting the

specific worker location is not recommended as it leads to privacy concerns (To, Ghinita, and Shahabi 2014), we show that approximate location is a reasonable alternative that influences task selection on a cross-device platform. We also observe that time of the day is generally not a primary concern for workers except that, unsurprisingly, the task acceptance rate is much lower during the night (12.00AM - 6.00AM).

While mobile crowd work is common (Chi, Batra, and Hsu 2018), an estimated over 40% of our workers have never used the mobile version of MTurk platform. On the other hand, voice-based crowd platforms are not yet commercially available (Hettiachchi et al. 2020a). This limited understanding and exposure to voice-based and other alternative crowdsourcing platforms can be a reason behind less pronounced interaction effects concerning contextual factors and devices.

Our results also indicate that parameters specific to the HIT such as task type, the number of HITs available, and task time estimation, still play a vital role in task selection. Our findings are in line with the crowdsourcing literature (Daniel et al. 2018; Martin et al. 2014) and further confirms that such relationships extend into various crowd work scenarios.

Integrated Cross-Device Crowdsourcing Platforms

Audio related tasks, like audio annotation and speech transcription, are common in current crowdsourcing marketplaces (Difallah et al. 2015) which has led to an increased exploration of crowdsourcing via voice-interaction (Vashistha, Sethi, and Anderson 2017; Hettiachchi et al. 2020a). On the other hand, smartphones and other mobile computing devices are capable of handling performance intensive tasks and are suitable for sustained work (Chi, Batra, and Hsu 2018).

Crowdsourcing platforms have seen an increase in the number of tasks related to mobile apps. Research has also shown that there is potential to use crowdsourcing for tasks that extend beyond screen-based devices, such as virtual reality experiments or application testing (Ma et al. 2018). Some platforms, such as *Prolific*, even allow mobile app installs as part of the assigned tasks. However, our qualitative results highlight that crowd workers find it inconvenient to switch between devices to complete a task. By allowing workers to browse tasks, accept and work on different devices, a cross-platform crowd marketplace can mitigate the required effort to switch between devices and create a positive crowd work experience for workers.

Literature also reports that workers exhibit multi-tasking behaviour and engage in other tasks like watching TV and chatting online while completing crowd work (Chandler, Mueller, and Paolacci 2014). In fact, some workers prefer to multi-task even though it is not always desired by task requesters (Lascau et al. 2019). Working on devices like smartphones and smart speakers could easily allow workers to facilitate this multi-tasking work style as compared to a workstation. In addition, an always-on device like the smart speaker or a ubiquitous device like the smartphone is helpful for workers in terms of handling interruptions and working in short sessions.

Given the steady growth in crowd work population (Difallah, Filatova, and Ipeirotis 2018) and the availability of a wide array of tasks, we anticipate that crowdsourcing platforms will gradually shift towards natively supporting different types of devices. For example, the popular crowdsourcing platform Amazon MTurk is aiming to increase task compatibility on mobile devices [2] and is well-positioned to extend their platform to smart speakers in the future through the increasingly ubiquitous Amazon Alexa.

Limitations

We acknowledge several limitations of our study. First, workers who participated in our study have not experienced a fully functional voice-based crowd platform. It is possible that this lack of exposure impacted their decision to either accept or reject tasks on smart speakers. Second, our qualitative data originates from a subset of workers who took part in the main task. We invited 94 workers for the post-survey through a custom qualification in MTurk from which 30 workers completed the task. Third, we do not investigate all possible contextual factors and focus primarily on ones that have been shown to impact crowd work. Nevertheless, we tested over 5,000 unique HITs in our study, which provides a wide array of potential crowd work scenarios. Additional factors would vastly increase this number and lead to an overly complex study design.

Conclusion

In this paper, we present a study on MTurk aimed at better understanding crowd workers' preferences regarding accepting or rejecting tasks under varying contexts. Our results indicate that task acceptance rate is 80.5% on personal computers, 77.3% on smartphones and 70.7% on digital voice assistants. We also show that contextual factors such as workers approximate location and social context influence their willingness to accept tasks presented on different devices. Further, we discuss how an integrated crowdsourcing platform that operates across different types of devices can bring benefits to crowd workers by allowing for flexibility in terms of work location, convenient task initiation. Further, we argue that the findings of our work can contribute towards creating effective task assignment strategies for future cross-device crowdsourcing platforms. However, further work is needed to examine how task performance varies across devices as well as developing appropriate cross-platform task matching mechanisms.

References

Acer, U. G.; Broeck, M. v. d.; Forlivesi, C.; Heller, F.; and Kawsar, F. 2019. Scaling Crowdsourcing with Mobile Workforce: A Case Study with Belgian Postal Service. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3(2):35:1–35:32.

[2] https://blog.mturk.com/now-you-can-complete-hits-on-the-new-worker-website-6fab0da9ca80

Bates, D.; Mächler, M.; Bolker, B.; and Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1).

Braun, V., and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3(2):77–101.

Chandler, J.; Mueller, P.; and Paolacci, G. 2014. Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods* 46(1):112–130.

Chatzimilioudis, G.; Konstantinidis, A.; Laoudias, C.; and Zeinalipour-Yazti, D. 2012. Crowdsourcing with Smartphones. *IEEE Internet Computing* 16(5):36–44.

Chi, P.-Y. P.; Batra, A.; and Hsu, M. 2018. Mobile crowdsourcing in the wild. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct - MobileHCI '18*, 410–415. New York, New York, USA: ACM.

Chilton, L. B.; Horton, J. J.; Miller, R. C.; and Azenkot, S. 2010. Task search in a human computation market. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, 1–9. ACM.

Daniel, F.; Kucherbaev, P.; Cappiello, C.; Benatallah, B.; and Allahbakhsh, M. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Computing Surveys* 51(1):1–40.

Difallah, D. E.; Catasta, M.; Demartini, G.; Ipeirotis, P. G.; and Cudré-Mauroux, P. 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, 238–247. Geneva, Switzerland: IW3C2.

Difallah, D.; Filatova, E.; and Ipeirotis, P. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, 135–143. ACM.

Gadiraju, U.; Checco, A.; Gupta, N.; and Demartini, G. 2017. Modus Operandi of Crowd Workers: The Invisible Role of Microtask Work Environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1(3):1–29.

Gadiraju, U.; Demartini, G.; Kawase, R.; and Dietze, S. 2019. Crowd Anatomy Beyond the Good and Bad: Behavioral Traces for Crowd Worker Modeling and Pre-selection. *Computer Supported Cooperative Work (CSCW)* 28(5):815–841.

Gadiraju, U.; Kawase, R.; and Dietze, S. 2014. A Taxonomy of Microtasks on the Web. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, 218–223. ACM.

Goncalves, J.; Ferreira, D.; Hosio, S.; Liu, Y.; Rogstadius, J.; Kukka, H.; and Kostakos, V. 2013. Crowdsourcing on the Spot: Altruistic Use of Public Displays, Feasibility, Performance, and Behaviours. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, 753–762. ACM.

Goncalves, J.; Feldman, M.; Hu, S.; Kostakos, V.; and Bernstein, A. 2017. Task Routing and Assignment in Crowdsourcing Based on Cognitive Abilities. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, 1023–1031. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Goyal, T.; Mcdonnell, T.; Kutlu, M.; Elsayed, T.; and Lease, M. 2018. Your Behavior Signals Your Reliability: Modeling Crowd Behavioral Traces to Ensure Quality Relevance Annotations. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*, HCOMP, 41–49.

Gummidi, S. R. B.; Xie, X.; and Pedersen, T. B. 2019. A survey of spatial crowdsourcing. *ACM Transactions on Database Systems* 44(2).

Hara, K.; Adams, A.; Milland, K.; Savage, S.; Callison-Burch, C.; and Bigham, J. P. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 449:1–449:14. ACM.

Hassani, A.; Haghighi, P. D.; and Jayaraman, P. P. 2015. Context-Aware Recruitment Scheme for Opportunistic Mobile Crowdsensing. In *2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS)*, 266–273. IEEE.

Hettiachchi, D.; van Berkel, N.; Hosio, S.; Kostakos, V.; and Goncalves, J. 2019. Effect of Cognitive Abilities on Crowdsourcing Task Performance. In *Human-Computer Interaction – INTERACT 2019*, 442–464. Springer.

Hettiachchi, D.; Sarsenbayeva, Z.; Allison, F.; van Berkel, N.; Dingler, T.; Marini, G.; Kostakos, V.; and Goncalves, J. 2020a. "Hi! I am the Crowd Tasker" Crowdsourcing through Digital Voice Assistants. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20. New York, NY, USA: ACM.

Hettiachchi, D.; van Berkel, N.; Kostakos, V.; and Goncalves, J. 2020b. CrowdCog: A Cognitive Skill based System for Heterogeneous Task Assignment and Recommendation in Crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 4(CSCW2).

Hosio, S.; Goncalves, J.; Lehdonvirta, V.; Ferreira, D.; and Kostakos, V. 2014. Situated crowdsourcing using a market model. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST 2014, 55–64.

Hosio, S. J.; Karppinen, J.; Takala, E.-P.; Takatalo, J.; Goncalves, J.; van Berkel, N.; Konomi, S.; and Kostakos, V. 2018. Crowdsourcing Treatments for Low Back Pain. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, volume 2018-April, 1–12. New York, New York, USA: ACM Press.

Ikeda, K., and Hoashi, K. 2017. Crowdsourcing GO: Effect of worker situation on mobile crowdsourcing performance. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1142–1153. ACM.

Irani, L. C., and Silberman, M. S. 2013. Turkopticon: Inter-

rupting Worker Invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, 611. ACM Press.

Kaplan, T.; Saito, S.; Hara, K.; and Bigham, J. P. 2018. Striving to Earn More: A Survey of Work Strategies and Tool Use Among Crowd Workers. In *Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*. AAAI.

Kazai, G.; Kamps, J.; and Milic-Frayling, N. 2012. The face of quality in crowdsourcing relevance labels. In *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*, CIKM '12, 2583. ACM.

Kittur, A.; Nickerson, J. V.; Bernstein, M.; Gerber, E.; Shaw, A.; Zimmerman, J.; Lease, M.; and Horton, J. 2013. The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, 1301–1318. ACM.

Lascau, L.; Gould, S. J. J.; Cox, A. L.; Karmannaya, E.; and Brumby, D. P. 2019. Monotasking or multitasking: Designing for crowdworkers' preferences. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–14. New York, New York, USA: ACM.

Ma, X.; Cackett, M.; Park, L.; Chien, E.; and Naaman, M. 2018. Web-Based VR Experiments Powered by the Crowd. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 33–43. ACM Press.

Martin, D.; Hanrahan, B. V.; O'Neill, J.; and Gupta, N. 2014. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*, CSCW '14, 224–235. ACM Press.

Matejka, J.; Glueck, M.; Grossman, T.; and Fitzmaurice, G. 2016. The Effect of Visual Appearance on the Performance of Continuous Sliders and Visual Analogue Scales. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5421–5432. New York, NY, USA: ACM.

Mavridis, P.; Gross-Amblard, D.; and Miklós, Z. 2016. Using Hierarchical Skills for Optimized Task Assignment in Knowledge-Intensive Crowdsourcing. In *Proceedings of the 25th International Conference on World Wide Web - WWW '16*, 843–853. ACM Press.

Saito, S.; Nakano, T.; Chiang, C. W.; Kobayashi, T.; Savage, S.; and Bigham, J. P. 2019. TurkScanner: Predicting the hourly wage of microtasks. In *The Web Conference 2019 - Proceedings of the World Wide Web Conference*, WWW 2019, 3187–3193.

To, H.; Ghinita, G.; and Shahabi, C. 2014. Framework for protecting worker location privacy in spatial crowdsourcing. *Proceedings of the VLDB Endowment*.

Vashistha, A.; Garg, A.; and Anderson, R. 2019. ReCall: Crowdsourcing on Basic Phones to Financially Sustain Voice Forums. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 169:1–169:13. ACM.

Vashistha, A.; Sethi, P.; and Anderson, R. 2017. Respeak: A Voice-based, Crowd-powered Speech Transcription System. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, 1855–1866. ACM.

Vashistha, A.; Sethi, P.; and Anderson, R. 2018. BSpeak: An Accessible Voice-based Crowdsourcing Marketplace for Low-Income Blind People. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, CHI '18, 1–13. ACM.

Williams, A. C.; Mark, G.; Milland, K.; Lank, E.; and Law, E. 2019. The perpetual work life of crowdworkers: How tooling practices increase fragmentation in crowdwork. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW).

Chapter 7

Crowdsourcing through Digital Voice Assistants

7.1 Introduction

The findings of the study presented in Chapter 6 led to two important conclusions. First, crowd workers are willing to accept certain crowdsourcing tasks on devices that are not their regular work devices (e.g., desktop and laptop computers). Second, their willingness to accept tasks varies depending on their context, which can be potentially utilised for cross-device task assignment. In this study, we hypothesise that a voice-based crowdsourcing platform that operates through a voice assistant can provide greater flexibility to crowd workers. Particularly, voice assistants can be accessed through smart speakers that passively sit in one's home or through mobile phones.

We built 'Crowd Tasker', a voice-based crowdsourcing platform that operates through a digital voice assistant (Google Assistant). We conducted a controlled lab study to compare task performance in voice-based approach to regular screen-based crowdsourcing platforms. Our findings confirm that for native English speakers, crowdsourcing task accuracy on a voice-interface is not significantly different to task performance on a regular screen-interface. Further, we conducted a field deployment to explore how workers engage with voice-based crowdsourcing platforms in a more natural environment. Our results show that voice-based crowdsourcing can provide greater flexibility to workers by allowing them to initiate tasks quickly and easily at opportune moments and complete tasks while attending to other simple household tasks.

Finally, we discuss how voice-based crowdsourcing can complement traditional crowdsourcing platforms and provide more benefits to workers. We present design guidelines for implementing voice-based crowdsourcing platforms, and elaborate on selecting and designing voice-based crowd tasks. The attached publication (Article IV) provides additional details regarding our implementation, experimental setup, and outcomes.

7.2 Article IV

Hettiachchi, D., Sarsenbayeva, Z., Allison, F., van Berkel, N., Dingler, T., Marini, G., Kostakos, V., Goncalves, J. (2020). "Hi! I am the Crowd Tasker" Crowdsourcing through Digital Voice Assistants. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–14). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3313831.3376320

Ethics ID: 1954377, The University of Melbourne Human Ethics Advisory Group.

"Hi! I am the Crowd Tasker" Crowdsourcing through Digital Voice Assistants

Danula Hettiachchi[1], **Zhanna Sarsenbayeva**[1], **Fraser Allison**[1], **Niels van Berkel**[2], **Tilman Dingler**[1], **Gabriele Marini**[1], **Vassilis Kostakos**[1], **Jorge Goncalves**[1]
[1]The University of Melbourne, Melbourne, Australia
[2]University College London, London, United Kingdom
[1]first.last@unimelb.edu.au, [2]n.vanberkel@ucl.ac.uk

ABSTRACT
Inspired by the increasing prevalence of digital voice assistants, we demonstrate the feasibility of using voice interfaces to deploy and complete crowd tasks. We have developed *Crowd Tasker*, a novel system that delivers crowd tasks through a digital voice assistant. In a lab study, we validate our proof-of-concept and show that crowd task performance through a voice assistant is comparable to that of a web interface for voice-compatible and voice-based crowd tasks for native English speakers. We also report on a field study where participants used our system in their homes. We find that crowdsourcing through voice can provide greater flexibility to crowd workers by allowing them to work in brief sessions, enabling multi-tasking, and reducing the time and effort required to initiate tasks. We conclude by proposing a set of design guidelines for the creation of crowd tasks for voice and the development of future voice-based crowdsourcing systems.

Author Keywords
crowdsourcing; smart speakers; digital voice assistants; voice user interface

CCS Concepts
•**Human-centered computing** → **Interaction devices;** •**Information systems** → **Crowdsourcing;**

INTRODUCTION
Despite the growing popularity of digital voice assistants (such as Alexa, Siri, Google Assistant, and Cortana), they are predominantly used for low-complexity tasks such as setting timers, playing music, checking the weather or regulating a thermostat [4, 44]. Yet, the increasing sophistication of digital voice assistants enables the possibility that more complex tasks, or even sustained work could be conducted through conversational interfaces. Gartner has predicted that 25% of digital workers will use conversational agents on a daily basis by 2021, and that 25% of employee interactions with business applications will be through voice by 2023[1]. This impending shift towards digital voice assistant-enabled work has the potential to instigate voice-based crowdsourcing as a complementary means to conduct crowd work, rather than a replacement to current approaches (*e.g.,* use of online platforms) [33]. Currently, crowd work is nearly always conducted through a screen-based interface such as a desktop computer or a smartphone, and mostly by workers in their own homes [5]. The hands-free and eyes-free nature of voice interaction could be beneficial to these workers–particularly those that juggle crowd work with other responsibilities at home–by allowing them to complete tasks while doing other things around the home. Also, voice-assistants are a promising way to attract new crowd workers, who are only available to complete small amounts of work at opportune moments.

For example, digital voice assistants can allow users to access crowd work more quickly and conveniently by simply talking to the voice assistant whenever they want to work, rather than having to sit at a desk, log in to a device, launch a browser, and finally select a task [33]. These steps can accumulate a substantial amount of lost time if the user is alternating between work and other activities throughout the day. Furthermore, voice interfaces can make crowd work more accessible to users with vision or motor disabilities that make it difficult for them to engage in screen-based work [62]. On the other hand, not all types of crowd tasks are suited for voice-interaction as they may contain indispensable visual elements or involve complex workflows [18].

While previous research has explored speech transcription through smartphone-based voice input [61, 62], these studies involved ad-hoc systems with a single task. The proposed systems do not provide the capability to browse and launch a wider range of crowdsourcing tasks solely using voice commands. Furthermore, there is no prior work investigating the potential of digital voice assistants or smart speakers for crowd work. To facilitate voice-based crowd work, we developed Crowd Tasker, a novel stand-alone voice crowdsourcing application that can be accessed through any device that supports Google Assistant. To assess whether worker performance using voice input is comparable to a regular web interface,

[1]https://www.gartner.com/en/newsroom/press-releases/2019-01-09-gartner-predicts-25-percent-of-digital-workers-will-u

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376320

we conduct a lab study with 30 participants. We test two types of crowd tasks: voice-compatible (sentiment analysis, comprehension, and text moderation), and voice-based (audio annotation, speech transcription, and emotion labelling), and find that for most tasks, worker accuracy does not significantly differ between the voice assistant and a regular web interface for native English speakers. Subsequently, we conduct a field deployment with another set of 12 participants who completed tasks using a voice assistant in their homes over the course of one week. The aim of the field deployment was to better understand emergent user behaviour and to assess if data quality suffers when completing crowd tasks through a voice assistant when the user is in a less controlled environment. Our results show that participant contributions were of similar quality to those in the lab study. In addition, participants reported that they initiated tasks at opportune moments and worked in brief sessions, while also multitasking when convenient.

Based on our findings, we propose a set of guidelines for the design of future voice-based crowdsourcing systems as well as best practices for creating voice-compatible crowd tasks.

RELATED WORK

Voice Interaction and Digital Voice Assistants

While voice interaction technologies have been developed for a number of decades, there has been renewed interest in the topic with the popularity and growing availability of digital voice assistants. Recent work by Bentley *et al.* [4] examines the use of digital voice assistants in 88 households. Their speech log analysis reveals that users engage with smart speakers through short sessions throughout the day as opposed to using the device for longer periods of time to complete a series of tasks. Furthermore, they show that users more frequently utilise smart speakers when compared to phone-based voice assistants. They identify Music, Information (*e.g.,* asking for spellings) and Automation (*e.g.,* turning off lights) as most frequently used command categories.

Several studies have compared voice input to manual input for the same task and report that voice input rates well on engagement, but poorly on usability and sense of control [2, 45]. Due to a lack of typical user interaction signals like mouse clicks and scroll movements, measuring and evaluating user satisfaction on voice interfaces greatly differs from traditional screen based interfaces. In a recent survey, Kocaballi *et al.* [39] examine a number of studies that aim to understand and measure user experience in conversational interfaces. For example, Hashemi *et al.* [32] propose to model user satisfaction by creating intent sensitive word embeddings or by representing user interactions as a sequence of user intents. The literature also proposes design guidelines that can create better voice user interfaces [15, 16, 46, 47]. However, research highlights that voice interfaces require better theories and more design guidelines, due to persistent usability issues [14, 48].

Further, research shows that assimilation bias can have an impact on performance in voice user interfaces. In a study where participants were asked to use a voice based calendar application, Myers *et al.* [49] report that participants with increased experience with voice user interfaces took less time with tasks. In addition to the experience, language proficiency is known to impact the usability of digital voice assistants. Pyae *et al.* [52] report that native English speakers had a better overall user experience when compared to non-native English speakers when using Google Home devices. Research has also shown that matching the personality of the voice assistant and the user's expectations can result in higher likeability and trust for assistants [6]. In a study involving older adults, Chattaraman *et al.* [9] report that users' internet competency and the digital assistant's conversational style can have significant interaction effects on social (*e.g.,* trust in the system), functional (*e.g.,* perceived ease of using the system), and behavioural intent outcomes. Several other studies have also confirmed that people respond differently to synthesised voices depending on how they sound and whether they are polite [12, 13].

While voice interaction is associated with numerous benefits, literature also looks at several negative aspects. Researchers have investigated different privacy concerns of using digital voice assistants [42]. This research has led to studies that aim to mitigate potential attacks, such as the work by Kwak *et al.* [41] that distinguish genuine voice commands from potential voice based attacks. In addition, voice interaction is not considered socially acceptable in all public situations [53].

Crowdsourcing with Audio and Speech Data

There exists a wide range of crowdsourcing tasks that use speech or audio data [21]. Such tasks require workers to listen to audio data and/or provide answers through voice input. For instance, crowdsourcing has been used to gather speech data from different local dialects [43], rate speech data for assessing speech disorders [7], annotate audio data [22, 25], and annotate speech data for training automatic speech recognition systems [8]. In a speech sound rating task, Byun *et al.* [7] state that the inability to standardise equipment or playback is a major limitation when using an online crowdsourcing platform like Amazon Mechanical Turk. There are also numerous other tasks, such as sentiment analysis and moderation, that can be completed via voice input although they typically contain text data and text responses.

Vashistha *et al.* [61] introduced 'Respeak', a mobile application that uses voice input for crowdsourcing speech transcription tasks. In the study, participants listen to short audio clips and repeat what they had heard. In a deployment with 25 university students in India, the study shows that audio files could be transcribed with a word error rate of 8.6% for Hindi and 15.2% for Indian English. The application uses Google's Android Speech Recognition API to generate transcripts of user utterances. An extension of the proposed application was also successfully used to crowdsource speech transcription tasks from visually impaired users [62] and through basic phones [60]. However, all three studies are limited to speech transcription and none of them are fully functional hands-free voice interfaces that have the capability to browse available tasks, launch tasks, and check progress.

In a vision paper, Hettiachchi *et al.* [33] propose that it is feasible to use smart speakers for crowdsourcing and discuss potential benefits like low cost of entry, ubiquitous nature,

efficiency and accessibility. They also highlight several challenges such as privacy concerns, integration issues, and impact on data quality when multitasking. We extend this work with an empirical evaluation, where we present a functional voice interaction application for crowdsourcing with several different tasks, and evaluate the system using both a lab study and a field deployment.

Crowd Worker Context

In most crowdsourcing platforms, such as MTurk, Figure Eight, and Prolific, crowd workers actively select and launch tasks they wish to work on. This model typically introduces higher latencies for tasks that require workers with specific skills (*e.g.,*Translation) [20]. As a solution, several studies have investigated the possibility of proactively delivering tasks to workers instead of waiting for them to initiate the task [1, 36]. In mobile crowdsourcing, Acer *et al.* [1] investigate how worker mobility patterns, workflow, and behavioural attributes can be used to identify opportune moments to deliver tasks to mobile crowdworkers. The study aims to embed crowdsourcing tasks to workers' daily routine and reports increased worker response rate and accuracy. In crowdsourcing, task requesters also aim to capture the cognitive surplus of workers, which is described as the free time of individuals who are capable of contributing to a task [55]. Different techniques can be used to tap into the cognitive surplus. Goncalves *et al.* [27, 29] show that interactive public displays can be successfully used to gather input from people who are idling at public spaces, while Hosio *et al.* [35] demonstrated the feasibility of a situated crowdsourcing system. In another example, Skorupska *et al.* [57] show that older adults can contribute to a transcription task while watching a movie.

By using digital voice assistants for a broader spectrum of crowd tasks, we aim to reduce the complexity of initiating crowd work. By doing so, this is likely to lead to a better utilisation of cognitive surplus and opportune moments for crowdsourcing purposes.

CROWD TASKER SYSTEM

To enable crowdsourcing through digital voice assistants, we developed Crowd Tasker, an application for Google Assistant which prompts crowd tasks to users and stores responses. We opted for Google Assistant as it has the largest market share in Digital Voice Assistants [50], and allows us to easily deploy our application to both smart speakers and smartphones. We used Dialogflow[2] and the NodeJS client library for Actions on Google[3] to process user utterances and manage the crowd task flow. Using Dialogflow we mapped users' voice input to a set of pre-configured intents that lead to different actions. An intent represents an end-user's intention for one conversation turn. It also allowed us to activate different intents based on the context, such as a previous response by the user. Figure 1 shows the different intents we developed considering main use cases of online crowd work along with their flow within the application. We iteratively improved our system prototype to provide a unified user experience.

[2]https://dialogflow.com/

[3]https://github.com/actions-on-google/actions-on-google-nodejs

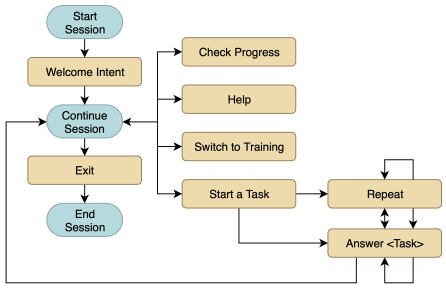

Figure 1: Intents and session flow of Crowd Tasker system.

Our application works with Google Assistant, and users can launch it by saying 'Talk to Crowd Tasker'. The application does not use any visual prompts and can be initiated from any device that supports Google Assistant (*e.g.,* smartphone, smart speaker, smart watch). Users can check what kind of tasks are available, start a particular task, review their progress, and ask to repeat questions. The system also asks users to provide a clear response or seek help and listen to instructions when it cannot understand the user's utterance. Below is an example for a conversation between Crowd Tasker and a user.

Participant : Okay Google, talk to Crowd Tasker
Crowd Tasker: Hi, I am the Crowd Tasker, What can I do for you today?
Participant: What kind of tasks do you have?
Crowd Tasker: There are six available tasks: Sentiment analysis, Comprehension, [...] and Emotion labelling. Which task do you want to complete?
Participant: Please start Audio Annotation task.
Crowd Tasker: <audio clip playing>
Participant: A dog barking
[Conversation continues]

Crowdsourcing Tasks

We created six unique crowdsourcing tasks under two categories: voice-compatible (*i.e.,* typically text-based tasks, but can be adapted for voice input) and voice-based (*i.e.,* audio or speech related). The voice-compatible tasks are sentiment analysis, comprehension, and text moderation. The voice-based tasks are audio annotation, speech transcription and emotion labelling. Task selection was informed by a crowd task taxonomy [24] and task availability [19] from prior work to be representative of typical tasks available in crowdsourcing platforms. Each task contained 10 different questions with varying complexity extracted from datasets that are extensively used in the literature.

Sentiment Analysis: Participants are asked to identify the sentiment of a given sentence. We adopt the task from prior work that investigates crowd task performance [26, 28, 34]. Sentiment can be classified as either 'positive', 'negative' or 'neutral'.

Comprehension: Participants provide an answer to a question based on a provided text. Questions are based on different Wikipedia articles. For the task deployed through the voice assistant, we use English speech data from The Spoken Wikipedia project [40].

Text Moderation: Workers are asked to label text messages as 'spam' or 'not spam'. Data is extracted from the SMS Spam Collection [3]. In the web interface, the message is presented in text format, whereas when using the voice assistant, participants listen to the message as generated by Google's text-to-speech service.

Audio Annotation: In this task, participants are asked to provide a label that describes a sound they hear. All the audio clips and ground-truth labels are extracted from the Freesound Data set [22]. An answer is considered accurate if it matches any of the valid keywords for the clip. For example, for a clip of a moving horse carriage, terms such as horse and cart are considered as valid answers.

Speech Transcription: In the speech transcription task, participants listen to a short audio clip (average length of 3 seconds) which contains an utterance of an English speaker. Participants are asked to clearly speak out or type in what they heard. Speech data and transcripts are sourced from the Noisy speech database [59]. We use the Levenstein distance [17] to calculate the accuracy of each answer.

Emotion Labelling: For the emotion labelling task, we use the Multimodal EmotionLines Dataset [51], which contains short utterances of different people from a popular TV show. We extract audio clips and ground-truth labels for two people. Workers are asked to categorise the emotion of each utterance as either 'anger', 'disgust', 'fear', 'joy', 'sad', or 'surprise'.

STUDY

Lab Study

We conducted a lab study to compare crowd task performance through web (*i.e.,* using a regular graphical user interface) vs. digital voice assistants (*i.e.,* using a voice interface). Hence, we also built a simple web application that replicates the task completion interface of a typical crowdsourcing platform. The system was developed using Python (Django framework) and connected to the Crowd Tasker database that contains task and performance data.

We recruited 30 participants through a university-wide online notice board, using two eligibility constraints: we only recruited native or fluent English speakers, and only participants who have used digital voice assistants. During screening, participants reported whether they used digital assistants frequently (daily or more than few times a week), occasionally (few times a week), or rarely (few times a month). We balanced the use of voice assistants by recruiting 10 participants for each category (30 participants in total). Participants were compensated with a $20 gift voucher.

Participants completed tasks under two conditions: using a desktop-based web interface (Figure 2), and a Google Home smart speaker. Initially, participants completed a training round in which they completed one question from each task for both conditions. We then counter-balanced the order of the experimental conditions, and randomised the completion order of all tasks a priori. Each task contains 10 questions, and participants answered 5 questions per condition. Table 1 summarises tasks under each condition. Finally, participants completed a short exit interview to discuss their experience, and we probed them about convenience, perception of the two conditions, and task difficulty.

	Question		**Answer**	
Task	*Web*	*VA*	*Web*	*VA*
Sentiment Analysis	Read	Listen	Button	Speak
Comprehension	Read	Listen	Type	Speak
Text Moderation	Read	Listen	Button	Speak
Audio Annotation	Listen	Listen	Type	Speak
Speech Transcription	Listen	Listen	Type	Speak
Emotion Labelling	Listen	Listen	Button	Speak

Table 1: Crowdsourcing Tasks

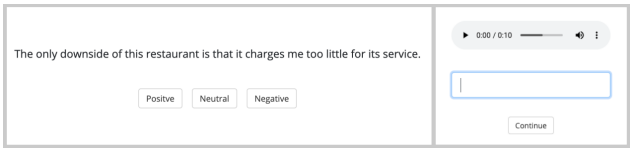

Figure 2: Screenshots of the web interface for Emotion Labelling (left) and Comprehension (right) tasks

Field Deployment

To further examine the feasibility of using digital voice assistants for crowdsourcing purposes we conducted a field deployment. Before proceeding with the field deployment, we made several enhancements to the system based on the feedback of participants of the lab study, including several workflow improvements. For instance, in the lab study, participants listened to the text segment prior to the question. We swapped the presentation order, so that participants could anticipate the relevant information when listening to the text. We also improved the timing for gaps in between spoken text to enhance the quality of overall conversation experience. For example, in the lab study, participants mentioned that they found it difficult to distinguish between instructions and the first question of a task due to absence of an appropriate time gap (similar to proximity in Gestalt Principles [30]. Finally, we added an intent, that allows participants to check their progress and know how many questions are remaining in each task.

We recruited 12 participants through our university's online notice board. Similar to the lab study, we set out eligibility constrains and recruited a population that is balanced in terms of participants' experience with voice assistants. Additionally, we did not recruit any of the participants who completed the lab study.

At the beginning of the study, we met our participants in person, provided them a Google Home Smart Speaker, and asked them to use it in their home for a period of 7 days. We also instructed participants on how to setup the device and use the application, and finally gave them a brief demonstration. We asked them to complete all the training tasks first. We then explained how they could complete tasks: through the provided smart speaker or using the digital voice assistant application on another device.

In the field deployment, participants were required to complete 6 tasks similar to the lab study. To replicate the reward mechanism of a standard crowdsourcing marketplace, we informed the participants that they would be compensated based on the number of tasks they complete in the study. Participants were given a gift voucher of up to $30 if they completed all available tasks. After a week, participants returned the smart speaker and took part in a short interview about their experience. We asked participants to report on the level of convenience, whether they were doing any other activity while completing tasks, and how they compare interacting through the smart speaker and another device.

RESULTS

Lab Study

30 participants (18 women and 12 men) completed the lab study. Participant age ranged from 18 to 38 years ($M = 25.7$, $SD = 5.7$). 8 participants were native English speakers, while the remaining participants were fluent English speakers. We did not observe any significant impact on task accuracy in terms of participant demographics (age and gender) or voice assistant usage. We also asked participants whether they had prior experience with particular voice assistant services such as Google Assistant, Siri, and Alexa. However, there was no significant effect on task performance from any of the indicators.

Web Interface vs. Voice Assistant

Differences in worker performance between the web interface and the voice assistant in terms of accuracy and task completion time are shown in Figure 3. A paired-sample t-test indicates that there is no significant difference in accuracy between the web and voice assistant conditions for the text moderation task ($t(29) = -0.90$), $p = 0.38$). Similarly, a Wilcoxon signed rank test revealed that there is no difference in task accuracy for the emotion labelling task ($Z = 89$, $p = 0.90$). Task accuracy is significantly higher in the web interface for all the remaining tasks.

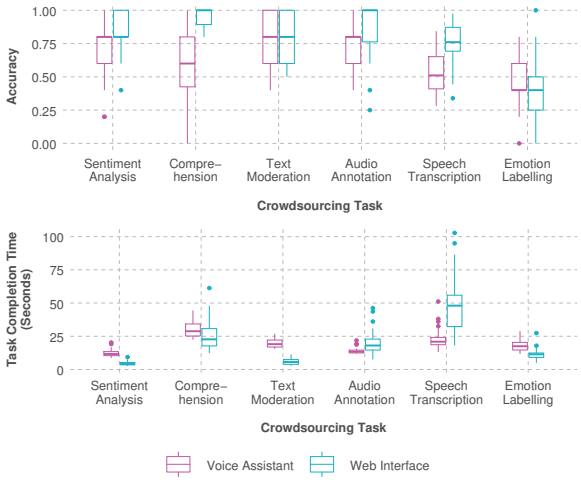

Figure 3: Worker accuracy (top) and task time (bottom) across crowd tasks when using the web and voice assistant

Paired t-tests indicated that task completion time is significantly lower in the voice assistant condition for two voice-based tasks, audio annotation ($t(29) = -3.62$, $p < 0.01$) and speech transcription ($t(29) = -6.33$, $p < 0.01$). For all 3 voice-compatible tasks and emotion labelling task, task completion time in voice assistant is significantly higher than the web interface.

Native English Speakers

As shown in Figure 4, when completing tasks through voice, native English speakers exhibit a higher task accuracy for most tasks when compared to the remaining participants.

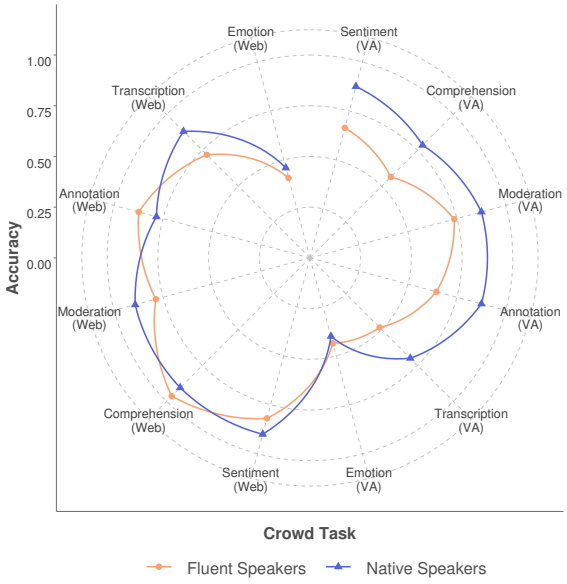

Figure 4: Native and fluent English speakers accuracy across crowd tasks when using voice assistants and web

We further examined the difference in task accuracy for native speakers. A paired-sample t-test was conducted to compare the accuracy in web and in voice assistant conditions. There was no significant difference in the accuracy for web and voice assistant conditions in sentiment analysis task ($t(7) = -0.42$, $p = 0.68$), moderation task ($t(7) = -0.11$, $p = 0.91$), audio annotation task ($t(7) = 0.77$, $p = 0.46$), and emotion labelling task ($t(7) = -0.39$, $p = 0.71$). As the accuracy scores of reading comprehension tasks were not normally distributed, we used a Wilcoxon signed rank test and found no significant difference in accuracy in web and voice assistant conditions ($Z = 1$, $p = 0.42$).

Native English speakers exhibited a similar variation to the general sample, considering task completion time. To compare the task completion time in the web interface and voice assistant, we conducted paired t-tests or Wilcoxon signed rank tests (when the sample was not normally distributed). For speech transcription, task completion time was significantly lower in voice assistant condition ($t(7) = -3.42$, $p = 0.01$). For audio annotation, there was no significant difference between two conditions ($t(7) = -1.05$, $p = 0.33$). For emotion labelling task and the remaining voice-compatible tasks, task

completion time was significantly higher in voice assistants when compared to web interface. Results from our lab study, for the general sample and the native English speakers are summarised in Table 2.

	Accuracy		Task Time	
Task	*General Sample*	*Native Speakers*	*General Sample*	*Native Speakers*
Sentiment Analysis	↓	-	↑	↑
Comprehension	↓	-	↑	↑
Text Moderation	-	-	↑	↑
Audio Annotation	↓	-	↓	-
Speech Transcription	↓	↓	↓	↓
Emotion Labelling	-	-	↑	↑

Arrows indicate that the measure is significantly higher (↑) or lower (↓) in the voice assistant when compared to the web interface. No statistically significant difference between conditions is indicated through a dash (-)

Table 2: Summary of statistical results of the lab study

Qualitative Data

To further extend these results, we present a qualitative analysis of our semi-structured interview data. Informing ourselves through the aforementioned quantitative results and the setup of the field study, we apply the general inductive approach to data analysis as defined by Thomas [58]. Two of the paper's authors (one of which conducted the interviews) independently analysed and coded the interview data – after which three of the authors agreed on the final set of themes. Given the study's exploratory character, we focus on the following themes; 'participant interaction', 'task suitability', and 'perceived usefulness'.

Participant interaction

Participants considered the specific advantages of using either a web interface or a voice assistant. The web interface was perceived as offering participants a higher level of control over their input, with one participant describing this as;

> P08: "I feel I am more accurate and precise when I type. I also feel I'm more in control in the web."

The web interface allowed participants to work at their own pace, without pressure from a timeout by the voice assistant.

> P10: "Voice has more pressure, I need to give a timely response. [Using the web interface,] I feel more relaxed as the pace is defined by me."

However, a number of participants ($n = 13$) found it easier, more efficient, or simply more enjoyable to speak out the answer as opposed to typing. Participants compared the interaction with the voice assistant to be more human-like as compared to input provided through a web interface.

> P10: "Also when the answer is too long then voice is easier because it saves time of typing."

Furthermore, the voice interface provided participants with benefits we did not initially consider. One participant stated that the use of voice commands requires less focus on the correct spelling of words.

> P15: "If you want me to type, I am not sure about the words (spellings), but I don't have to worry about that when speaking."

Task suitability

Given the wide range of tasks included in our study, we aimed to identify task suitability in relation to voice interaction. Several participants ($n = 10$) reported that they found it difficult to remember content when completing tasks through the voice assistant. Participants also mentioned that they were unable to memorise all options in the emotion labelling task.

> P18: "Speech transcription and comprehension tasks were harder on voice assistant, because I had to remember longer sentences."
>
> P16: "Emotion labelling was difficult because I didn't remember the emotions."

Therefore, when interacting through voice, participants preferred tasks with fewer options such as text moderation and sentiment analysis and tasks with short answers like audio annotation. A number of participants also mentioned that the use of voice control allowed them to respond quicker and accelerate the interaction.

> P24: "The rest of the tasks (apart from comprehension) were easier with the speaker, because it took off the effort of reading and typing manually."

Perceived usefulness

Although the study was carried out in a controlled lab environment, we were interested in the participants' perceived usefulness of a voice assistant in completing crowdsourcing tasks. Participants believed that the use of a voice assistant would enable them to simultaneously work on something else.

> P21: "Using the voice assistant, you can be busy and multitask but with web its not possible."

Furthermore, participants described scenarios in which they believe the use of a voice assistant would be useful, including examples such as leisure time, cleaning, and cooking.

> P26: "I think I will use voice assistant a lot during my leisure time. During the weekend I do household work so I can use the speaker."

Field Deployment

12 participants (5 women and 7 men) aged between 20 to 32 years old ($M = 26.7$, $SD = 3.8$) completed the field deployment study. Five participants were native English speakers, while the remaining participants were fluent English speakers. All participants completed the full set of tasks within the one week period.

In Figure 5, we can observe a similar task performance in the lab study and the field deployment. Our statistical analysis confirmed that for all the tasks, there is no significant difference in both the accuracy and task time between tasks completed through the smart speaker in the lab study and the field deployment.

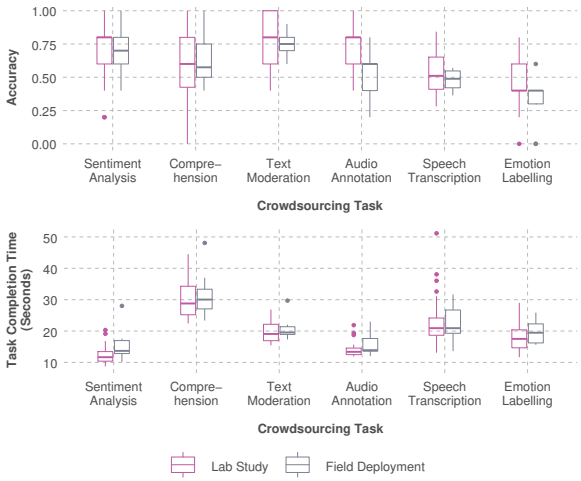

Figure 5: Task accuracy (top) and Task completion time (bottom) for participants in the lab study and in field deployment

To understand participants' voice assistant usage for crowd work, we further examined their usage patterns from the task completion data. Figure 6 shows the total number of question answered by all participants over the time of day.

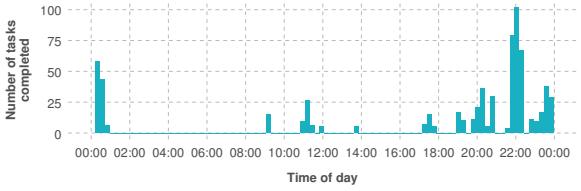

Figure 6: Total number of questions answered by participants over the time of the day

As exemplified in Figure 7, six of the participants completed tasks over more than one day. The other six participants completed all questions within a single day. Figure 8 shows the question completion over time for those participants. Although they completed all questions within the same day, we notice that they used multiple brief sessions to complete tasks with interruptions or breaks in between.

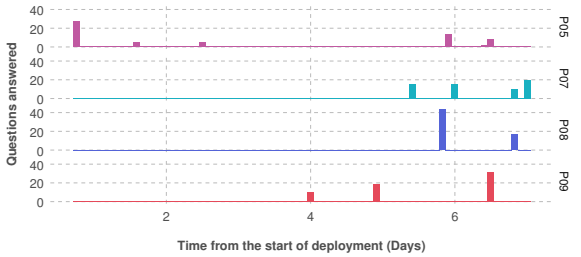

Figure 7: Task completion by day of deployment for four participants

While all participants had a smart speaker set up in their home, they were also given the option to complete tasks through digital voice assistants on their smartphones. In the field

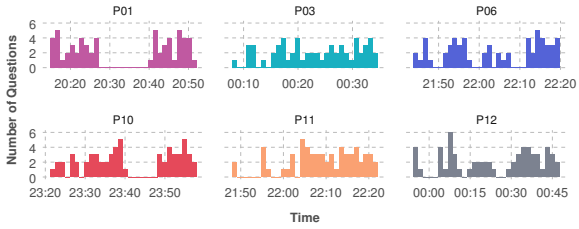

Figure 8: Task completion by time of day for six participants

deployment, 7 participants completed all tasks through smart speakers, whereas 3 participants used the smartphone for all the tasks. Only 2 participants used both devices, completing 35 and 30 questions through their smartphone.

Application logs indicate that participants used 129 sessions in total to interact with Crowd Tasker. On average participants used 13.07 queries per session. Table 3 presents a summary of user utterances indicated by corresponding intents matched by voice assistant during the field deployment. We notice a high number of repeats for Speech Transcription task. 'Check progress' intent has the highest number of matches after 'Start a task intent' and a higher exit rate, suggesting participants often checked their progress before closing the application.

Intent	*Num. of sessions*	*Num. of matches*	*Exit percentage*
Welcome Intent	129	130	13.1%
Switch to Training	24	31	9.7%
Start a Task	80	241	4.6%
Check Progress	59	142	16.2%
Check Available Tasks	21	31	6.5%
Sign-in (during briefing)	12	12	100%
Help	3	3	0%
Answer - Sentiment Analysis	26	129	3.9%
Answer - Comprehension	23	126	3.2%
Answer - Text Moderation	17	120	0%
Answer - Audio Annotation	22	128	0%
Answer - Speech Transcription	20	122	1.6%
Answer - Emotion Labelling	39	135	3.7%
Repeat	6	17	5.88%
Repeat (after starting a task)	21	43	9.3%
Repeat - Speech Transcription	13	55	1.8%
Repeat - Comprehension	6	24	0%
Repeat - Audio Annotation	2	3	0%

Table 3: Summary of intent matching

Qualitative Data

Similar to the qualitative analysis of our lab study, we again perform an inductive approach to the analysis of our semi-structured interviews obtained following the field deployment [58]. The focus of our analysis is on the practical aspects of crowdsourcing using a voice assistant *in situ*. We therefore focus on the following themes; 'ease of use', 'multitask behaviour', and the use of 'smartphone vs smart speaker'.

Ease of use

Our in-the-wild deployment highlighted issues in interacting with the device. For example, the voice assistant could occasionally not recognise participants' utterances due to accent, background noise, or volume level. An issue we previously did not consider was the interaction between participant devices. Some participants reported that multiple devices (*e.g.,*both smartphone and smart speaker) were activated when issuing the command to initiate the voice assistant.

> P07: "My phone got activated when I spoke to the speaker."

Furthermore, the level of voice recognition differed between participants, occasionally hampering the participant's ability to complete tasks.

> P03: "Sometimes the speaker had trouble understanding my accent or it didn't pick up my voice."

Generally, however, our participants ($n = 8$) highlighted that completing tasks through voice was convenient and the system was easy to use. In particular, the launching of tasks was seen as straightforward in comparison to the use of a computer.

> P02: "It was quick and easy to complete tasks through the voice assistant."
>
> P10: "Good thing about voice assistant is that I could quickly start tasks."

Multitask behaviour

Participants mentioned that they had to change their attention depending on the task. These participants ($n = 7$) were able to attend to other tasks or switch context while going through tasks.

> P09: "I was interrupted once. I had to talk to my flatmate."
>
> P06: "I was folding laundry while doing the tasks. I really didn't have any problem."

Some participants ($n = 3$) also mentioned that they initiated Crowd Tasker at opportune moments such as during idling, followed by a routine task, or when they needed a distraction from a particular task.

> P03: "I was free when I started it (Crowd Tasker) and I was sitting on a couch. I was occasionally checking my smartphone for notifications while doing the tasks."
>
> P02: "Probably I did after dinner. I was paying full attention but watching something in between tasks."
>
> P07: "Even when I was using the speaker, I was helping my friend arrange the house. I was also in the middle of an assignment. Because I wanted a distraction, I started the task."

Smartphone vs Smart Speaker

Participants who opted to use their smartphone instead of the smart speaker appreciated the fact that they can get a visual confirmation of their utterances.

> P06: "If I am using the phone, I know if Google understood me correctly."

Participants also highlighted that it was beneficial to see the question on the screen and then answer through voice when the task was too complex.

> P11: "Half of these tasks are easy with voice. For others it might be good to use voice assistant in phone, so you can see the question but still can answer through voice."

As most participants ($n = 7$) chose to use the smart speaker for all the tasks, they commented on positives of the speaker such as better audio quality and voice recognition distance as compared to smartphones.

DISCUSSION

Our study is the first to systematically investigate the possibility of using voice-only interfaces, such as smart assistants, for crowdsourcing purposes with a variety of tasks. Through a lab study and a field deployment we are able to demonstrate the feasibility of this approach, and at the same time highlight a number of remaining research and design challenges.

Our work is at the nexus of the literature on crowdsourcing and voice interaction. These have been largely distinct, each having a long tradition of design guidelines, best practice suggestions, and research findings. As we discuss here, we find that the usability of voice interaction needs to be carefully thought through when developing voice interfaces for crowd work, particularly by taking into account the nature of crowd work. For instance, we find that crowd work requires humans to provide answers to the agent's questions, whereas typically with voice assistants it is humans who provide the questions and agents the answers. This poses a number of challenges that we highlight in our discussion.

Crowdsourcing through Voice

We show that crowd tasks can be completed through a voice assistant with an acceptable level of speed and accuracy, and that a voice assistant can provide crowd workers with greater flexibility in how they approach crowd tasks compared to a regular web interface. Indeed, participants were faster at completing free-form answer tasks with Crowd Tasker than with a typical web interface. Prior work reports that high quality data could be obtained by using voice input for crowdsourcing speech transcription tasks [60, 61, 62]. While we were able to obtain reasonable data quality for our transcription task, as detailed in Table 2, results from our lab study indicate that other tasks are better-suited for voice-based crowdsourcing systems. For crowdsourcing platforms, it would be more productive to create a unified system that issues online and voice-based crowd work depending on the type of task. Perhaps, Amazon Web Services is best positioned to achieve this as they own a crowdsourcing platform (Mechanical Turk) as well as a voice assistant service (Alexa). Other crowdsourcing platforms can also plausibly create voice assistant applications that tie into their own market. This would also require further research that explores dynamic task assignment to either a web or voice interfaces.

In contrast to previous systems that involve calling a phone number to access an interactive voice response (IVR) application [60] or providing voice input through a smartphone appli-

cation [61, 62], accessing CrowdTasker through an always-on microphone can bring numerous benefits to users. The qualitative results of the field deployment highlight that participants were able to utilise their cognitive surplus by initiating tasks at opportune moments [55]. The hands-free interaction through the speaker also allowed participants to multitask while completing crowd tasks. In Figure 6, we observe that participants completed most tasks outside regular working hours. We also note that there is no statistically significant difference in performance with the voice assistant between the lab study and the field deployment, suggesting that multitasking did not have a negative impact on the data quality.

Although we recruited fluent English speakers for our study, we observed a significant difference in task accuracy between native and non-native English speakers. Our findings are in line with prior work that states that native English speakers have a better overall experience with smart speakers [52]. From our qualitative analysis, we understand that this difference is due to recognition problems on both sides of the interaction: the voice assistant being less successful at recognising non-native English speech, and the non-native English speakers being less able to comprehend the voice assistant's speech due to tone, accent, and speech rate. Therefore, when using voice assistants for crowd work, language competency of the worker will play an important role. Using language-based pre-selection mechanisms or support for multiple languages could be feasible solutions to mitigate this factor. We also note that language skills are important for a wide array of crowdsourcing tasks in regular web-based crowdsourcing platforms, so this problem is not unique to the voice interface.

Developing Voice-based Crowdsourcing Platforms

While our study reveals promising results for the use of voice assistants for crowd tasks, it also identifies several factors that undermine the user experience and data quality. We discuss these challenges and propose ways to address them in the development of future voice-based crowdsourcing systems.

Optimising workflow to provide control for workers

Our participants mentioned that they felt less in control when using the voice assistant. This is consistent with prior studies that have found voice interaction to be associated with a lower subjective sense of control in both smart-home interfaces [45] and digital games [2]. We propose three features to reduce this perceived lack of control. First, the voice assistant should repeat the entire or part of a task question upon request. Crowd Tasker gives the user an option to repeat an entire question, but the qualitative feedback from our study highlights that this feature should be extended to provide more granular control. For instance, each question in the comprehension task consisted of two parts: a question and a sentence. In the current implementation Crowd Tasker repeats both parts when asked, forcing the user to listen to seconds of potentially irrelevant content when they may only want to check a single word. In a future implementation, it would be useful for the user to be able to request only a specific part to be repeated.

Second, workers should be able to stop and resume tasks at any point. Although Crowd Tasker was designed to provide a predefined number of questions in each task, it will exit the application if it receives no response from the worker after repeated prompts. When the worker returns to the voice assistant, they can resume from the last completed question. For voice-based crowd work, we recommend including more task checkpoints than in web-interfaces.

Third, for certain task types, it is useful if the worker can skip certain sections or interrupt the voice assistant while it is speaking. For example, in the text moderation task, workers had to listen to the entire message even if they had figured out their answer within the first few words. Currently, popular commercial voice systems provide limited interruptibility, but future systems that allow for more interruptions are likely to provide a more pleasing user experience for crowdsourced voice work.

Finally, due to the nature of voice interaction, it can be beneficial to ask participants a question first, and then provide them the relevant stimuli to complete the task. In screen-based crowd work, this information is often given in the reverse order, but due to the visual nature of those tasks it is possible for workers to quickly switch between task description and stimulus. With a voice-only modality, our participants preferred to be asked the question first, so that they know what to look for when listening to a stimulus, effectively having a reduced working memory load. This ultimately reduces the need to repeat the task description, improve accuracy, and result to greater subjective satisfaction.

Handling responses

When designing and developing Crowd Tasker, handling responses for questions that require free-form answers proved to be particularly challenging. We also note that during the field deployment, specific worker commands such as asking to repeat the question were, on occasion, erroneously captured as answers. As a possible solution to mitigate errors, we propose that future systems should allow users to listen to and revert their answer if necessary.

Payment

The payment mechanism in a voice-based crowdsourcing system can be similar to existing online crowdsourcing platforms where workers are paid per completed task [18], with each task having a maximum time limit. In a voice-based system, prompts should be added to indicate if a worker runs out of time. When estimating task times for payments, it is important to consider the time taken to playback the question or prompt, to ensure fair compensation [31].

Task allocation and recommendation

In conventional crowdsourcing platforms, workers need to browse and select the task they wish to attempt. This process takes a considerable amount of time and effort [11]. Similarly, browsing through a large number of tasks is not desirable in a voice interface. Therefore, it is critical to allocate or recommend a handful of relevant tasks to workers when delivering tasks through voice. There is a large body of work that analyses different worker attributes [34, 37], and behavioural traces [23, 54] that can be utilised to match tasks to workers. In addition, our participants mentioned that they had to repeatedly ask for available tasks as they had to use the specific task

name to initiate the task. Thus, a feature that automatically starts a relevant task for the user will be particularly useful in a voice-based crowdsourcing system.

Selecting tasks for voice

Our findings show that certain tasks are more appropriate for voice interaction than others. We initially anticipated that inherently voice-based tasks (such as audio annotation, speech transcription, and emotion labelling) would be more suitable than tasks that are voice-compatible but text-based (such as sentiment analysis, comprehension and text moderation). However, our analysis suggests that other factors play a more critical role in determining task suitability, such as demand on working memory and task complexity. We discuss such factors extensively in the following section.

Designing Voice Interaction for Crowd Tasks

While there exists a significant amount of research regarding conversational interfaces, our study shows that crowd work is a peculiar case. Whereas in traditional conversational interaction the user may be prompted to talk about their desires and preferences, in the crowd work scenario users are typically prompted to talk about stimuli they have just heard, which increases their cognitive load. Minimising strain on users is crucial, as satisfaction with voice assistants has been shown to vary according to the level of effort involved in the tasks for which they are used [38]. Therefore, based on our results, we highlight several important considerations for designing voice systems for crowd work.

Shortening questions and answers

Voice interfaces often place higher demand on users' short-term and working memory compared to graphical interfaces, due to both the lack of visible cues and the fact that speech uses more of these cognitive resources than hand-eye coordination does [56]. This was manifested in our study, as participants struggled with those tasks that required them to hold information in mind while completing an action through voice. Accuracy was decreased and subjective reported effort was increased for Speech Transcription and Comprehension in particular, as these tasks involved working with samples of speech that were too long to hold in working memory. Hence, we suggest as a general rule that the amount of information presented in a single conversation turn of a voice-based crowd task should be kept to a minimum.

The same holds for longer answers. For tasks such as Speech Transcription, which required long responses, participants preferred to type out the recording using the web interface rather than dictate it by voice, as typing made it easier to transcribe short chunks at a time, rather than attempting to transcribe the entire sentence at once. Therefore, in addition to short prompts and questions, it is also important to keep the user's required responses as short as possible when designing crowd tasks for voice. When it is necessary to work with longer sentences, such as in transcription, it is recommended to break them into smaller sub-tasks [10].

Reducing the number of options in multi-label tasks

Our study had three tasks that required participants to choose a predefined option as the answer. Of these, participants found the emotion labelling task particularly difficult as the voice condition required them to remember 6 different response options. In contrast, many participants mentioned the sentiment analysis (options: 'positive', 'negative', 'neutral') and text moderation (options: 'spam', 'not spam') tasks as being no more difficult with voice than with the web interface. Therefore, we recommend that voice-based crowd tasks should provide only a small set of options for the user to choose from. Where multiple labels are necessary, a task decomposition technique could be used to transform the multi-label task into multiple binary labelling tasks [63]. For some tasks, such as emotion labelling, it may be feasible for the system to obtain an open answer, and then map that answer on to a hidden set of options using natural language processing techniques.

Limitations

We acknowledge three limitations in our study. First, our evaluation is limited to one voice assistant. While there are several other services available, we decided to use Google Assistant due to the complexity introduced to the study design when using multiple services. In addition, Google Assistant currently holds the largest market share [50] and there is currently no substantial difference in workflow or the ability to recognise voice across the major services.

Second, our participants are not regular crowd workers and have only limited experience with crowdsourcing. While it would be more relevant to evaluate this system with crowd workers, there are numerous practical difficulties in deploying a system of this nature in the wild.

Third, our field deployment is limited to a week as we used the same questions from the lab study for better comparability. A more longitudinal future study that recruits more participants can reveal further insights on designing commercial voice-based crowdsourcing systems.

CONCLUSION

To investigate the feasibility of using digital voice assistants for crowdsourcing, we developed Crowd Tasker, a novel stand-alone crowdsourcing system that runs on Google Assistant. Through a lab study, we report that for native English speakers, there is no significant difference in task accuracy when completing five types of tasks through a regular web interface and a voice interface while task completion time varies depending on the task. Further, through a field deployment, we show that participants were able to complete tasks conveniently through voice at their home. They were able to multi-task, launch tasks at opportune moments, and resume their work if interrupted or distracted while using the voice assistant.

We identify several approaches on how to optimise voice-driven workflow, handle responses, recommend or allocate tasks, and select tasks when developing voice-based crowdsourcing systems. We anticipate that our work will lay the foundation for different research avenues to explore voice-based crowd work as a noteworthy addition to the existing crowdsourcing eco-system, and help create more accessible and convenient platforms for crowd workers.

REFERENCES

[1] Utku Günay Acer, Marc van den Broeck, Claudio Forlivesi, Florian Heller, and Fahim Kawsar. 2019. Scaling Crowdsourcing with Mobile Workforce: A Case Study with Belgian Postal Service. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2, Article 35 (June 2019), 32 pages. DOI: http://dx.doi.org/10.1145/3328906

[2] Fraser Allison, Joshua Newn, Wally Smith, Marcus Carter, and Martin Gibbs. 2019. Frame Analysis of Voice Interaction Gameplay. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 393, 14 pages. DOI: http://dx.doi.org/10.1145/3290605.3300623

[3] Tiago A. Almeida, José María Gómez Hidalgo, and Tiago P. Silva. 2013. Towards SMS Spam Filtering: Results under a New Dataset. *International Journal of Information Security Science* 2, 1 (2013), 1 – 18.

[4] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 91 (Sept. 2018), 24 pages. DOI:http://dx.doi.org/10.1145/3264901

[5] Janine Berg. 2015. Income security in the on-demand economy: Findings and policy lessons from a survey of crowdworkers. *Comp. Lab. L. & Pol'y J.* 37 (2015), 543.

[6] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At Your Service: Designing Voice Assistant Personalities to Improve Automotive User Interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 40, 11 pages. DOI: http://dx.doi.org/10.1145/3290605.3300270

[7] Tara McAllister Byun, Peter F. Halpin, and Daniel Szeredi. 2015. Online crowdsourcing for efficient rating of speech: A validation study. *Journal of Communication Disorders* 53 (2015), 70 – 83. DOI: http://dx.doi.org/10.1016/j.jcomdis.2014.11.003

[8] Chris Callison-Burch and Mark Dredze. 2010. Creating Speech and Language Data with Amazon's 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, Stroudsburg, PA, USA, 1–12.

[9] Veena Chattaraman, Wi-Suk Kwon, Juan E. Gilbert, and Kassandra Ross. 2019. Should AI-Based, conversational digital assistants employ social- or task-oriented interaction style? A task-competency and reciprocity perspective for older adults. *Computers in Human Behavior* 90 (2019), 315 – 330. DOI: http://dx.doi.org/10.1016/j.chb.2018.08.048

[10] Justin Cheng, Jaime Teevan, Shamsi T. Iqbal, and Michael S. Bernstein. 2015. Break It Down: A Comparison of Macro- and Microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 4061–4064. DOI: http://dx.doi.org/10.1145/2702123.2702146

[11] Lydia B. Chilton, John J. Horton, Robert C. Miller, and Shiri Azenkot. 2010. Task Search in a Human Computation Market. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '10)*. ACM, New York, NY, USA, 1–9. DOI: http://dx.doi.org/10.1145/1837885.1837889

[12] Leigh Clark. 2018. Social Boundaries of Appropriate Speech in HCI: A Politeness Perspective. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference (HCI '18)*. BCS Learning & Development Ltd., Swindon, UK, Article 76, 5 pages. DOI:http://dx.doi.org/10.14236/ewic/HCI2018.76

[13] Leigh Clark, João Cabral, and Benjamin Cowan. 2018. The CogSIS Project: Examining the Cognitive Effects of Speech Interface Synthesis. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference (HCI '18)*. BCS Learning & Development Ltd., Swindon, UK, Article 169, 3 pages. DOI: http://dx.doi.org/10.14236/ewic/HCI2018.170

[14] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R. Cowan. 2019a. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* (09 2019). DOI: http://dx.doi.org/10.1093/iwc/iwz016

[15] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019b. What Makes a Good Conversation?: Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 475, 12 pages. DOI: http://dx.doi.org/10.1145/3290605.3300705

[16] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What Can I Help You with?": Infrequent Users' Experiences of Intelligent Personal Assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '17)*. ACM, New York, NY, USA, Article 43, 12 pages. DOI: http://dx.doi.org/10.1145/3098279.3098539

[17] Fred J. Damerau. 1964. A Technique for Computer Detection and Correction of Spelling Errors. *Commun. ACM* 7, 3 (1964), 171–176. DOI: http://dx.doi.org/10.1145/363958.363994

[18] 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)*. ACM, New York, NY, USA, 617–617. DOI: http://dx.doi.org/10.1145/2740908.2744109

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

[20] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2016. Scheduling Human Intelligence Tasks in Multi-Tenant Crowd-Powered Systems. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 855–865. DOI: http://dx.doi.org/10.1145/2872427.2883030

[21] Maxine Eskenazi, Gina-Anne Levow, Helen Meng, Gabriel Parent, and David Suendermann. 2013. *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*. John Wiley & Sons.

[22] Eduardo Fonseca, Jordi Pons, Xavier Favory, Frederic Font, Dmitry Bogdanov, Andrés Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. 2017. Freesound Datasets: a platform for the creation of open audio datasets. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)*. Suzhou, China, 486–493.

[23] Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. 2018. Crowd Anatomy Beyond the Good and Bad: Behavioral Traces for Crowd Worker Modeling and Pre-selection. *Computer Supported Cooperative Work (CSCW)* (Jun 2018). DOI: http://dx.doi.org/10.1007/s10606-018-9336-y

[24] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. 2014. A Taxonomy of Microtasks on the Web. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media (HT '14)*. ACM, New York, NY, USA, 218–223. DOI: http://dx.doi.org/10.1145/2631775.2631819

[25] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 776–780. DOI: http://dx.doi.org/10.1109/ICASSP.2017.7952261

[26] Jorge Goncalves, Michael Feldman, Subingqian Hu, Vassilis Kostakos, and Abraham Bernstein. 2017. Task Routing and Assignment in Crowdsourcing Based on Cognitive Abilities. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1023–1031. DOI: http://dx.doi.org/10.1145/3041021.3055128

[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. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*. ACM, New York, NY, USA, 753–762. DOI: http://dx.doi.org/10.1145/2493432.2493481

[28] Jorge Goncalves, Simo Hosio, Niels van Berkel, Furqan Ahmed, and Vassilis Kostakos. 2017. CrowdPickUp: Crowdsourcing Task Pickup in the Wild. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 51 (Sept. 2017), 22 pages. DOI: http://dx.doi.org/10.1145/3130916

[29] Jorge Goncalves, Hannu Kukka, Iván Sánchez, and Vassilis Kostakos. 2016. Crowdsourcing Queue Estimations in Situ. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 1040–1051. DOI: http://dx.doi.org/10.1145/2818048.2819997

[30] Lisa Graham. 2008. Gestalt theory in interactive media design. *Journal of Humanities & Social Sciences* 2, 1 (2008).

[31] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 449, 14 pages. DOI: http://dx.doi.org/10.1145/3173574.3174023

[32] Seyyed Hadi Hashemi, Kyle Williams, Ahmed El Kholy, Imed Zitouni, and Paul A. Crook. 2018. Measuring User Satisfaction on Smart Speaker Intelligent Assistants Using Intent Sensitive Query Embeddings. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. ACM, New York, NY, USA, 1183–1192. DOI: http://dx.doi.org/10.1145/3269206.3271802

[33] Danula Hettiachchi, Niels van Berkel, Tilman Dingler, Fraser Allison, Vassilis Kostakos, and Jorge Goncalves. 2019a. Enabling Creative Crowd Work through Smart Speakers. In *Workshop on Designing Crowd-powered Creativity Support Systems (CHI '19 Workshop)*. 1–5. DOI:http://dx.doi.org/10.5281/zenodo.2648986

[34] Danula Hettiachchi, Niels van Berkel, Simo Hosio, Vassilis Kostakos, and Jorge Goncalves. 2019b. Effect of Cognitive Abilities on Crowdsourcing Task Performance. In *Human-Computer Interaction – INTERACT 2019*. Springer International Publishing, Cham, 442–464. DOI: http://dx.doi.org/10.1007/978-3-030-29381-9_28

[35] Simo Hosio, Jorge Goncalves, Vili Lehdonvirta, Denzil Ferreira, and Vassilis Kostakos. 2014. Situated Crowdsourcing Using a Market Model. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST '14)*. ACM, New York, NY, USA, 55–64. DOI: http://dx.doi.org/10.1145/2642918.2647362

[36] Thivya Kandappu, Nikita Jaiman, Randy Tandriansyah, Archan Misra, Shih-Fen Cheng, Cen Chen, Hoong Chuin Lau, Deepthi Chander, and Koustuv Dasgupta. 2016. TASKer: Behavioral Insights via Campus-based Experimental Mobile Crowd-sourcing. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 392–402. DOI:http://dx.doi.org/10.1145/2971648.2971690

[37] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2012. The Face of Quality in Crowdsourcing Relevance Labels: Demographics, Personality and Labeling Accuracy. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 2583–2586. DOI:http://dx.doi.org/10.1145/2396761.2398697

[38] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding User Satisfaction with Intelligent Assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (CHIIR '16)*. ACM, New York, NY, USA, 121–130. DOI: http://dx.doi.org/10.1145/2854946.2854961

[39] Ahmet Baki Kocaballi, Liliana Laranjo, and Enrico Coiera. 2019. Understanding and Measuring User Experience in Conversational Interfaces. *Interacting with Computers* 31, 2 (05 2019), 192–207. DOI: http://dx.doi.org/10.1093/iwc/iwz015

[40] Arne Köhn, Florian Stegen, and Timo Baumann. 2016. Mining the Spoken Wikipedia for Speech Data and Beyond. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (23-28). European Language Resources Association (ELRA), Paris, France.

[41] Il-Youp Kwak, Jun Ho Huh, Seung Taek Han, Iljoo Kim, and Jiwon Yoon. 2019. Voice Presentation Attack Detection Through Text-Converted Voice Command Analysis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 598, 12 pages. DOI: http://dx.doi.org/10.1145/3290605.3300828

[42] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, Are You Listening?: Privacy Perceptions, Concerns and Privacy-seeking Behaviors with Smart Speakers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 102 (Nov. 2018), 31 pages. DOI:http://dx.doi.org/10.1145/3274371

[43] Adrian Leemann, Marie-José Kolly, and David Britain. 2018. The English Dialects App: The creation of a crowdsourced dialect corpus. *Ampersand* 5 (2018), 1 – 17. DOI: http://dx.doi.org/10.1016/j.amper.2017.11.001

[44] Irene Lopatovska, Katrina Rink, Ian Knight, Kieran Raines, Kevin Cosenza, Harriet Williams, Perachya Sorsche, David Hirsch, Qi Li, and Adrianna Martinez. 2019. Talk to me: Exploring user interactions with the Amazon Alexa. *Journal of Librarianship and Information Science* 51, 4 (2019), 984–997. DOI: http://dx.doi.org/10.1177/0961000618759414

[45] Michal Luria, Guy Hoffman, and Oren Zuckerman. 2017. Comparing Social Robot, Screen and Voice Interfaces for Smart-Home Control. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 580–628. DOI: http://dx.doi.org/10.1145/3025453.3025786

[46] Christine Murad, Cosmin Munteanu, Leigh Clark, and Benjamin R. Cowan. 2018. Design Guidelines for Hands-free Speech Interaction. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI '18)*. ACM, New York, NY, USA, 269–276. DOI:http://dx.doi.org/10.1145/3236112.3236149

[47] Christine Murad, Cosmin Munteanu, Benjamin R. Cowan, and Leigh Clark. 2019. Revolution or Evolution? Speech Interaction and HCI Design Guidelines. *IEEE Pervasive Computing* 18, 02 (apr 2019), 33–45. DOI: http://dx.doi.org/10.1109/MPRV.2019.2906991

[48] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 6, 7 pages. DOI: http://dx.doi.org/10.1145/3173574.3173580

[49] Chelsea M. Myers, Anushay Furqan, and Jichen Zhu. 2019. The Impact of User Characteristics and Preferences on Performance with an Unfamiliar Voice User Interface. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 47, 9 pages. DOI:http://dx.doi.org/10.1145/3290605.3300277

[50] Christi Olson and Kelli Kemery. 2019. Voice Report: From answers to action: customer adoption of voice technology and digital assistants. (2019). https://advertiseonbing-blob.azureedge.net/blob/bingads/media/insight/whitepapers/2019/04%20apr/voice-report/bingads_2019voicereport.pdf

[51] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 527–536. DOI:http://dx.doi.org/10.18653/v1/P19-1050

[52] Aung Pyae and Paul Scifleet. 2018. Investigating Differences Between Native English and Non-native English Speakers in Interacting with a Voice User Interface: A Case of Google Home. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction (OzCHI '18)*. ACM, New York, NY, USA, 548–553. DOI: http://dx.doi.org/10.1145/3292147.3292236

[53] Julie Rico. 2010. Evaluating the Social Acceptability of Multimodal Mobile Interactions. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems (CHI EA '10)*. ACM, New York, NY, USA, 2887–2890. DOI:http://dx.doi.org/10.1145/1753846.1753877

[54] Jeffrey M. Rzeszotarski and Aniket Kittur. 2011. Instrumenting the Crowd: Using Implicit Behavioral Measures to Predict Task Performance. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST '11)*. ACM, New York, NY, USA, 13–22. DOI: http://dx.doi.org/10.1145/2047196.2047199

[55] Clay Shirky. 2010. *Cognitive surplus: How technology makes consumers into collaborators*. Penguin, UK.

[56] Ben Shneiderman. 2000. The Limits of Speech Recognition. *Commun. ACM* 43, 9 (Sept. 2000), 63–65. DOI:http://dx.doi.org/10.1145/348941.348990

[57] Kinga Skorupska, Manuel Nunez, Wieslaw Kopec, and Radoslaw Nielek. 2018. Older Adults and Crowdsourcing: Android TV App for Evaluating TEDx Subtitle Quality. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 159 (Nov. 2018), 23 pages. DOI: http://dx.doi.org/10.1145/3274428

[58] David R. Thomas. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation* 27, 2 (2006), 237–246. DOI: http://dx.doi.org/10.1177/1098214005283748

[59] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. 2016. Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System Using Deep Recurrent Neural Networks. In *Interspeech*. 352–356. DOI: http://dx.doi.org/10.21437/Interspeech.2016-159

[60] Aditya Vashistha, Abhinav Garg, and Richard Anderson. 2019. ReCall: Crowdsourcing on Basic Phones to Financially Sustain Voice Forums. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 169, 13 pages. DOI: http://dx.doi.org/10.1145/3290605.3300399

[61] Aditya Vashistha, Pooja Sethi, and Richard Anderson. 2017. Respeak: A Voice-based, Crowd-powered Speech Transcription System. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1855–1866. DOI:http://dx.doi.org/10.1145/3025453.3025640

[62] Aditya Vashistha, Pooja Sethi, and Richard Anderson. 2018. BSpeak: An Accessible Voice-based Crowdsourcing Marketplace for Low-Income Blind People. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 57, 13 pages. DOI: http://dx.doi.org/10.1145/3173574.3173631

[63] Yan Yan, Romer Rosales, Glenn Fung, and Jennifer G. Dy. 2011. Active Learning from Crowds. In *Proceedings of the 28th International Conference on International Conference on Machine Learning (ICML'11)*. ACM, New York, NY, USA, 1161–1168.

Chapter 8

Discussion

The original research contributions of this thesis are presented in Chapters 4, 5, 6, and 7. This chapter reflects on these research contributions in relation to prior work and how they answer the proposed research questions outlined in Chapter 1. In addition, we detail future directions for research on crowdsourcing task assignment and provide a summary of the limitations of the studies presented in this thesis.

8.1 Worker Cognitive Ability and Crowdsourcing Data Quality

Ensuring that crowd data is of high quality is of great importance as the generated input can feed into critical research and commercial applications, including decision-making systems [165]. As a result, there is a large body of literature on data quality in crowdsourcing [27]. In Chapter 2, we dissect such data quality improvement methods into pre-execution, online methods, and post-processing methods and discuss how online task assignment methods can bring more value to the crowdsourcing process.

We also highlight that existing crowdsourcing platforms provide limited task assignment capabilities. Existing assignment methods generally fail to accomplish the desired traits for practical implementation. They are either complex to implement, not cost-effective, not able to provide significant data quality gains, or only applicable for a specific type of tasks. Thus, in this thesis, we explore novel crowdsourcing task assignment methods that can potentially overcome the aforementioned limitations.

This thesis set out to explore the utility of certain worker attributes on matching workers with crowdsourcing tasks, due to their broader applicability and scalability when compared to other approaches such as using current answer information [49, 101], worker behaviour [69, 149] and gold standard data [119]. In particular, we investigate worker cognitive ability and context. Our preference for using worker cognitive ability was motivated by several key advantages. We can objectively measure cognitive ability using cognitive tests which are fast-paced and relatively less time consuming compared to language tests and personality tests [99, 120]. It is also relatively difficult for a worker to distort cognitive test outcomes. In Chapters 4 and 5, we explore the relationship between crowdsourcing task performance and cognitive tests and aim to answer the following research question.

RQ1: How can we improve crowdsourcing data quality by assigning tasks using online cognitive tests?

First, in Chapter 4, we demonstrate that despite using a limited number of distinct trials within a test, brief online cognitive tests yield specific test effects (e.g., Stroop effect). Thus, we can reliably use them for testing worker cognitive ability in a crowdsourcing platform [25]. In addition, for repeated testing, we can use different cognitive tests that measure the same executive function (e.g., Stroop, Simon, go/no-go, and other tests for Inhibition Control [33]) to avoid workers getting familiar with specific tests.

Prior work has only reported the link between offline paper-based cognitive tests and crowdsourcing task performance [59]. In Chapter 4, we show that such performance correlations exist in an online setting. More importantly, we explain how we can use the primary executive functions of the brain (i.e., inhibition control, switching, and working memory) to build a relationship between specific crowdsourcing tasks and cognitive tests. According to Psychology literature, each cognitive test measures a specific executive function (e.g., Stroop test measure inhibition control) and crowdsourcing tasks have underlying skill requirements that can be mapped to the primary executive functions [33]. We demonstrate that our hypothesis is accurate by examining the feature importance scores of our predictive models. Unlike our method, previous attempts to use worker attributes for task matching does not establish a direct rationale between attributes and task outcomes [141], making it difficult to extend the assignment method for a broader range of tasks. Furthermore, in Chapter 5, we outline several practical considerations for using cognitive test based task assignment new types of tasks. Practitioners could use a pilot deployment, refer to literature on executive functions, or examine the task similarity with known tasks to determine the task-test relationship for new tasks.

In Chapters 4 and 5, we show that we can gain data quality improvements by assigning tasks based on cognitive test outcomes. Particularly, in Chapter 5 we report significant data quality improvements in classification, sentiment analysis, transcription and proofreading tasks when compared to a baseline assignment where workers select the tasks they like.

Another important finding that we report in Chapter 4 is the variation in worker task performance across different tasks. Our evidence strengthens the notion that separating the worker pool into 'good' and 'bad' workers is not highly suited for crowd task matching [51]. For example, such methods often unnecessarily penalise workers who are good at specific types of tasks but not generally competent across all tasks. This thesis sheds light on how to capture this underlying task-worker compatibility systematically. Furthermore, in crowdsourcing experiments presented in Chapters 4 and 5, we see that cognitive ability based task assignment can ensure that a large majority of the worker pool get assigned to at least one crowdsourcing task.

8.2 Dynamic Online Task Assignment

Determining the relationship between crowdsourcing task performance and cognitive ability, and proposing a task assignment method is not sufficient to ensure that such a method would work in an online dynamic platform. This is also a significant limitation in many previous task matching approaches, where they only propose a theoretical model and test with synthetic data (e.g., [7, 14, 77]), or propose a worker accuracy estimation

method (e.g., [65, 99]). Therefore, this thesis examines dynamic task assignment using worker cognitive skills as proposed in our second research question.

RQ2: How can we achieve dynamic online task assignment using worker cognitive ability?

In Chapter 4, we consider each task and select a subset of workers who are more likely to produce high-quality contributions based on their cognitive test outcomes. Our analysis shows a clear difference in task accuracy between the selected and remaining workers.

Extending the cognitive ability based method to a real-time dynamic task assignment scenario poses additional challenges that we consider in Chapter 5. First, workers sequentially arrive and uptake tasks. Thus, we cannot make task assignment decisions by considering all workers and feasible assignments. Second, a worker may continue to complete an arbitrary number of questions in the same task. Third, testing workers involves an upfront cost, which may negatively impact the overall task cost if many workers are not assigned to tasks. In Chapter 5, we present and evaluate 'CrowdCog', a cognitive skill based task assignment and recommendation system that aim to overcome aforementioned challenges. CrowdCog assigns a limited number of cognitive tests when a worker arrives, and attempts to match the worker immediately with an available task once the tests are completed. Instead of testing all executive functions, we greedily assign a compatible task to maximise the quality gain while balancing the testing cost. Our experimental results presented in Chapter 5 accompany a cost-analysis which demonstrates that CrowdCog can recover testing cost and produce high-quality labels in an economical manner compared to a baseline assignment.

In a dynamic task assignment scenario, CrowdCog is also capable of using different quality thresholds to cater for requester needs. If a requester aims for high data quality, they can pre-configure a higher selection threshold in CrowdCog such that only highly compatible workers are selected for a given task. Conversely, if they want to collect answers as quickly as possible with modest quality gains, they can use a lower quality threshold. In Chapter 5, we also describe various question assignment and plurality assignment methods that can improve data quality. When comparing CrowdCog to a state-of-the-art question assignment method [175] in Chapter 5, we show that accuracy gains are similar among the two methods. Unlike task assignment methods, question assignment methods rely on evaluating current answer distribution regularly as workers submit answers and are limited to specific task types. Therefore, our task assignment method that works across a broader range of tasks would be the preferred method for most cases. While it is not recommended for platform-wide implementation, practitioners could also implement question assignment together with task assignment to achieve increased output accuracy.

In addition to our task assignment method, we also present and evaluate a task recommendation method. Our results show that workers are more likely to accept a recommended task. Also, workers completed tasks more accurately when attempting a recommended task. Crowdsourcing literature argues that task assignment can restrict and

undermine crowd worker autonomy. This shortcoming can be mitigated by employing task recommendation [56].

8.3 Crowd Worker Context and Task Assignment

Through the third research question, we aim to understand and incorporate worker context information for task assignment, which is highly applicable when considering crowd tasks performed through alternative work devices, like smartphones and smart speakers.

RQ3: How can we use the context of the worker to assign tasks effectively?

In Chapter 6, we show that worker context is an important factor for task assignment. There are differences in task acceptance patterns based on the context and work device, implying that workers prefer different devices depending on their context, such as current location. For example, when accepting tasks presented through smartphones, the acceptance rate when workers are at their primary workstation is lower than when at home but away from the workstation. Similarly, task characteristics such as expected task completion time are important in matching tasks to devices. We show that workers are reluctant to accept overly long tasks (i.e., expected time is 10 minutes) in smart speakers when compared to other devices. Thus, we can provide greater flexibility to workers by making suitable tasks available in different contexts and devices. We envision a future where crowd tasks can be directed to a wider range of appropriate work devices, such as smart watches [3] and smart speakers with screens and touch interaction.

Our findings concerning crowdsourcing through digital voice assistant devices reported in Chapter 7 further strengthens the argument for cross-device crowdsourcing. Participants of our field deployment stated that they were able to launch tasks at opportune moments and multitask while completing crowd tasks, which is also supported by the exploratory analysis presented. In addition, by extending crowd work to a wide array of devices and modalities, we can increase the accessibility of crowdsourcing platforms [163]. For instance, a visually impaired person could engage in crowd work through a voice-based crowdsourcing platform that runs on a smart speaker.

We evaluate our 'CrowdTasker' application that runs on a digital voice assistant and demonstrate the feasibility of voice-based crowd work in Chapter 7. Our results show no significant difference in crowdsourcing task accuracy for native English speakers when completing tasks through voice-interface and standard web-interface. Furthermore, we report similar task performance in our field study where participants completed tasks at their home, highlighting the practicality of a voice-based crowdsourcing system.

Developing crowdsourcing platforms that work with non-standard modalities is challenging [3, 60] and there is only limited prior work that examines voice-based crowd work [164]. In Chapter 7, we discuss specific design considerations for developing a voice-based crowdsourcing platform. Several workflow adjustments can be implemented to provide more control for workers. For example, workers should be able to request the voice-assistant to repeat content at granular levels, stop and resume tasks at any point,

and skip certain content or interrupt the voice assistant while speaking. As navigating and browsing tasks through voice is more time-consuming than screen-based interaction, task assignment or recommendation is more beneficial for voice-based crowdsourcing. Particularly, with the possibility to directly start the task, it can help worker launch suitable tasks at opportune moments tapping into their cognitive surplus [155] and complete crowd tasks in brief sessions allowing them to manage crowd work better with other household commitments. Additional features such as allowing workers to review and modify their response can also help.

Designing crowdsourcing tasks for voice-based platforms is also not trivial. In Chapter 7, we argue that practitioners should carefully identify crowdsourcing tasks that work better with voice. Our analysis suggests that working memory demand and task complexity play a critical role in determining task suitability. Tasks created for voice-interaction should be concise such that workers can process information and respond without requiring to use a high amount of cognitive resources. Similarly, multiple-choice questions should have fewer response options or should be converted to binary labelling questions.

8.4 Limitations and Future Directions

When discussing the future of crowd work, Kittur et al. [105] identify task assignment as one of the key elements that can improve the value and meaning of crowd work. While task assignment has been increasingly researched in recent years, we do not see widespread adoption of task assignment strategies in commercial crowdsourcing platforms [27]. In this thesis, we show that task assignment through fast-paced cognitive tests can yield significant data quality improvements. While cognitive tests and our assignment method can be readily incorporated into crowd platforms, future work should investigate ways to integrate other generalisable quality indicators. In addition, our work highlights the feasibility of voice-based crowd work as a novel direction of flexible crowd work that appeals to both researchers and commercial platforms. Further, we show how context influences task acceptance in a cross-device scenario that warrants further investigation into the potential benefits of cross-device task assignment. In this section, we reflect on limitations with current approaches and in our work, and discuss how future research could address them to promote the practical use of task assignment.

One of the critical limitations of many task assignment methods is that they fail to work across a broader range of tasks. Thus, there is little incentive for crowdsourcing platforms to implement or facilitate such methods. Similar to our cognitive ability based task assignment method, future work could explore more generalisable methods that do not directly depend on the task. Instead of using binary or scalar quality indicators, researchers could explore and establish relationships between worker signals and different types of crowd tasks to achieve this.

Research should also focus on how to address the cold start issue in crowdsourcing task assignment. Particularly, task requesters often do not have the luxury of collecting large volumes of training data or accessing and analysing past worker records before employing an assignment method. Our cognitive ability based task assignment method

aims to mitigate this by building generic assignment models that do not rely on individual contributions. However, similar to any supervised learning method, our approach requires limited training data through pilot implementations. New methods that explore unsupervised approaches with generic models would be more favourable to requesters.

Moreover, integrating different worker accuracy estimation methods and task assignment strategies is another feasible research direction that can further improve the value and utility of assignment methods. For example,Barbosa and Chen [10] attempt to integrate worker demographics and related attributes, and show that we can improve data quality by allowing requesters to pre-specify the workforce diversity or uniformity. This thesis shows how worker cognitive ability (Chapters 4 and 5) and context (Chapter 6) are useful for task assignment. Similarly, prior research shows how personality [98] and task-specific qualification tests [130] are good indicators of worker performance. Future work could investigate how to encapsulate different test scores and attributes to provide a unified estimation of worker accuracy. A prudent strategy is to implement a test marketplace, where task requesters could publish different tests and surveys that other requesters can use.

In this thesis, we show how alternative crowd work platforms like voice-based crowdsourcing are feasible and potentially beneficial to workers. With the increasing availability of crowd tasks like audio annotation [37], rating speech data [127], virtual reality experiments [122], integrated cross-device crowdsourcing platforms would be a useful addition to the crowdsourcing ecosystem. However, future work is needed to explore and evaluate dynamic task assignment methods that involve different worker contexts and work devices.

While crowdsourcing is an effective method to harness large volumes of training data for machine learning models [165], different biases (e.g., population bias, presentation bias) can be introduced through crowdsourced data collection process [128, 138]. While biases can be identified [83] and reduced in post-processing steps such as aggregation [93], future research should explore how task assignment methods can proactively manage such biases [58].

Furthermore, due to limited features and the competitive nature in crowdsourcing platforms, workers tend to use several third-party tools to increase their productivity [96], leading to task switching behaviour and increased fragmentation in work-life balance [170]. It is important to consider worker factors, and develop approaches that can potentially help workers manage their work (e.g., task scheduling approaches that help reduce context switching [40], flexible ways of conducting crowd work as presented in Chapter 7).

Finally, fair compensation for crowd workers is another important aspect [151, 169]. However, it is not sufficient to ensure that worker earnings meet the minimum hourly pay rate, requesters and platforms need to help them minimise the idle time in between jobs. In fact, task assignment can help reduce task search time by matching workers to compatible tasks. Future work could explore and quantify how such factors are improved through task assignment. Furthermore, assignment methods should explore task matching at a more granular level [51, 73, 101] than simply identifying 'good' or 'bad' workers [149]. This will be particularly beneficial for inexperienced workers as well as others who may not be universally good at all tasks.

Chapter 9
Conclusion

This thesis investigates worker cognitive ability and context-based task assignment for improving data quality in crowdsourcing. We present insights into building a relationship between specific crowdsourcing tasks and cognitive tests using executive functions of the brain. Our crowdsourcing experiments show that task assignment through cognitive abilities leads to significant performance gains across different crowdsourcing tasks such as sentiment analysis, transcription, classification and counting. Furthermore, the thesis considers practical task assignment challenges and presents an online crowdsourcing framework that can assign and recommend tasks based on worker cognitive abilities. Task requesters can readily adapt our method to assign a broader range of crowdsourcing tasks.

With the growing availability of devices like smartphones and smart speakers, crowd workers are able to work from non-traditional setups. We also present insights into task assignment in such a cross-device crowdsourcing scenario where worker context becomes an important determinant for matching workers with relevant tasks. Particularly, results from our crowdsourcing experiment demonstrates that workers are willing to uptake various crowdsourcing tasks on smartphones and smart speakers similar to standard desktop computers. We argue that crowd platforms can provide benefits to workers by selecting appropriate tasks based on worker context and the device at hand.

Furthermore, the thesis presents and evaluates a voice-based crowdsourcing platform where workers can complete tasks by interacting with a digital voice assistant. We demonstrate that when native English speakers complete voice-compatible (e.g., sentiment analysis) and voice-based (e.g., audio annotation) crowd tasks via voice-interface, the output accuracy is not significantly different from completing tasks through a standard web-interface. We offer guidelines for creating voice-based crowdsourcing platforms and designing crowd tasks for voice-interaction.

We discuss how future crowdsourcing platforms could incorporate multiple worker performance estimation factors in a unified assignment framework, use task assignment to reduce biases in collected data, and provide seamless crowdsourcing experience to workers through cross-device task assignment. Ultimately, we anticipate that the findings presented in this thesis will lead to crowdsourcing platforms that are more accessible and flexible for workers, and more productive for task requesters.

References

[1] à Campo, S., Khan, V., Papangelis, K., and Markopoulos, P. "Community heuristics for user interface evaluation of crowdsourcing platforms". In: *Future Generation Computer Systems* vol. 95 (June 2019), pp. 775–789. DOI: 10.1016/j.future.2018.02.028.

[2] Abraham, I., Alonso, O., Kandylas, V., Patel, R., Shelford, S., and Slivkins, A. "How Many Workers to Ask?: Adaptive Exploration for Collecting High Quality Labels". In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '16. New York, NY, USA: ACM, 2016, pp. 473–482. DOI: 10.1145/2911451.2911514.

[3] Acer, U. G., Broeck, M. v. d., Forlivesi, C., Heller, F., and Kawsar, F. "Scaling Crowdsourcing with Mobile Workforce: A Case Study with Belgian Postal Service". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* vol. 3.no. 2 (June 2019), 35:1–35:32. DOI: 10.1145/3328906.

[4] Alagarai Sampath, H., Rajeshuni, R., and Indurkhya, B. "Cognitively Inspired Task Design to Improve User Performance on Crowdsourcing Platforms". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '14. New York, NY, USA: ACM, 2014, pp. 3665–3674. DOI: 10.1145/2556288.2557155.

[5] Alonso, O. and Baeza-Yates, R. "Design and implementation of relevance assessments using crowdsourcing". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 6611 LNCS. Springer Verlag, 2011, pp. 153–164. DOI: 10.1007/978-3-642-20161-5_16.

[6] Ambati, V., Vogel, S., and Carbonell, J. "Collaborative workflow for crowdsourcing translation". In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. CSCW '12. New York, New York, USA: ACM Press, 2012, p. 1191. DOI: 10.1145/2145204.2145382.

[7] Assadi, S., Hsu, J., and Jabbari, S. "Online Assignment of Heterogeneous Tasks in Crowdsourcing Markets". In: *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing*. HCOMP '15. AAAI Press, 2015.

[8] Awad, E., Dsouza, S., Bonnefon, J., Shariff, A., and Rahwan, I. "Crowdsourcing moral machines". In: *Communications of the ACM* vol. 63.no. 3 (Feb. 2020), pp. 48–55. DOI: 10.1145/3339904.

[9] Baba, Y. and Kashima, H. "Statistical Quality Estimation for General Crowdsourcing Tasks". In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '13. New York, NY, USA: ACM, 2013, pp. 554–562. DOI: 10.1145/2487575.2487600.

[10] Barbosa, N. M. and Chen, M. "Rehumanized Crowdsourcing: A Labeling Framework Addressing Bias and Ethics in Machine Learning". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. New York, NY, USA: ACM, May 2019, pp. 1–12. DOI: 10.1145/3290605.3300773.

[11] Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* vol. 67.no. 1 (2015). DOI: 10.18637/jss.v067.i01.

[12] Berkel, N. van, Goncalves, J., Hettiachchi, D., Wijenayake, S., Kelly, R. M., and Kostakos, V. "Crowdsourcing Perceptions of Fair Predictors for Machine Learning: A Recidivism Case Study". In: *Proc. ACM Hum.-Comput. Interact.* vol. 3.no. CSCW (Nov. 2019). DOI: 10.1145/3359130.

[13] Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., and Panovich, K. "Soylent: A Word Processor with a Crowd Inside". In: *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*. UIST '10. New York, NY, USA: ACM, 2010, pp. 313–322. DOI: 10.1145/1866029.1866078.

[14] Boim, R. et al. "Asking the Right Questions in Crowd Data Sourcing". In: *2012 IEEE 28th International Conference on Data Engineering*. ICDE '12. IEEE, Apr. 2012, pp. 1261–1264. DOI: 10.1109/ICDE.2012.122.

[15] Braun, V., Clarke, V., Hayfield, N., and Terry, G. "Thematic analysis". In: *Handbook of Research Methods in Health Social Sciences*. 2019. DOI: 10.1007/978-981-10-5251-4_103.

[16] Brawley, A. M. and Pury, C. L. "Work experiences on MTurk: Job satisfaction, turnover, and information sharing". In: *Computers in Human Behavior* vol. 54 (Jan. 2016). DOI: 10.1016/j.chb.2015.08.031.

[17] Cao, C. C., She, J., Tong, Y., and Chen, L. "Whom to ask? Jury selection for decision making tasks on micro-blog services". In: *Proceedings of the VLDB Endowment* vol. 5.no. 11 (July 2012), pp. 1495–1506. DOI: 10.14778/2350229.2350264.

[18] Celis, L. E., Reddy, S. P., Singh, I. P., and Vaya, S. "Assignment Techniques for Crowdsourcing Sensitive Tasks". In: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. CSCW '16. New York, NY, USA: ACM, Feb. 2016, pp. 836–847. DOI: 10.1145/2818048.2835202.

[19] Checco, A., Bates, J., and Demartini, G. "Adversarial Attacks on Crowdsourcing Quality Control". In: *Journal of Artificial Intelligence Research* (2020). DOI: 10.1613/jair.1.11332.

[20] Checco, A., Bates, J., and Demartini, G. "All That Glitters is Gold – An Attack Scheme on Gold Questions in Crowdsourcing". In: *Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing*. HCOMP '18. AAAI Press. 2018.

[21] Chen, C., Meng, X., Zhao, S., and Fjeld, M. "ReTool: Interactive microtask and workflow design through demonstration". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. New York, NY, USA: ACM, May 2017, pp. 3551–3556. DOI: 10.1145/3025453.3025969.

[22] Chen, Q., Bragg, J., Chilton, L. B., and Weld, D. S. "Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. New York, NY, USA: ACM, May 2019, pp. 1–14. DOI: 10.1145/3290605.3300761.

[23] Chen, X., Lin, Q., and Zhou, D. "Optimistic Knowledge Gradient Policy for Optimal Budget Allocation in Crowdsourcing". In: *Proceedings of the 30th International Conference on Machine Learning*. Vol. 28. Proceedings of Machine Learning Research 3. Atlanta, Georgia, USA: PMLR, 2013, pp. 64–72.

[24] Cheng, J., Teevan, J., Iqbal, S. T., and Bernstein, M. S. "Break It Down: A Comparison of Macro- and Microtasks". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. New York, New York, USA: ACM Press, 2015, pp. 4061–4064. DOI: 10.1145/2702123.2702146.

[25] Crump, M. J. C., McDonnell, J. V., and Gureckis, T. M. "Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research." In: *PLoS ONE* vol. 8.no. 3 (2013), pp. 1–18. DOI: 10.1371/journal.pone.0057410.

[26] Dai, P., Rzeszotarski, J. M., Paritosh, P., and Chi, E. H. "And Now for Something Completely Different: Improving Crowdsourcing Workflows with Micro-Diversions". In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW '15. New York, NY, USA: ACM, 2015, pp. 628–638. DOI: 10.1145/2675133.2675260.

[27] Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., and Allahbakhsh, M. "Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions". In: *ACM Computing Surveys* vol. 51.no. 1 (Apr. 2018), pp. 1–40. DOI: 10.1145/3148148.

[28] Das Sarma, A., Parameswaran, A., and Widom, J. "Towards Globally Optimal Crowdsourcing Quality Management". In: *Proceedings of the 2016 International Conference on Management of Data*. Vol. 26-June-20. SIGMOD '16. New York, New York, USA: ACM Press, 2016, pp. 47–62. DOI: 10.1145/2882903.2882953.

[29] Dawid, A. P. and Skene, A. M. "Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm". In: *Applied Statistics* (1979). DOI: 10.2307/2346806.

[30] Demartini, G. "Hybrid human–machine information systems: Challenges and opportunities". In: *Computer Networks* vol. 90 (2015), pp. 5–13. DOI: https://doi.org/10.1016/j.comnet.2015.05.018.

[31] Demartini, G., Difallah, D. E., and Cudré-Mauroux, P. "Large-scale linked data integration using probabilistic reasoning and crowdsourcing". In: *VLDB Journal* (2013). DOI: 10.1007/s00778-013-0324-z.

[32] Demartini, G., Difallah, D. E., and Cudré-Mauroux, P. "ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking". In: *Proceedings of the 21st international conference on World Wide Web*. WWW '12. New York, New York, USA: ACM Press, 2012, pp. 469–478. DOI: 10.1145/2187836.2187900.

[33] Diamond, A. "Executive Functions". In: *Annual Review of Psychology* vol. 64.no. 1 (2013), pp. 135–168. DOI: 10.1146/annurev-psych-113011-143750.

[34] Dickerson, J. P., Sankararaman, K. A., Srinivasan, A., and Xu, P. "Assigning tasks to workers based on historical data: Online task assignment with two-sided arrivals". In: *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS* vol. 1 (2018), pp. 318–326.

[35] Difallah, D., Checco, A., Demartini, G., and Cudré-Mauroux, P. "Deadline-Aware Fair Scheduling for Multi-Tenant Crowd-Powered Systems". In: *ACM Transactions on Social Computing* vol. 2.no. 1 (Feb. 2019), pp. 1–29. DOI: 10.1145/3301003.

[36] Difallah, D., Filatova, E., and Ipeirotis, P. "Demographics and Dynamics of Mechanical Turk Workers". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. WSDM '18. New York, NY, USA: ACM, Feb. 2018, pp. 135–143. DOI: 10.1145/3159652.3159661.

[37] Difallah, D. E., Catasta, M., Demartini, G., Ipeirotis, P. G., and Cudré-Mauroux, P. "The Dynamics of Micro-Task Crowdsourcing: The Case of Amazon MTurk". In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15. Geneva, Switzerland: IW3C2, 2015, pp. 238–247. DOI: 10.1145/2736277.2741685.

[38] Difallah, D. E., Demartini, G., and Cudré-Mauroux, P. "Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms". In: *CEUR Workshop Proceedings*. 2012.

[39] Difallah, D. E., Demartini, G., and Cudré-Mauroux, P. "Pick-a-crowd: Tell Me What You Like, and I'll Tell You What to Do". In: *Proceedings of the 22nd International Conference on World Wide Web*. WWW '13. New York, NY, USA: ACM, 2013, pp. 367–374. DOI: 10.1145/2488388.2488421.

[40] Difallah, D. E., Demartini, G., and Cudré-Mauroux, P. "Scheduling Human Intelligence Tasks in Multi-Tenant Crowd-Powered Systems". In: *Proceedings of the 25th International Conference on World Wide Web*. WWW '16. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2016, pp. 855–865. DOI: 10.1145/2872427.2883030.

[41] Dimara, E., Bezerianos, A., and Dragicevic, P. "Narratives in Crowdsourced Evaluation of Visualizations: A Double-Edged Sword?" In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. New York, NY, USA: ACM, 2017, pp. 5475–5484. DOI: 10.1145/3025453.3025870.

[42] Doan, A., Ramakrishnan, R., and Halevy, A. Y. "Crowdsourcing Systems on the World-Wide Web". In: *Commun. ACM* vol. 54.no. 4 (Apr. 2011), pp. 86–96. DOI: 10.1145/1924421.1924442.

[43] Dow, S. P., Kulkarni, A., Klemmer, S., and Hartmann, B. "Shepherding the Crowd Yields Better Work". In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. CSCW '12. New York, NY, USA: ACM, 2012, pp. 1013–1022. DOI: 10.1145/2145204.2145355.

[44] Downs, J. S., Holbrook, M. B., Sheng, S., and Cranor, L. F. "Are Your Participants Gaming the System?: Screening Mechanical Turk Workers". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '10. New York, NY, USA: ACM, 2010, pp. 2399–2402. DOI: 10.1145/1753326.1753688.

[45] Drapeau, R., Chilton, L. B., and Weld, D. S. "MicroTalk: Using Argumentation to Improve Crowdsourcing Accuracy". In: *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing*. HCOMP '16. AAAI Press, 2016.

[46] Eickhoff, C. "Cognitive Biases in Crowdsourcing". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. WSDM '18. New York, NY, USA: ACM, 2018, pp. 162–170. DOI: 10.1145/3159652.3159654.

[47] Eickhoff, C., Harris, C. G., Vries, A. P. de, and Srinivasan, P. "Quality Through Flow and Immersion: Gamifying Crowdsourced Relevance Assessments". In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '12. New York, NY, USA: ACM, 2012, pp. 871–880. DOI: 10.1145/2348283.2348400.

[48] Ekstrom, R. B., Dermen, D., and Harman, H. H. *Manual for kit of factor-referenced cognitive tests*. Vol. 102. Princeton, NJ, USA: Educational Testing Service, 1976.

[49] Fan, J., Li, G., Ooi, B. C., Tan, K.-l., and Feng, J. "iCrowd: An Adaptive Crowdsourcing Framework". In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. SIGMOD '15. New York, NY, USA: ACM, 2015, pp. 1015–1030. DOI: 10.1145/2723372.2750550.

[50] Gadiraju, U., Checco, A., Gupta, N., and Demartini, G. "Modus Operandi of Crowd Workers : The Invisible Role of Microtask Work Environments". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* vol. 1.no. 3 (Sept. 2017), pp. 1–29. DOI: 10.1145/3130914.

[51] Gadiraju, U., Demartini, G., Kawase, R., and Dietze, S. "Crowd Anatomy Beyond the Good and Bad: Behavioral Traces for Crowd Worker Modeling and Pre-selection". In: *Computer Supported Cooperative Work: CSCW: An International Journal* vol. 28.no. 5 (Sept. 2019), pp. 815–841. DOI: 10.1007/s10606-018-9336-y.

[52] Gadiraju, U., Fetahu, B., and Kawase, R. "Training workers for improving performance in Crowdsourcing Microtasks". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2015. DOI: 10.1007/978-3-319-24258-3_8.

[53] Gadiraju, U., Fetahu, B., Kawase, R., Siehndel, P., and Dietze, S. "Using Worker Self-Assessments for Competence-Based Pre-Selection in Crowdsourcing Microtasks". In: *ACM Transactions on Computer-Human Interaction* vol. 24.no. 4 (Aug. 2017), pp. 1–26. DOI: 10.1145/3119930.

[54] Gadiraju, U., Kawase, R., and Dietze, S. "A Taxonomy of Microtasks on the Web". In: *Proceedings of the 25th ACM Conference on Hypertext and Social Media*. HT '14. New York, NY, USA: ACM Press, 2014, pp. 218–223. DOI: 10.1145/2631775.2631819.

[55] Gadiraju, U., Yang, J., and Bozzon, A. "Clarity is a Worthwhile Quality: On the Role of Task Clarity in Microtask Crowdsourcing". In: *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. HT '17. New York, New York, USA: ACM Press, July 2017, pp. 5–14. DOI: 10.1145/3078714.3078715.

[56] Geiger, D. and Schader, M. "Personalized task recommendation in crowdsourcing information systems - Current state of the art". In: *Decision Support Systems* (2014). DOI: 10.1016/j.dss.2014.05.007.

[57] Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., and Wilmer, J. B. "Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments". In: *Psychonomic Bulletin and Review* vol. 19.no. 5 (Oct. 2012), pp. 847–857. DOI: 10.3758/s13423-012-0296-9.

[58] Goel, N. and Faltings, B. "Crowdsourcing with Fairness, Diversity and Budget Constraints". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '19. New York, NY, USA: ACM, Jan. 2019, pp. 297–304. DOI: 10.1145/3306618.3314282.

[59] Goncalves, J., Feldman, M., Hu, S., Kostakos, V., and Bernstein, A. "Task Routing and Assignment in Crowdsourcing Based on Cognitive Abilities". In: *Proceedings of the 26th International Conference on World Wide Web*. WWW '17. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2017, pp. 1023–1031. DOI: 10.1145/3041021.3055128.

[60] Goncalves, J., Ferreira, D., Hosio, S., Liu, Y., Rogstadius, J., Kukka, H., and Kostakos, V. "Crowdsourcing on the Spot: Altruistic Use of Public Displays, Feasibility, Performance, and Behaviours". In: *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. UbiComp '13. New York, New York, USA: ACM Press, 2013, p. 753. DOI: 10.1145/2493432.2493481.

[61] Goncalves, J., Hosio, S., Berkel, N. van, Ahmed, F., and Kostakos, V. "CrowdPickUp". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* vol. 1.no. 3 (Sept. 2017), pp. 1–22. DOI: 10.1145/3130916.

[62] Goncalves, J., Hosio, S., Ferreira, D., and Kostakos, V. "Game of words: Tagging places through crowdsourcing on public displays". In: *Proceedings of the 2014 conference on Designing interactive systems*. DIS '14. New York, NY, USA: ACM, June 2014, pp. 705–714. DOI: 10.1145/2598510.2598514.

[63] Goncalves, J., Hosio, S., Rogstadius, J., Karapanos, E., and Kostakos, V. "Motivating participation and improving quality of contribution in ubiquitous crowdsourcing". In: *Computer Networks* vol. 90 (Oct. 2015), pp. 34–48. DOI: 10.1016/j.comnet.2015.07.002.

[64] Goncalves, J., Pandab, P., Ferreira, D., Ghahramani, M., Zhao, G., and Kostakos, V. "Projective testing of diurnal collective emotion". In: *UbiComp 2014 - Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 2014. DOI: 10.1145/2632048.2636067.

[65] Goyal, T., Mcdonnell, T., Kutlu, M., Elsayed, T., and Lease, M. "Your Behavior Signals Your Reliability: Modeling Crowd Behavioral Traces to Ensure Quality Relevance Annotations". In: *The Sixth AAAI Conference on Human Computation and Crowdsourcing*. HCOMP '18 Hcomp. AAAI Press, 2018, pp. 41–49.

[66] Gummidi, S. R. B., Xie, X., and Pedersen, T. B. "A survey of spatial crowdsourcing". In: *ACM Transactions on Database Systems* vol. 44.no. 2 (2019). DOI: 10.1145/3291933.

[67] Guo, S., Parameswaran, A., and Garcia-Molina, H. "So Who Won?: Dynamic Max Discovery with the Crowd". In: *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. SIGMOD '12. New York, NY, USA: ACM, 2012, pp. 385–396. DOI: 10.1145/2213836.2213880.

[68] Gureckis, T. M. et al. "psiTurk: An open-source framework for conducting replicable behavioral experiments online". In: *Behavior Research Methods* vol. 48.no. 3 (Sept. 2016), pp. 829–842. DOI: 10.3758/s13428-015-0642-8.

[69] Han, S., Dai, P., Paritosh, P., and Huynh, D. "Crowdsourcing Human Annotation on Web Page Structure". In: *ACM Transactions on Intelligent Systems and Technology* vol. 7.no. 4 (Apr. 2016), pp. 1–25. DOI: 10.1145/2870649.

[70] Hara, K., Le, V., and Froehlich, J. "Combining Crowdsourcing and Google Street View to Identify Street-Level Accessibility Problems". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '13. New York, NY, USA: ACM, 2013, pp. 631–640. DOI: 10.1145/2470654.2470744.

[71] Hassani, A., Haghighi, P. D., and Jayaraman, P. P. "Context-Aware Recruitment Scheme for Opportunistic Mobile Crowdsensing". In: *2015 IEEE 21st International Conference on Parallel and Distributed Systems*. ICPADS '15. IEEE, Dec. 2015, pp. 266–273. DOI: 10.1109/ICPADS.2015.41.

[72] Hettiachchi, D., Berkel, N. van, Hosio, S., Kostakos, V., and Goncalves, J. "Effect of Cognitive Abilities on Crowdsourcing Task Performance". In: *Human-Computer Interaction – INTERACT 2019*. Cham: Springer International Publishing, 2019, pp. 442–464. DOI: 10.1007/978-3-030-29381-9_28.

[73] Hettiachchi, D., Berkel, N. van, Kostakos, V., and Goncalves, J. "CrowdCog: A Cognitive Skill based System for Heterogeneous Task Assignment and Recommendation in Crowdsourcing". In: *Proceedings of the ACM on Human-Computer Interaction* vol. 4.no. CSCW2 (Oct. 2020), pp. 1–22. DOI: 10.1145/3415181.

[74] Hettiachchi, D. and Goncalves, J. "Towards Effective Crowd-Powered Online Content Moderation". In: *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*. New York, NY, USA: ACM, Dec. 2019, pp. 342–346. DOI: 10.1145/3369457.3369491.

[75] Hettiachchi, D., Sarsenbayeva, Z., Allison, F., Berkel, N. van, Dingler, T., Marini, G., Kostakos, V., and Goncalves, J. ""Hi! I am the Crowd Tasker" Crowdsourcing through Digital Voice Assistants". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. New York, NY, USA: ACM Press, 2020. DOI: 10.1145/3313831.3376320.

[76] Hettiachchi, D., Wijenayake, S., Hosio, S., Kostakos, V., and Goncalves, J. "How Context Influences Cross-Device Task Acceptance in Crowd Work". In: *Proceedings of the Eighth AAAI Conference on Human Computation and Crowdsourcing*. HCOMP'20. AAAI Press, 2020, pp. 53–62.

[77] Ho, C. J., Jabbari, S., and Vaughan, J. W. "Adaptive task assignment for crowdsourced classification". In: *30th International Conference on Machine Learning*. ICML '13 vol. 28.no. Part 1 (2013), pp. 534–542.

[78] Ho, C. J. and Vaughan, J. W. "Online task assignment in crowdsourcing markets". In: *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence Online*. Vol. 1. AAAI Press, 2012, pp. 45–51.

[79] Hosio, S., Goncalves, J., Kostakos, V., and Riekki, J. "Crowdsourcing public opinion using urban pervasive technologies: Lessons from real-life experiments in Oulu". In: *Policy and Internet* (2015). DOI: 10.1002/poi3.90.

[80] Hosio, S., Goncalves, J., Lehdonvirta, V., Ferreira, D., and Kostakos, V. "Situated crowdsourcing using a market model". In: *Proceedings of the 27th annual ACM symposium on User interface software and technology*. UIST '14. New York, NY, USA: ACM, Oct. 2014, pp. 55–64. DOI: 10.1145/2642918.2647362.

[81] Howe, J. "The Rise of Crowdsourcing". In: *Wired Magazine* (2006). DOI: 10.1086/599595.

[82] Hsueh, P.-Y., Melville, P., and Sindhwani, V. "Data quality from crowdsourcing: a study of annotation selection criteria". In: *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*. 2009, pp. 27–35.

[83] Hu, X. et al. "Crowdsourcing Detection of Sampling Biases in Image Datasets". In: *Proceedings of The Web Conference 2020*. WWW '20. New York, NY, USA: ACM, Apr. 2020, pp. 2955–2961. DOI: 10.1145/3366423.3380063.

[84] Huang, S.-W. and Fu, W.-T. "Enhancing Reliability Using Peer Consistency Evaluation in Human Computation". In: *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. Ed. by Dasgupta, S. and McAllester, D. Vol. 28. CSCW '13 4. New York, NY, USA: ACM, Jan. 2013, pp. 639–648. DOI: 10.1145/2441776.2441847.

[85] Hung, N. Q. V., Thang, D. C., Weidlich, M., and Aberer, K. "Minimizing Efforts in Validating Crowd Answers". In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. SIGMOD '15. New York, NY, USA: ACM, 2015, pp. 999–1014. DOI: 10.1145/2723372.2723731.

[86] Ikeda, K. and Hoashi, K. "Crowdsourcing GO: Effect of worker situation on mobile crowdsourcing performance". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. 2017, pp. 1142–1153. DOI: 10.1145/3025453.3025917.

[87] Ikeda, K., Morishima, A., Rahman, H., Roy, S. B., Thirumuruganathan, S., Amer-Yahia, S., and Das, G. "Collaborative crowdsourcing with Crowd4U". In: *Proceedings of the VLDB Endowment* vol. 9.no. 13 (2015), pp. 1497–1500. DOI: 10.14778/3007263.3007293.

[88] Ipeirotis, P. G. and Gabrilovich, E. "Quizz: targeted crowdsourcing with a billion (potential) users". In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW '14. ACM. 2014, pp. 143–154.

[89] Irani, L. C. and Silberman, M. S. "Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '13. New York, NY, USA: ACM, Apr. 2013, pp. 611–620. DOI: 10.1145/2470654.2470742.

[90] Joglekar, M., Garcia-Molina, H., and Parameswaran, A. "Comprehensive and reliable crowd assessment algorithms". In: *Proceedings - International Conference on Data Engineering*. ICDE '15. IEEE, 2015. DOI: 10.1109/ICDE.2015.7113284.

[91] John, O. P., Naumann, L. P., and Soto, C. J. "Paradigm shift to the integrative big five trait taxonomy". In: *Handbook of personality: Theory and research* vol. 3.no. 2 (2008), pp. 114–158.

[92] Kairam, S. and Heer, J. "Parting Crowds: Characterizing Divergent Interpretations in Crowdsourced Annotation Tasks". In: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. CSCW '16. New York, New York, USA: ACM Press, 2016, pp. 1635–1646. DOI: 10.1145/2818048.2820016.

[93] Kamar, E., Kapoo, A., and Horvitz, E. "Identifying and Accounting for Task-Dependent Bias in Crowdsourcing". In: *Proceedings, The Third AAAI Conference on Human Computation and Crowdsourcing*. HCOMP '15. AAAI Press, 2015.

[94] Kamhoua, G. A., Pissinou, N., Iyengar, S. S., Beltran, J., Miller, J., Kamhoua, C. A., and Njilla, L. L. "Approach to detect non-adversarial overlapping collusion in crowdsourcing". In: *2017 IEEE 36th International Performance Computing and Communications Conference, IPCCC 2017*. 2018. DOI: 10.1109/PCCC.2017.8280462.

[95] Kang, Q. and Tay, W. P. "Sequential Multi-class Labeling in Crowdsourcing: A Ulam-renyi Game Approach". In: *Proceedings of the International Conference on Web Intelligence*. WI '17. New York, NY, USA: ACM, 2017, pp. 245–251. DOI: 10.1145/3106426.3106446.

[96] Kaplan, T., Saito, S., Hara, K., and Bigham, J. "Striving to earn more: a survey of work strategies and tool use among crowd workers". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 6. HCOMP '18 1. 2018.

[97] Karger, D. R., Oh, S., and Shah, D. "Efficient crowdsourcing for multi-class labeling". In: *Performance Evaluation Review* vol. 41.no. 1 SPEC. ISS. (2013), pp. 81–92. DOI: 10.1145/2494232.2465761.

[98] Kazai, G., Kamps, J., and Milic-Frayling, N. "The face of quality in crowdsourcing relevance labels". In: *Proceedings of the 21st ACM international conference on Information and knowledge management*. CIKM '12. New York, New York, USA: ACM Press, 2012, p. 2583. DOI: 10.1145/2396761.2398697.

[99] Kazai, G., Kamps, J., and Milic-Frayling, N. "Worker Types and Personality Traits in Crowdsourcing Relevance Labels". In: *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. CIKM '11. New York, NY, USA: ACM, 2011, pp. 1941–1944. DOI: 10.1145/2063576.2063860.

[100] Kazai, G. and Zitouni, I. "Quality Management in Crowdsourcing using Gold Judges Behavior". In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. WSDM '16. New York, New York, USA: ACM Press, 2016, pp. 267–276. DOI: 10.1145/2835776.2835835.

[101] Khan, A. R. and Garcia-Molina, H. "CrowdDQS: Dynamic Question Selection in Crowdsourcing Systems". In: *Proceedings of the 2017 ACM International Conference on Management of Data*. SIGMOD '17. New York, NY, USA: ACM, 2017, pp. 1447–1462. DOI: 10.1145/3035918.3064055.

[102] KhudaBukhsh, A. R., Carbonell, J. G., and Jansen, P. J. "Detecting Non-Adversarial Collusion in Crowdsourcing". In: *Second AAAI Conference on Human Computation and Crowdsourcing*. HCOMP '14. AAAI Press, 2014.

[103] Kittur, A., Chi, E. H., and Suh, B. "Crowdsourcing User Studies with Mechanical Turk". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '08. New York, NY, USA: ACM, 2008, pp. 453–456. DOI: 10.1145/1357054.1357127.

[104] Kittur, A., Khamkar, S., André, P., and Kraut, R. "CrowdWeaver: Visually managing complex crowd work". In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. CSCW '12. New York, New York, USA: ACM Press, 2012, p. 1033. DOI: 10.1145/2145204.2145357.

[105] Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., and Horton, J. "The future of crowd work". In: *Proceedings of the 2013 conference on Computer supported cooperative work*. CSCW '13. New York, New York, USA: ACM Press, 2013, p. 1301. DOI: 10.1145/2441776.2441923.

[106] Kittur, A., Smus, B., Khamkar, S., and Kraut, R. E. "CrowdForge: Crowdsourcing complex work". In: *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. UIST '11. 2011. DOI: 10.1145/2047196.2047202.

[107] Kobren, A., Tan, C. H., Ipeirotis, P. G., and Gabrilovich, E. "Getting More for Less: Optimized Crowdsourcing with Dynamic Tasks and Goals". In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15. New York, New York, USA: ACM Press, 2015, pp. 592–602. DOI: 10.1145/2736277.2741681.

[108] Kristof, A. L. "Person-organization fit: an integrative review of its conceptualizations, measurement, and implications." In: *Personnel Psychology* vol. 49.no. 1 (1996), pp. 1–49.

[109] Kuang, L., Zhang, H., Shi, R., Liao, Z., and Yang, X. "A spam worker detection approach based on heterogeneous network embedding in crowdsourcing platforms". In: *Computer Networks* vol. 183 (Dec. 2020), p. 107587. DOI: 10.1016/j.comnet.2020.107587.

[110] Kulkarni, A., Can, M., and Hartmann, B. "Collaboratively Crowdsourcing Workflows with Turkomatic". In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. CSCW '12. New York, NY, USA: ACM, 2012, pp. 1003–1012. DOI: 10.1145/2145204.2145354.

[111] Kumai, K., Matsubara, M., Shiraishi, Y., Wakatsuki, D., Zhang, J., Shionome, T., Kitagawa, H., and Morishima, A. "Skill-and-Stress-Aware Assignment of Crowd-Worker Groups to Task Streams". In: *Sixth AAAI Conference on Human Computation and Crowdsourcing*. HCOMP '18. AAAI Press, 2018, pp. 88–97.

[112] Lazar, J., Feng, J. H., and Hochheiser, H. *Research Methods in Human-Computer Interaction*. 2017. DOI: 10.1016/b978-0-444-70536-5.50047-6.

[113] Le, J., Edmonds, A., Hester, V., and Biewald, L. "Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution". In: *Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation* (2010).

[114] Li, G., Wang, J., Zheng, Y., and Franklin, M. J. "Crowdsourced Data Management: A Survey". In: *IEEE Transactions on Knowledge and Data Engineering* vol. 28.no. 9 (Sept. 2016), pp. 2296–2319. DOI: 10.1109/TKDE.2016.2535242.

[115] Li, H., Zhao, B., and Fuxman, A. "The Wisdom of Minority: Discovering and Targeting the Right Group of Workers for Crowdsourcing". In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW '14. New York, NY, USA: ACM, 2014, pp. 165–176. DOI: 10.1145/2566486.2568033.

[116] Little, G., Chilton, L. B., Goldman, M., and Miller, R. C. "Exploring iterative and parallel human computation processes". In: *Workshop Proceedings - Human Computation Workshop 2010*. HCOMP '10. 2010. DOI: 10.1145/1837885.1837907.

[117] Liu, Q., Ihler, A. T., and Steyvers, M. "Scoring Workers in Crowdsourcing: How Many Control Questions are Enough?" In: *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc., 2013, pp. 1914–1922.

[118] Liu, Q., Peng, J., and Ihler, A. T. "Variational Inference for Crowdsourcing". In: *Advances in Neural Information Processing Systems*. Vol. 25. 2012, pp. 692–700.

[119] Liu, X., Lu, M., Ooi, B. C., Shen, Y., Wu, S., and Zhang, M. "CDAS: a crowdsourcing data analytics system". In: *Proceedings of the VLDB Endowment* vol. 5.no. 10 (June 2012), pp. 1040–1051. DOI: 10.14778/2336664.2336676.

[120] Lykourentzou, I., Antoniou, A., Naudet, Y., and Dow, S. P. "Personality Matters: Balancing for Personality Types Leads to Better Outcomes for Crowd Teams". In: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. CSCW '16. New York, NY, USA: ACM, 2016, pp. 260–273. DOI: 10.1145/2818048.2819979.

[121] Ma, F. et al. “FaitCrowd: Fine Grained Truth Discovery for Crowdsourced Data Aggregation”. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’15. New York, NY, USA: ACM, 2015, pp. 745–754. DOI: 10.1145/2783258.2783314.

[122] Ma, X., Cackett, M., Park, L., Chien, E., and Naaman, M. “Web-Based VR Experiments Powered by the Crowd”. In: *Proceedings of the 2018 World Wide Web Conference*. WWW ’18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, pp. 33–43.

[123] Manam, V. K. C. and Quinn, A. J. “WingIt: Efficient refinement of unclear task instructions”. In: *Sixth AAAI Conference on Human Computation and Crowdsourcing*. Vol. 6. HCOMP ’18 1. AAAI Press, 2018.

[124] Mao, A., Kamar, E., Chen, Y., Horvitz, E., Schwamb, M. E., Lintott, C. J., and Smith, A. M. “Volunteering Versus Work for Pay: Incentives and Tradeoffs in Crowdsourcing”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 1. HCOMP ’13 1. AAAI Press, Nov. 2013.

[125] Marston, W. M. *Emotions of normal people*. Vol. 158. Routledge, 2013.

[126] Mavridis, P., Gross-Amblard, D., and Miklós, Z. “Using Hierarchical Skills for Optimized Task Assignment in Knowledge-Intensive Crowdsourcing”. In: *Proceedings of the 25th International Conference on World Wide Web*. WWW ’16. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, Apr. 2016, pp. 843–853. DOI: 10.1145/2872427.2883070.

[127] McAllister Byun, T., Halpin, P. F., and Szeredi, D. “Online crowdsourcing for efficient rating of speech: A validation study”. In: *Journal of Communication Disorders* (2015). DOI: 10.1016/j.jcomdis.2014.11.003.

[128] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. *A survey on bias and fairness in machine learning*. 2019.

[129] Miles M.B & Huberman, A. *An expanded sourcebook: Qualitative data analysis (2nd Edition)*. Sage Publications, 1994.

[130] Mitra, T., Hutto, C., and Gilbert, E. “Comparing Person- and Process-centric Strategies for Obtaining Quality Data on Amazon Mechanical Turk”. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI ’15. New York, New York, USA: ACM Press, 2015, pp. 1345–1354. DOI: 10.1145/2702123.2702553.

[131] Mizusawa, K., Tajima, K., Matsubara, M., Amagasa, T., and Morishima, A. “Efficient Pipeline Processing of Crowdsourcing Workflows”. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, Oct. 2018, pp. 1559–1562. DOI: 10.1145/3269206.3269292.

[132] Mo, K., Zhong, E., and Yang, Q. “Cross-task Crowdsourcing”. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’13. New York, NY, USA: ACM, 2013, pp. 677–685. DOI: 10.1145/2487575.2487593.

[133] Mo, L., Cheng, R., Kao, B., Yang, X. S., Ren, C., Lei, S., Cheung, D. W., and Lo, E. “Optimizing Plurality for Human Intelligence Tasks”. In: *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. CIKM ’13 October. New York, NY, USA: ACM, 2013, pp. 1929–1938. DOI: 10.1145/2505515.2505755.

[134] Morschheuser, B., Hamari, J., and Koivisto, J. “Gamification in Crowdsourcing: A Review”. In: *2016 49th Hawaii International Conference on System Sciences (HICSS)*. IEEE, Jan. 2016, pp. 4375–4384. DOI: 10.1109/HICSS.2016.543.

[135] Moshfeghi, Y., Huertas-Rosero, A. F., and Jose, J. M. “Identifying Careless Workers in Crowdsourcing Platforms”. In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. SIGIR ’16. New York, NY, USA: ACM, July 2016, pp. 857–860. DOI: 10.1145/2911451.2914756.

[136] Musthag, M. and Ganesan, D. “Labor Dynamics in a Mobile Micro-task Market”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’13. New York, NY, USA: ACM, 2013, pp. 641–650. DOI: 10.1145/2470654.2470745.

[137] Oleson, D., Sorokin, A., Laughlin, G., Hester, V., Le, J., and Biewald, L. “Programmatic gold: Targeted and scalable quality assurance in crowdsourcing”. In: *AAAI Workshop - Technical Report*. 2011.

[138] Olteanu, A., Castillo, C., Diaz, F., and Kıcıman, E. “Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries”. In: *Frontiers in Big Data* (2019). DOI: 10.3389/fdata.2019.00013.

[139] Papadopoulos, D. P., Uijlings, J. R. R., Keller, F., and Ferrari, V. “Extreme Clicking for Efficient Object Annotation”. In: *2017 IEEE International Conference on Computer Vision*. ICCV ’17. IEEE, 2017, pp. 4940–4949. DOI: 10.1109/ICCV.2017.528.

[140] Park, S., Shoemark, P., and Morency, L. “Toward Crowdsourcing Micro-Level Behavior Annotations: The Challenges of Interface, Training, and Generalization”. In: *Proceedings of the 19th International Conference on Intelligent User Interfaces*. IUI ’14. New York, NY, USA: ACM, 2014, pp. 37–46. DOI: 10.1145/2557500.2557512.

[141] Peer, E., Vosgerau, J., and Acquisti, A. “Reputation as a sufficient condition for data quality on Amazon Mechanical Turk”. In: *Behavior research methods* vol. 46.no. 4 (Dec. 2014), pp. 1023–1031. DOI: 10.3758/s13428-013-0434-y.

[142] Qiu, C., Squicciarini, A. C., Carminati, B., Caverlee, J., and Khare, D. R. “CrowdSelect: Increasing Accuracy of Crowdsourcing Tasks through Behavior Prediction and User Selection”. In: CIKM ’16 (2016), pp. 539–548. DOI: 10.1145/2983323.2983830.

[143] Rangi, A. and Franceschetti, M. “Multi-armed bandit algorithms for crowdsourcing systems with online estimation of workers’ ability”. In: *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems*. AAMAS ’18. 2018, pp. 1345–1352.

[144] Raykar, V. C. and Yu, S. “Eliminating spammers and ranking annotators for crowdsourced labeling tasks”. In: *Journal of Machine Learning Research* vol. 13.no. Feb (2012), pp. 491–518.

[145] Raykar, V. C., Yu, S., Zhao, L. H., Jerebko, A., Florin, C., Valadez, G. H., Bogoni, L., and Moy, L. “Supervised Learning from Multiple Experts: Whom to Trust When Everyone Lies a Bit”. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML ’09. New York, NY, USA: ACM Press, 2009, pp. 889–896. DOI: 10.1145/1553374.1553488.

[146] Retelny, D., Bernstein, M. S., and Valentine, M. A. “No Workflow Can Ever Be Enough”. In: *Proceedings of the ACM on Human-Computer Interaction* vol. 1.no. CSCW (Dec. 2017), pp. 1–23. DOI: 10.1145/3134724.

[147] Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J., and Vukovic, M. “An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets.” In: *Proceedings of the Fifth International AAAI Conference on Web and Social Media*. Vol. 11. ICWSM ’11. California, USA: AAAI, 2011, pp. 17–21.

[148] Ross, J. and Tomlinson, B. “Who are the Crowdworkers? Shifting Demographics in Mechanical Turk”. In: *CHI ’10 Extended Abstracts on Human Factors in Computing Systems*. 2010, pp. 2863–2872.

[149] Rzeszotarski, J. M. and Kittur, A. “Instrumenting the Crowd: Using Implicit Behavioral Measures to Predict Task Performance”. In: *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. UIST ’11. New York, NY, USA: ACM, 2011, pp. 13–22. DOI: 10.1145/2047196.2047199.

[150] Saberi, M., Hussain, O. K., and Chang, E. “An online statistical quality control framework for performance management in crowdsourcing”. In: *Proceedings of the International Conference on Web Intelligence*. WI ’17. New York, New York, USA: ACM Press, 2017, pp. 476–482. DOI: 10.1145/3106426.3106436.

[151] Salehi, N., Irani, L. C., Bernstein, M. S., Alkhatib, A., Ogbe, E., Milland, K., and Clickhappier. "We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI'15. ACM. New York, NY, USA: ACM, Apr. 2015, pp. 1621–1630. DOI: 10.1145/2702123.2702508.

[152] Schaekermann, M., Joslin, G. O., Larson, K., and Edith, L. A. "Resolvable vs. Irresolvable disagreement: A study on worker deliberation in crowd work". In: *Proceedings of the ACM on Human-Computer Interaction* (2018). DOI: 10.1145/3274423.

[153] Schmitz, H. and Lykourentzou, I. "Online Sequencing of Non-Decomposable Macrotasks in Expert Crowdsourcing". In: *ACM Transactions on Social Computing* vol. 1.no. 1 (2018), pp. 1–33. DOI: 10.1145/3140459.

[154] Shaw, A. D., Horton, J. J., and Chen, D. L. "Designing incentives for inexpert human raters". In: *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. CSCW '11. New York, New York, USA: ACM Press, 2011, p. 275. DOI: 10.1145/1958824.1958865.

[155] Shirky, C. *Cognitive surplus: How technology makes consumers into collaborators*. UK: Penguin, 2010.

[156] Siddharthan, A., Lambin, C., Robinson, A. M., Sharma, N., Comont, R., O'mahony, E., Mellish, C., and Van Der Wal, R. "Crowdsourcing without a crowd: Reliable online species identification using Bayesian models to minimize crowd size". In: *ACM Transactions on Intelligent Systems and Technology* vol. 7.no. 4 (2016). DOI: 10.1145/2776896.

[157] Singer, Y. and Mittal, M. "Pricing mechanisms for crowdsourcing markets". In: *Proceedings of the 22nd international conference on World Wide Web*. WWW '13. New York, New York, USA: ACM Press, 2013, pp. 1157–1166. DOI: 10.1145/2488388.2488489.

[158] Stol, K. J. and Fitzgerald, B. "Two's company, three's a crowd: A case study of crowdsourcing software development". In: *Proceedings - International Conference on Software Engineering*. 1. IEEE Computer Society, May 2014, pp. 187–198. DOI: 10.1145/2568225.2568249.

[159] Su, H., Deng, J., and Fei-Fei, L. "Crowdsourcing annotations for visual object detection". In: *AAAI Workshop - Technical Report*. 2012.

[160] Tong, Y., Zhou, Z., Zeng, Y., Chen, L., and Shahabi, C. "Spatial crowdsourcing: a survey". In: *The VLDB Journal* vol. 29.no. 1 (Jan. 2020). DOI: 10.1007/s00778-019-00568-7.

[161] Tran-Thanh, L., Dong Huynh, T., Rosenfeld, A., Ramchurn, S. D., and Jennings, N. R. "Crowdsourcing complex workflows under budget constraints". In: *Proceedings of the National Conference on Artificial Intelligence*. AAAI '15. AAAI Press, 2015, pp. 1298–1304.

[162] Tu, J., Cheng, P., and Chen, L. "Quality-Assured Synchronized Task Assignment in Crowdsourcing". In: *IEEE Transactions on Knowledge and Data Engineering* vol. 4347.no. c (2019), pp. 1–1. DOI: 10.1109/tkde.2019.2935443.

[163] Vashistha, A., Sethi, P., and Anderson, R. "BSpeak: An Accessible Voice-based Crowdsourcing Marketplace for Low-Income Blind People". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. New York, NY, USA: ACM, 2018, 57:1–57:13. DOI: 10.1145/3173574.3173631.

[164] Vashistha, A., Sethi, P., and Anderson, R. "Respeak: A Voice-based, Crowd-powered Speech Transcription System". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. New York, NY, USA: ACM, 2017, pp. 1855–1866. DOI: 10.1145/3025453.3025640.

[165] Vaughan, J. W. "Making better use of the crowd: How crowdsourcing can advance machine learning research". In: *The Journal of Machine Learning Research* vol. 18.no. 1 (2017), pp. 7026–7071.

[166] Venanzi, M., Guiver, J., Kazai, G., Kohli, P., and Shokouhi, M. "Community-based Bayesian aggregation models for crowdsourcing". In: *WWW 2014 - Proceedings of the 23rd International Conference on World Wide Web*. New York, New York, USA: ACM Press, Apr. 2014, pp. 155–164. DOI: 10.1145/2566486.2567989.

[167] Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., and Movellan, J. "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise". In: *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*. 2009.

[168] Whiting, M. E. et al. "Crowd Guilds: Worker-Led Reputation and Feedback on Crowdsourcing Platforms". In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '17. New York, NY, USA: ACM, 2017, pp. 1902–1913. DOI: 10.1145/2998181.2998234.

[169] Whiting, M. E., Hugh, G., and Bernstein, M. S. "Fair Work: Crowd Work Minimum Wage with One Line of Code". In: *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing*. Vol. 7. HCOMP 1. 2019, pp. 197–206.

[170] Williams, A. C., Mark, G., Milland, K., Lank, E., and Law, E. "The perpetual work life of crowdworkers: How tooling practices increase fragmentation in crowdwork". In: *Proceedings of the ACM on Human-Computer Interaction* vol. 3.no. CSCW (2019). DOI: 10.1145/3359126.

[171] Yuan, D., Li, G., Li, Q., and Zheng, Y. "Sybil defense in crowdsourcing platforms". In: *International Conference on Information and Knowledge Management, Proceedings*. 2017. DOI: 10.1145/3132847.3133039.

[172] Zhao, Z., Cheng, J., Wei, F., Zhou, M., Ng, W., and Wu, Y. "SocialTransfer: Transferring social knowledge for cold-start crowdsourcing". In: *CIKM 2014 - Proceedings of the 2014 ACM International Conference on Information and Knowledge Management* (2014), pp. 779–788. DOI: 10.1145/2661829.2661871.

[173] Zheng, Y., Li, G., and Cheng, R. "DOCS: Domain-aware crowdsourcing system using knowledge bases". In: *Proceedings of the VLDB Endowment* vol. 10.no. 4 (2016), pp. 361–372. DOI: 10.14778/3025111.3025118.

[174] Zheng, Y., Li, G., Li, Y., Shan, C., and Cheng, R. "Truth inference in crowdsourcing". In: *Proceedings of the VLDB Endowment* vol. 10.no. 5 (Jan. 2017), pp. 541–552. DOI: 10.14778/3055540.3055547.

[175] Zheng, Y., Wang, J., Li, G., Cheng, R., and Feng, J. "QASCA: A Quality-Aware Task Assignment System for Crowdsourcing Applications". In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. SIGMOD '15. New York, NY, USA: ACM, 2015, pp. 1031–1046. DOI: 10.1145/2723372.2749430.

[176] Zhou, D., Platt, J. C., Basu, S., and Mao, Y. *Learning from the Wisdom of Crowds by Minimax Entropy*. Tech. rep. 2012, pp. 2195–2203.

[177] Zhu, H., Dow, S. P., Kraut, R. E., and Kittur, A. "Reviewing Versus Doing: Learning and Performance in Crowd Assessment". In: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW '14. New York, NY, USA: ACM, 2014, pp. 1445–1455. DOI: 10.1145/2531602.2531718.

[178] Zhuang, H., Parameswaran, A., Roth, D., and Han, J. "Debiasing Crowdsourced Batches". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15. New York, NY, USA: ACM, 2015, pp. 1593–1602. DOI: 10.1145/2783258.2783316.

[179] Zhuang, M. and Gadiraju, U. "In What Mood Are You Today? An Analysis of Crowd Workers' Mood, Performance and Engagement Mengdie". In: *Proceedings of the 10th ACM Conference on Web Science*. WebSci '19. New York, New York, USA: ACM Press, 2019, pp. 373–382. DOI: 10.1145/3292522.3326010.