IT'S GETTING CROWDED! IMPROVING THE EFFECTIVENESS OF MICROTASK CROWDSOURCING

Von der Fakultät für Elektrotechnik und Informatik
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des Grades

A dissertation submitted to Leibniz Universität Hannover

fulfilling the requirements for the degree of

DOKTOR DER NATURWISSENSCHAFTEN

Dr. rer. nat.

genehmigte Dissertation von

M. Sc. Ujwal Kumar Gadiraju

geboren am 7. October 1988, in Abu Dhabi, UAE

Hannover, Deutschland,
2017

Referent: Prof. Dr. techn. Wolfgang Nejdl
Ko-Referent: Prof. Dr. Gianluca Demartini
Ko-Referent: Prof. Dr. Michael Rohs
Tag der Promotion: 24.10.2017

ZUSAMMENFASSUNG

Microtask Crowdsourcing hat sich als gut geeignete Methode zur Erwerbung von menschlichem Input auf Abruf hervorgetan und findet verbreitete Anwendung für die Lösung zahlreicher Probleme. Bekannte Beispiele sind unter anderem Umfragen und das Erstellen von Inhalten und Bildbeschreibungen. Dabei haben sich eine Reihe von Herausforderungen aufgetan, die überwunden werden müssen, um das wahre Potential dieses Modells zu nutzen. In dieser Arbeit identifizieren wir drei wesentliche Herausforderungen, mit dem Ziel, die Effektivität von Microtask Crowdsourcing zu verbessern. Die erste Herausforderung ist das bislang begrenzte Verständnis von Crowdsourcing-Aufgaben und Crowdworker-Charakteristiken. Das Verständnis der Dynamiken von Crowdsourcing-Aufgaben und das Verhalten von Crowdworkern, die zu Aufgaben beitragen, kann eine zentrale Rolle für das effektive Aufgabendesign spielen. Zweitens sind aktuelle Mechanismen zur Vorauswahl von Crowdworkern simplifizierend und unzulänglich. Das Auswählen von Crowdworkern mit den gewünschten Fähigkeiten, ohne Wissen über die Qualität der vergangenen Arbeit des Workers, ist eine Herausforderung. Benötigt werden aussagekräftigere Indikatoren für die Kompetenz von Crowdworkern sowie effektivere Mechanismen zur Vorauswahl. Drittens ist die Betrachtung von Faktoren, die die Qualität der in Microtasks geleisteten Arbeit beeinflussen bislang unvollständig. Daher ist es wichtig, die unterschiedlichen Aspekte, die Crowd Work beeinflussen, umfassend zu verstehen. In dieser Dissertation stellen wir uns den aufgezeigten Herausforderungen und zeigen neue Methoden, um existierende Probleme zu überwinden. Unsere zentralen Beiträge sind im Folgenden beschrieben.

- *Erweitern des aktuellen Verständnisses von Aufgabentypen, Worker-Verhalten und Qualitätskontrolle* — Unsere Ergebnisse einer umfassenden Studie mit 1.000 Crowdworkern auf der CrowdFlower-Plattform erweitern das aktuelle Verständnis von Crowdsourcing-Microtasks und dazugehörigem Crowdworker-Verhalten. Wir stellen ein zwei-Ebenen-Kategorisierungsschema für Microtasks vor und geben Einblicke in die Affinität von Workern bezüglich bestimmter Aufgaben, den Aufwand, der für das Abschlieend von Aufgaben verschiedener Typen nötig ist und die Zufriedenheit von Workern in Hinblick auf den finanziellen Anreiz. Wir haben die verbreiteten destruktiven Aktivitäten auf Crowdsourcing-Plattformen analysiert und das Verhalten von vertrauenswürdigen und nicht vertrauenswürdigen Crowdworkern studiert, insbesondere im Hinblick auf Umfragen. Um die Qualität der Ergebnisse insgesamt zu verbessern, haben wir verhaltensspezifische Metriken vorgestellt, die genutzt werden können, um ungewollte und potentiell

destruktive Aktivitäten in Crowsourcing-Aufgaben zu messen und zu unterbinden. Auf Basis dieser Aspekte haben wir Richtlinien für das effektive Design von Crowdsourcing-Aufgaben festgelegt.

- *Neue Mechanismen für die Vorauswahl von Crowdworkern* — Wir haben zwei verschiedene neue Methoden zur Vorauswahl von Crowdworkern vorgestellt. Die Methoden übertreffen State-of-the-Art-Ansätze über verschiedene Typen von Aufgaben hinweg. Zunächst definieren wir eine datengetriebene Worker-Typologie. Basierend auf Verhaltensmustern von Crowdworkern auf niedriger Ebene, stellen wir Features zur Modellierung und Vorauswahl von Workern vor. Unsere Ergebnisse ziehen wichtige Auswirkungen für Crowdsourcing-Systeme nach sich, bei denen der Verhaltenstyp eines Crowdworkers vor der Teilnahme unbekannt ist. Weiterhin liefern wir Gründe für die kompetenzbasierte Vorauswahl in Crowdsourcing-Plattformen. Wir zeigen die Auswirkung von falscher Selbsteinschätzung auf reale Microtasks und stellen eine neue Methode zur Vorauswahl von Crowdworkern vor, die die Präzision von Worker-Selbsteinschätzungen berücksichtigt. Unsere Ergebnisse bestätigen, dass Anforderer auf Crowdsourcing-Plattformen durch das zusätzliche Berücksichtigen der Präzision von Worker-Selbsteinschätzungen in der Vorauswahlphase deutlich profitieren können.

- *Aufdecken versteckter Faktoren, die Crowd Work beeinflussen: der Fall von Aufgabenklarheit und Arbeitsumgebungen* — Worker auf Microtask Crowdsourcing-Plattformen streben nach einem Gleichgewicht zwischen dem Bedürfnis des finanziellen Einkommens und dem Bedürfnis nach hohem Ansehen. Dieses Gleichgewicht ist oft bedroht durch schlecht entworfene Aufgaben, da Worker versuchen, diese abzuschlieen, obwohl sie nur über ein unzureichendes Verständnis der zu leistenden Arbeit verfügen. Wir haben die Rolle der Aufgabenklarheit als eine charakterisierende Eigenschaft von Aufgaben im Crowdsourcing aufgezeigt und ein neues Modell für Aufgabenklarheit basierend auf Ziel- und Rollen-Klarheit vorgestellt. Wir haben aufgedeckt, dass Aufgabenklarheit von Crowdworkern einheitlich wahrgenommen wird und durch den Typ der Aufgabe beeinflusst wird. Anschlieend haben wir ein Set von Features zum Erreichen von Aufgabenklarheit vorgestellt und die erstellten Labels genutzt, um ein überwachtes Machine Learning-Modell zur Vorhersage von Aufgabenklarheit zu trainieren und zu validieren. Ein anderer Aspekt, der innerhalb des Microtask Crowdsourcing bislang unscheinbar geblieben ist, ist der Aspekt der Arbeitsumgebungen, definiert durch Hardware und Software, die Crowdworkern zur Verfügung steht. Mittels mehrerer Studien haben wir den signifikanten Einfluss von Arbeitsumgebungen auf das Ergebnis von Crowd Work aufgezeigt. Unsere Ergebnisse

deuten darauf hin, dass Crowd Worker eine Vielzahl von Arbeitsumgebungen verwenden, die die Qualität der Arbeit beeinflussen. Abhängig vom Design der User Interface-Elemente in Microtasks haben wir herausgefunden, dass einige Arbeitsumgebungen Crowdworker besser unterstützen als andere.

Schlagworte: Crowdsourcing, Human Computation, Microtasks, Crowd Workers, Verhalten, Performance, Qualität, Vorauswahl, Aufgabentypen, Aufgabenklarheit, Arbeitsumgebungen

ABSTRACT

Microtask crowdsouring has emerged as an excellent means to acquire human input on demand, and has found widespread application in solving a variety of problems. Popular examples include surveys, content creation, acquisition of image annotations, etc. However, there are a number of challenges that need to be overcome to realize the true potential of this paradigm. With an aim to improve the effectiveness of microtask crowdsourcing, we identify three main challenges to address within the scope of this thesis. The first challenge is the limited understanding of crowdsourced tasks and crowd worker characteristics. Understanding the dynamics of tasks that are crowdsourced and the behavior of workers contributing to tasks, can play a vital role in effective task design. Secondly, current worker pre-selection mechanisms are simplistic and inadequate. It is challenging to recruit workers with desirable skills in the absence of historical knowledge regarding the quality of work produced. There is a need for stronger indicators of worker competence and more effective pre-selection mechanisms. Finally, there has been an incomplete consideration of factors that influence and shape the quality of work produced in crowdsourced microtasks. It is important to fully understand different aspects that influence crowd work. In this dissertation, we tackle the aforementioned challenges and propose novel methods to overcome existing problems in each case. Our main contributions are described below.

- *Advancing the Current Understanding of Task Types, Worker Behavior and Quality Control* — Our findings from an extensive study of 1,000 workers on CrowdFlower advance the current understanding of crowdsourced microtasks and corresponding worker behavior. We propose a two-level categorization scheme for microtasks and revealed insights into the task affinity of workers, effort exerted to complete tasks of various types, and worker satisfaction with the monetary incentives. We analyze the prevalent malicious activity on crowdsourcing platforms and study the behavior exhibited by trustworthy and untrustworthy workers, particularly on crowdsourced surveys. To improve the overall quality of results, we propose behavioral metrics that can be used to measure and counter undesirable or potentially malicious activity in crowdsourced tasks. Considering these aspects, we prescribe guidelines for the effective design of crowdsourced tasks.

- *Novel Mechanisms for Worker Pre-selection* — We propose two distinct and novel methods for worker pre-selection that outperform state-of-the-art approaches across different types of tasks. First, we define a

data-driven worker typology. By relying on low-level behavioral traces of workers, we propose behavioral features for worker modeling and pre-selection. Our findings bear important implications for crowdsourcing systems where a worker's behavioral type is unknown prior to participation. Next, we make a case for competence-based pre-selection in crowdsourcing marketplaces. We show the implications of flawed self-assessments on real-world microtasks, and propose a novel worker pre-selection method that considers accuracy of worker self-assessments. Our results confirm that requesters in crowdsourcing platforms can greatly benefit by additionally considering the accuracy of worker self-assessments in the pre-screening phase.

- *Revealing Hidden Factors that Affect Crowd Work: The Cases of Task Clarity and Work Environments* — Workers of microtask crowdsourcing marketplaces strive to find a balance between the need for monetary income and the need for high reputation. Such balance is often threatened by poorly formulated tasks, as workers attempt their execution despite a sub-optimal understanding of the work to be done. We unearth the role of *task clarity* as a characterising property of tasks in crowdsourcing, and propose a novel model for task clarity based on the *goal* and *role* clarity constructs. We reveal that task clarity is coherently perceived by crowd workers, and is affected by the type of the task. We then propose a set of features to capture task clarity, and used the acquired labels to train and validate a supervised machine learning model for task clarity prediction. Another aspect that has remained largely invisible in microtask crowdsourcing is that of *work environments*; defined as the hardware and software affordances at the disposal of crowd workers which are used to complete microtasks on crowdsourcing platforms. Through multiple studies, we reveal the significant role of work environments in the shaping of crowd work. Our findings indicate that crowd workers are embedded in a variety of work environments which influence the quality of work produced. Depending on the design of UI elements in microtasks, we found that some work environments support crowd workers more than others.

Keywords: Crowdsourcing, Human Computation, Microtasks, Crowd Workers, Behavior, Performance, Quality, Pre-selection, Task Type, Task Clarity, Work Environments

FOREWORD

The studies presented in this thesis have been published at various conferences or journals, as follows.

In Chapter 2 and Chapter 3, we describe contributions included in:

- Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze and Gianluca Demartini. Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In ***CHI'15****: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1631-1640, 2015 (Full Paper). [GKDD15a]

- Ujwal Gadiraju, Ricardo Kawase and Stefan Dietze. A Taxonomy of Microtasks on the Web. In ***HT'14****: Proceedings of the 25th ACM Conference on Hypertext & Social Media*, pages 218-223, 2014 (Short Paper). [GKD14b]

- Ujwal Gadiraju, Patrick Siehndel, Besnik Fetahu and Ricardo Kawase. Breaking Bad: Understanding Behavior of Crowd Workers in Categorization Microtasks. In ***HT'15****: Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 33-38, 2015 (Short Paper). [GSFK15]

- Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase and Stefan Dietze. Human Beyond the Machine: Challenges and Opportunities of Microtask Crowdsourcing. In ***IEEE IS****: IEEE Intelligent Systems*, volume 4, pages 81-85, 2015 (Journal Article). [GSFK15]

- Ujwal Gadiraju. Make Hay While the Crowd Shines: Towards Efficient Crowdsourcing on the Web. In ***WWW'15****: Proceedings of the 24th International Conference on World Wide Web Companion*, pages 493-497, 2015 (PhD Symposium). [GSFK15]

Chapter 4 is built upon the work published in:

- Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase and Stefan Dietze. Crowd Anatomy Beyond the Good and Bad – Behavioral Traces for Crowd Worker Modeling and Pre-selection. ***JCSCW****: Journal of Computer Supported Cooperative Work Special Issue on Crowd Dynamics: Conflicts, Contradictions, and Cooperation Issues in Crowdsourcing* (Journal Paper – *Under Review*).

Chapter 5 is built upon the work published in:

- Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel and Stefan Dietze. Using Worker Self-Assessments for Competence-based Pre-Selection in Crowdsourcing Microtasks. ***TOCHI**: ACM Transactions on Computer-Human Interaction* (Journal Paper). [GFK+17]

In Chapter 6 we include our research presented in:

- Ujwal Gadiraju, Jie Yang and Alessandro Bozzon. Clarity is a Worthwhile Quality – On the Role of Task Clarity in Microtask Crowdsourcing. ***HT'17**: Proceedings of the 28th ACM Conference on Hypertext & Social Media*, 2017 (Full Paper, ***Best Paper Award****). [GYB17]

Finally, in Chapter 7 we include our research presented in:

- Ujwal Gadiraju, Alessandro Checco, Neha Gupta and Gianluca Demartini. Modus Operandi of Crowd Workers : The Invisible Role of Microtask Work Environments. ***UBICOMP'17**: Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2017 (Journal Paper). [GCGD17]

During the course of my doctoral studies, I have collaborated with various researchers and published peer-reviewed papers investigating different areas of Information Retrieval and Web Science. Due to space constraints and the scope of this thesis, many of these papers are not discussed here. However, to provide an overall perspective of the research work carried out, a complete list of publications is presented below in a chronological order.

- Stefan Dietze, Jakob Beetz, Ujwal Gadiraju, Georgios Katsimpras, Raoul Wessel and René Berndt. Towards Preservation of Semantically Enriched Architectural Knowledge. In ***TPDL'13***: Proceedings of the 3rd International Workshop on Semantic Digital Archives co-located with 17th International Conference on Theory and Practice of Digital Libraries, 2013 (Full Workshop Paper). [DBG+13]
- Mihai Georgescu, Dang Duc Pham, Claudiu S. Firan, Ujwal Gadiraju and Wolfgang Nejdl. When in Doubt Ask the Crowd: Employing Crowdsourcing for Active Learning. In ***WIMS'14***: 4th International Conference on Web Intelligence, Mining and Semantics, 2014 (Full Paper). [GPF+14]

- Andrea Ceroni, Mihai Georgescu, Ujwal Gadiraju, Kaweh Djafari Naini and Marco Fisichella. Information Evolution in Wikipedia. In ***OpenSym'14***: Proceedings of The International Symposium on Open Collaboration, 2014 (Full Paper). [CGG+14]

- Ujwal Gadiraju, Ricardo Kawase and Stefan Dietze. Extracting Architectural Patterns from Web data. In ***ISWC'14***: Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, 2014 (Poster Paper). [GKD14a]

- Besnik Fetahu, Ujwal Gadiraju and Stefan Dietze. Crawl Me Maybe: Iterative Linked Dataset Preservation. In ***ISWC'14***: Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, 2014 (Poster Paper). [FGD14]

- Ujwal Gadiraju, Kaweh Djafari Naini, Andrea Ceroni, Mihai Georgescu, Dang Duc Pham and Marco Fisichella. WikipEvent: Temporal Event Data for the Semantic Web. In ***ISWC'14***: Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, 2014 (Demo Paper). [GNC+14]

- Ricardo Kawase, Patrick Siehndel and Ujwal Gadiraju. Technology Enhancing Learning: Past, Present and Future. In ***EC-TEL'14***: Proceedings of the Open Learning and Teaching in Educational Communities - 9th European Conference on Technology Enhanced Learning, 2014 (Full Paper). [KSG14]

- Ran Yu, Ujwal Gadiraju, Besnik Fetahu and Stefan Dietze. Adaptive Focused Crawling of Linked Data. In ***WISE'15***: Proceedings of the 16th International Conference on Web Information Systems Engineering, 2015 (Full Paper). [YGFD15]

- Ujwal Gadiraju, Stefan Dietze and Ernesto Diaz-Aviles. Ranking Buildings and Mining the Web for Popular Architectural Patterns. In ***WebSci'15***: Proceedings of the ACM Web Science Conference, 2015 (Full Paper). [GDD15]

- Besnik Fetahu, Ujwal Gadiraju and Stefan Dietze. Improving Entity Retrieval on Structured Data. In ***ISWC'15***: Proceedings of the 14th International Semantic Web Conference, 2015 (Full Paper). [FGD15]

- Ujwal Gadiraju, Besnik Fetahu and Ricardo Kawase. Training Workers for Improving Performance in Crowdsourcing Microtasks. In ***EC-TEL'15***: Proceedings of the Design for Teaching and Learning in a Networked World - 10th European Conference on Technology Enhanced Learning, 2015 (Full Paper). [GFK15a]

- Andrea Ceroni, Ujwal Gadiraju and Marco Fisichella. Improving Event Detection by Automatically Assessing Validity of Event Occurrence in Text. In ***CIKM'15***: Proceedings of the 24th ACM International Conference on Information and Knowledge Management, 2015 (Short Paper). [CGF15]

- Tuan A. Tran, Claudia Niederée, Nattiya Kanhabua, Ujwal Gadiraju and Avishek Anand. Balancing Novelty and Salience: Adaptive Learning to Rank Entities for Timeline Summarization of High-impact Events. In ***CIKM'15***: Proceedings of the 24th ACM International Conference on Information and Knowledge Management, 2015 (Full Paper). [TNK+15]

- Ujwal Gadiraju, Sebastian Möller, Martin Nöllenburg, Dietmar Saupe, Sebastian Egger-Lampl, Daniel Archambault and Brian Fisher. Crowdsourcing Versus the Laboratory: Towards Human-Centered Experiments Using the Crowd. In *Evaluation in the Crowd*. Crowdsourcing and Human-Centered Experiments - Dagstuhl Seminar 15481, Dagstuhl Castle, Germany, 2015, Revised Contributions (Book Chapter). [GMN+15]

- Jakob Beetz, Ina Blümel, Stefan Dietze, Besnik Fetahu, Ujwal Gadiraju, Martin Hecher, Thomas Krijnen, Michelle Lindlar, Martin Tamke, Raoul Wessel and Ran Yu. Enrichment and Preservation of Architectural Knowledge. 3D Research Challenges in Cultural Heritage II - How to Manage Data and Knowledge Related to Interpretative Digital 3D Reconstructions of Cultural Heritage, 2015 (Book Chapter). [BBD+16a]

- Ujwal Gadiraju, Patrick Siehndel and Stefan Dietze. Estimating domain specificity for effective crowdsourcing of link prediction and schema mapping. In ***WebSci'16***: Proceedings of the 8th ACM Conference on Web Science, 2016 (Short Paper). [GSD16a]

- Patrick Siehndel and Ujwal Gadiraju. Unlock the Stock: User Topic Modeling for Stock Market Analysis. In ***EDBT'16***: Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference, EDBT/ICDT Workshops, 2016 (Full Paper). [SG16]

- Ran Yu, Ujwal Gadiraju, Xiaofei Zhu, Besnik Fetahu and Stefan Dietze. Towards Entity Summarisation on Structured Web Markup. In ***ESWC'16***: ESWC 2016 Satellite Events, Heraklion, Crete,Greece, 2016, Revised Selected Papers. [YGZ+16]

- Ran Yu, Besnik Fetahu, Ujwal Gadiraju and Stefan Dietze. A Survey on Challenges in Web Markup Data for Entity Retrieval. In ***ISWC'16***: 15th International Semantic Web Conference (ISWC 2016) (Poster Paper). [YFGD16]

- Andrea Ceroni, Ujwal Gadiraju, Jan Matschke, Simon Wingert and Marco Fisichella. Where the Event Lies: Predicting Event Occurrence in Textual Documents. In ***SIGIR'16***: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016 (Demo Paper). [CGM+16]

- Ujwal Gadiraju, Gianluca Demartini, Djellel Eddine Difallah and Michele Catasta. It's getting crowded!: how to use crowdsourcing effectively for web science research. In ***WebSci'16***: Proceedings of the 8th ACM Conference on Web Science, WebSci 2016 (Tutorial). [GDDC16]

- Ujwal Gadiraju, Patrick Siehndel and Stefan Dietze. Estimating domain specificity for effective crowdsourcing of link prediction and schema mapping. In ***WebSci'16***: Proceedings of the 8th ACM Conference on Web Science, WebSci 2016 (Short Paper). [GSD16b]

- Jakob Beetz, Ina Blümel, Stefan Dietze, Besnik Fetahu, Ujwal Gadiraju, Martin Hecher, Thomas Krijnen, Michelle Lindlar, Martin Tamke, Raoul Wessel and Ran Yu. Enrichment and Preservation of Architectural Knowledge. In *3D Research Challenges in Cultural Heritage II - How to Manage Data and Knowledge Related to Interpretative Digital 3D Reconstructions of Cultural Heritage.* (Book Chapter). [BBD+16b]

- Ujwal Gadiraju and Stefan Dietze. Improving Learning Through Achievement Priming in Crowdsourced Information Finding Microtasks. In ***LAK'17***: Proceedings of the 17th International Learning Analytics & Knowledge Conference, 2017 (Full Paper). [GD17a]

- Ujwal Gadiraju and Ricardo Kawase. Improving Reliability of Crowdsourced Results by Detecting Crowd Workers with Multiple Identities. In ***ICWE'17***: Proceedings of the 17th International Conference on Web Engineering, ICWE 2017 (Full Paper). [GK17]

- Andrea Ceroni, Ujwal Gadiraju and Marco Fisichella. JustEvents: A Crowdsourced Corpus for Event Validation with Strict Temporal Constraints. In ***ECIR'17***: Proceedings of the 39th European Conference on IR Research, 2017 (Short Paper). [CGF17]

- Ran Yu, Ujwal Gadiraju, Besnik Fetahu and Stefan Dietze. FuseM: Query-Centric Data Fusion on Structured Web Markup. In ***ICDE'17***: 33rd IEEE International Conference on Data Engineering, 2017 (Short Paper). [CGF17]

> *"Read not to contradict and confute; nor to believe and take for granted; nor to find talk and discourse; but to weigh and consider. Some books are to be tasted, others to be swallowed, and some few to be chewed and digested: that is, some books are to be read only in parts, others to be read, but not curiously, and some few to be read wholly, and with diligence and attention."*
>
> — *Francis Bacon*

ACKNOWLEDGMENTS

First and foremost, I would like to dedicate this body of work to my parents, Vijay and Sarada Gadiraju, who have constantly inspired me to follow my interests, and supported me with equanimity through thick and thin. My sister, Pooja Gadiraju, and brother, Gaurav Gadiraju, have been two sturdy, steady and resolute pillars of my strength. Their encouragement when things did not go according to plan, joy and appreciation when they did, were indispensable constants throughout this eventful ride. I thank my brother-in-law, Venkat Attili, for the interest he has shown in my work. His genuine regard, and our lively discussions ushered me on. The unwavering love and confidence with which my family has enveloped me will remain my most cherished asset.

I am thankful to Prof. Wolfgang Nejdl, who set a high bar for excellence in research, and instilled a zeal to strive for the best in me. Together with Stefan Dietze, he allowed me the freedom to pursue research in *human computation* and *crowdsourcing*, a field which was not commonplace at L3S, but one that I found immediately and addictively appealing. It is with their relentless support through the course of this dissertation, that I was able to travel far and wide, exploring numerous opportunities and meeting inspiring people. The vibrant minds I encountered, have helped in shaping my ideas along the way, giving my research timely impetus. I thank Stefan for his candor and guidance over the years.

My friends in Hannover, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, Andrea Ceroni and Kaweh Djafari have made life as a PhD candidate lively and exciting. Our long vigils after office hours, battling the ever so swiftly encroaching deadlines, stood second best to our infectious ability to celebrate without occasion. Several of our interesting research ideas can be traced back to exhilarating conversations we had, and the occasional brainstorming sessions that we indulged in. I thank Besnik for seamlessly transitioning between the roles of a friendly advocate or a feisty critic, and for being my metaphorical ‘partner in crime’ all the way to this finish line. The commendable aptitude and efficiency with which Ricardo carried out his work was a learning that rubbed off on me, for this I extend my gratitude. It was with him that I first began working on *microtask crowdsourcing*, and I am glad that this partnership was forged.

No one has played a more certain hand in being an ally of mine for all things crowdsourcing, than Prof. Gianluca Demartini. Interacting with him

and the dream crowdsourcing teams he conjures at will, has emboldened my conviction that there is a permanent role for innate human intelligence despite the proliferation of machine intelligence. I have garnered several valuable learnings from Gianluca, for which I am particularly thankful.

I must express my deep gratitude to everyone who I have gotten to know through conferences, seminars, and other gatherings; especially those who I have had the pleasure to collaborate with closely. As the old adage goes, 'a man is a sum of the experiences he has had', and I stand just as grateful for the people I have crossed paths with, as I am for the amazing memories I have made over the last four years.

I wish to thank friends from older times who have made the challenge of completing a PhD less onerous with their sustained presence. Finally, for everyone who has shared even a small moment on this epic journey with me, for whom I have failed to muster characters and ink, this is an ode to you too.

I hope you enjoy the read as much as I enjoyed putting it all together.

— *Ujwal Gadiraju*

Hannover, Germany 17th July, 2017

"A scientist explores the world of phenomena by successive approximations. He knows that his data are not precise and that his theories must always be tested. It is quite natural that he tends to develop healthy skepticism, suspended judgment, and disciplined imagination."

— *Edwin Powell Hubble*

Contents

List of Figures

List of Tables

1

Crowdsourcing – All Things Big and Small

> *"The whole is greater than the sum of its parts."*
>
> — *Aristotle*

We live in a world today that is characterized by an unprecedented rate of technological progress. The ubiquity of electronic devices in the Internet age has led to an avalanche of user generated content. Several hundred millions of people around the globe share photos, videos and textual data on a daily basis. Governments around the world gather a deluge of data ranging from census data to incident reports. Businesses accumulate data regarding their customers; from detailed activity logs to interests and preferences. The list goes on. This availability of *big data* and analytical resources has led to several innovative developments over the last decade; improved healthcare, smart environments, aiding urban planning and traffic control, predictive analytics for economic forecasting, and many more. In this context, several new algorithms and systems have been developed that play important roles in the background.

The prowess of machines stems mainly from their large computational powers. Having said that, there are several problems that machines cannot solve due to their inability to deal effectively with certain abstract and complex concepts (for example, beauty or perception). In contrast, although humans lack similar computational capabilities, they can easily comprehend and deal with abstract and complex notions (for example, a human can take one look at a picture and say whether or not she finds it beautiful and why). The ability of humans and machines have for long been realized as being complementary. Human judgments are widely used to create training data for supervised machine learning algorithms. Relevance assessments by humans are used to build groundtruths that are then used to evaluate search and retrieval systems. Innate *human intelligence* therefore plays a pivotal role in realizing solutions to complex problems that machines alone would currently fail to solve. However, human input on a large scale has not always been as accessible as it is today.

The term crowdsourcing was coined in 2005 by Jeff Howe [How06], and proposed as an alternative to outsourcing. Howe defined it as the act of recruiting participants through an open call to complete a job that anyone on the Internet was capable of doing. Since then, the term has been used to describe a variety of activities where a group of participants make contributions towards a larger goal, with varying incentives and motivations. Popular examples include Wikipedia (where thousands of people around the world contribute towards building and curating what is now the world's largest and most widely used knowledge base), question answering communities like Quora, Stackoverflow, etc., citizen science initiatives such as Galaxy Zoo's[1] quest to identify and classify images of galaxies, and crowdfunding platforms like Kickstarter[2], among others.

1.1 Stepping Stones and Success Stories

To provide a glimpse of the potential use of crowdsourcing to solve a variety of real-world problems, we briefly present a few inspiring anecdotes from the past.

Although the word crowdsourcing itself was coined only a decade ago, early examples can be traced back to over a century ago [Sur05]. In 1907, at a county fair in Plymoth, people were encouraged to participate in a contest where they were set the task of estimating how much an ox on display would weigh after being slaughtered and dressed. 800 people participated in the contest, including butchers and farmers, some of whom were experts in judging the weight of cattle; others much less so. On analyzing the estimates, Sir Francis Galton, an English statistician, observed that the average estimate of 787 participants (13 were discarded for being illegible) was almost exactly accurate. In contrast, any individual estimate picked at random had a 50% chance of falling within a [-3.7%,+2.4%] interval of the average estimate [Gal07]. This early finding suggested the now popular notion of '*wisdom of crowds*'.

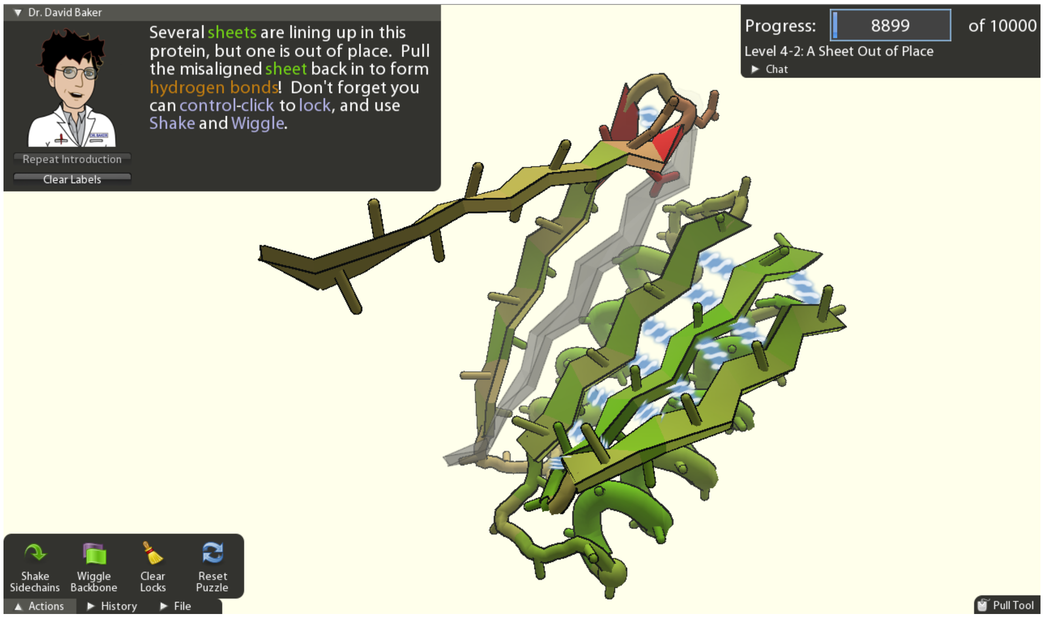
A century after Galton's surprising findings, a remarkable crowdsourcing effort unfortunately did not succeed in its purpose; the search for a noted computer scientist, Jim Gray [HT11]. In January, 2007, Jim Gray disappeared at sea while he was sailing alone with plans to scatter his mother's ashes nearly 30 miles outside San Francisco's Golden Gate. On realizing his disappearance, his friends and colleagues initiated ways to help authorities locate and rescue Gray. This evolved into an extraordinary search and rescue effort involving private planes, satellites, automated image analysis, ocean current simulations, and crowdsourced human computing, along with with the U.S. Coast Guard. The resulting team included graduate students, engineers, computer scientists, oceanographers, astronomers, venture capitalists, and entrepreneurs, many of whom had never made acquaintance. The efforts were supported by access to funds, technology, among other requisites, and a willingness to work relentlessly. Nearly 3

[1] http://www.galaxyzoo.org/

[2] http://www.kickstarter.com/

fruitless weeks later, the team agreed to call off the search.

In 2010 Cooper et al., researchers at the University of Washington showed that *Foldit*, a multiplayer online game that engages non-scientist players in solving hard prediction problems, provided very useful results that either matched or outperformed state-of-the-art algorithmically computed solutions [CKT+10]. Figure 1.1 presents a screenshot of the game interface during play. A year later in 2011, *Foldit* players helped to decipher the crystal structure of a monkey virus which causes human immunodeficiency virus infection and acquired immune deficiency syndrome (HIV/AIDS) [KDC+11]. This was a scientific problem that had remained unsolved for almost 15 years. While the puzzle was available for three weeks, non-scientist players remarkably produced an accurate 3D model of the enzyme in only ten days.



Several sheets are lining up in this protein, but one is out of place. Pull the misaligned sheet back in to form hydrogen bonds! Don't forget you can control-click to lock, and use Shake and Wiggle.

Progress: 8899 of 10000

Level 4-2: A Sheet Out of Place

Figure 1.1 A screen capture of the Foldit interface during play, corresponding to a puzzle. *Source:* http://lgdb.org/game/foldit.

In March 2014, Malaysian Airlines flight MH370 went missing enroute to Beijing from Kuala Lumpur. Only a few hours after the plane's disappearance, the search efforts were joined by nearly 2.5 million ordinary Internet users using the Tomnod website[3], in a crowdsourcing campaign that went viral [Fis14]. Figure 1.2 presents the interface on the Tomnod website which was used to try and locate the missing flight. By the end of the campaign, over 8 million users reportedly scanned 1,007,750 square kilometers of high resolution satellite imagery from the commercial satellite company, DigitalGlobe[4], and tagged millions of clues to help locate the missing aircraft. Search teams then investigated all the promising leads that were discovered. Despite the

[3] http://www.tomnod.com/

[4] http://www.digitalglobe.com/

incredible crowdsourced participation, the missing plane was not found.

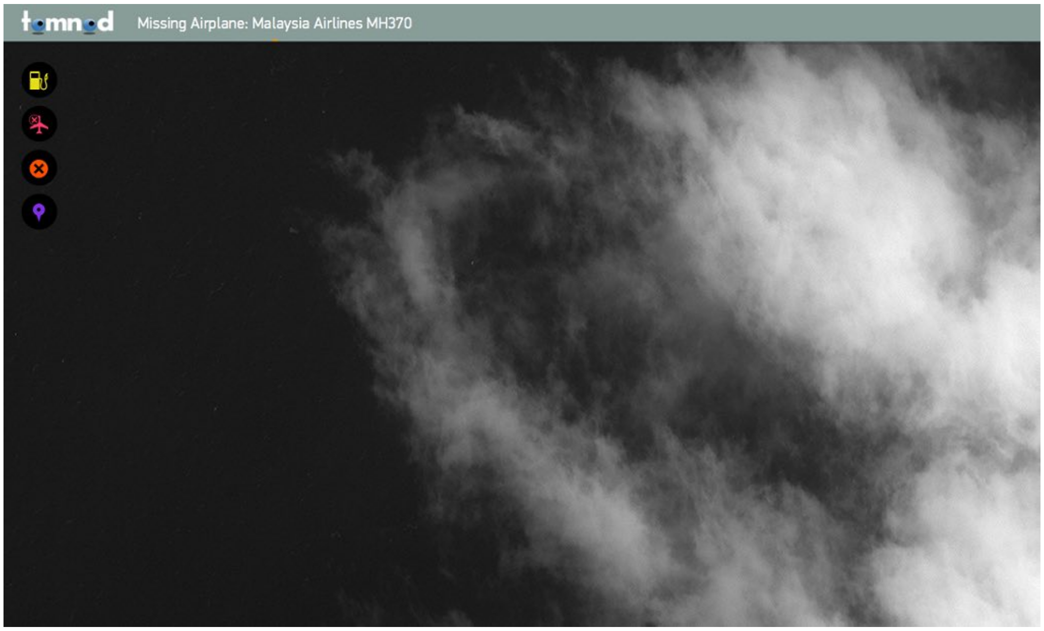


Figure 1.2 A screen capture of the Tomnod website showing the depicting the search interface for Malaysian Airlines flight MH370. *Source: [Fis14].*

As can be noted from the anecdotes discussed above, the motivation behind crowdsourced participation can vary greatly; from gamified incentives to altruism. In this thesis, we address a distinct realm of crowdsourcing where participants are primarily motivated by monetary rewards attached to the tasks, referred to as the '*paid crowdsourcing paradigm*'. We describe this context further in the following section.

1.2 Human Computation and Paid Crowdsourcing

Human computation has been defined as a paradigm for utilizing human processing power to solve problems that computers cannot yet solve [QB11]. It was first introduced at length by Von Ahn via the ESP Game [VAD04, VA08]. The author reflected that tasks like image recognition are trivial for humans, but challenge sophisticated algorithms. Although it is cumbersome and costly, manual labeling was the most suitable method for obtaining precise image descriptions. Von Ahn's ESP game showed that people could label images without realizing they were doing so, and that the experience could be made greatly enjoyable.

Human input can be acquired through a variety of implicit or intrinsic incentivization mechanisms; using gamification to encourage participation as in the ESP game, or altruism, as observed in the crowdsourcing campaign to find the missing Malaysian Airlines flight. As an alternative, participation can also be encouraged through explicit material incentives. The paid crowdsourcing paradigm has emerged as a result of the great need and high potential for acquiring human input to solve different problems. Some prominent examples are presented in the Figure 1.3. Freelancing platforms such as Upwork[5] or Freelancer[6] allow users to post jobs and find others capable of completing the work, in return for a corresponding monetary payment. Similarly, 99designs[7] and DesignCrowd[8] are online graphic design marketplaces.

Figure 1.3 Some examples of crowdsourcing platforms that serve as marketplaces for a diverse set of tasks ranging from creative ideation to human intelligence tasks (HITs), offering monetary incentives for participation.

Over the years, we have witnessed a surge in the adoption of the paid crowdsourcing paradigm to solve problems that require human intelligence at a large scale. This rise has been greatly propelled by microtask crowdsourcing platforms like Amazon's Mechanical Turk[9] (AMT) and CrowdFlower[10]. Microtask crowdsourcing platforms bring together *requesters* and *crowd workers* from around the world. *Requesters* are task administrators with specific needs and requirements who deploy tasks on crowdsourcing platforms to gather responses from participants. *Crowd workers* are participants willing to complete tasks on crowdsourcing platforms while meeting priorly stated requirements, in return for monetary rewards.

[5] http://www.upwork.com/
[6] http://www.freelancer.com/
[7] http://99designs.com/
[8] http://www.designcrowd.com/
[9] http://www.mturk.com/
[10] http://www.crowdflower.com/

1.3 Scope, Challenges and Contributions

Crowdsourcing solutions are gaining in popularity to solve problems that require human intelligence at a large scale. In the last decade there have been numerous applications of microtask crowdsourcing, spanning several domains in both research (from sociology to computer science) and for practical benefits across disciplines. Microtask crowdsourcing has unmistakably broken the barriers of qualitative and quantitative studies by providing a means to scale-up previously constrained laboratory studies and controlled experiments [HRZ11, PCI10]. Today, one can easily build ground truths for evaluation [GL10], access potential participants around the clock with diverse demographics at will [Ipe10b, DCD+15b], and all within an unprecedentedly short amount of time. This also comes with a number of challenges related to lack of control on participants and to data quality, some of which are addressed in this thesis, as described further on in this section. In this thesis, we do not address the realm of voluntary crowdsourcing, such as citizen science [BSP+14], serious games or games with a purpose [VAD08], wikis [DRH11], and so forth. Instead, we focus on solving some key problems in paid microtask crowdsourcing. Our work is influenced and propelled by the belief that there will always be streams of work needed by various people in the society, which cannot be satisfied by leveraging volutary participation or gamification, and for which paid channels will serve as the ideal source to meet the demands [KNB+13].

Some of the main challenges in microtask crowdsourcing, at the time when the different works presented in this thesis were carried out and published, are described below. We methodically address each of these challenges in this thesis, thereby either advancing the current understanding of crowd work or proposing solutions that outperform existing methods in each case.

- *Limited understanding of crowdsourced tasks and worker characteristics* — Spanning only the last decade, the field of microtask crowdsourcing is still nascent. Understanding the type of tasks that can be crowdsourced, as well as the landscape of microtasks that are popularly crowdsourced, will help design better platforms and improve marketplace dynamics between requesters and workers. Little is known and understood about worker behavior in microtask crowdsourcing platforms. Understanding how workers differ in their behavior, in a manner that determines the quality of their work can inform task design and pave way for more effective quality control mechanisms.

- *Inadequate means for worker pre-selection* — Often there is little or no historical data available corresponding to crowd workers, that can aid in predicting the quality of their work. In the absence of such indicators as well as in a bid to recruit workers with desirable skills and proven accuracy, requesters typically use either the peformance of workers in qualification tests or their performance in a smaller sample of the actual task, as criteria for worker selection in the pre-screening phase. While this serves as an approximation for recruiting desirable workers, there is a need for stronger indicators of worker competence and more effective pre-selection mechanisms.

- *Incomplete consideration of factors that influence quality related outcomes* — Quality control in microtask crowdsourcing has been arguably the most researched topic in the field. Yet, there has been a incomplete consideration of aspects that affect the quality of work that is produced. To promote fair and justified treatment of work that is produced on microtask crowdsourcing platforms, it is important to understand and consider all factors that shape the quality of work. This is largely lacking in current practice. Requesters and platforms typically consider only the final results that are produced by workers, without paying heed to *how* the results were produced. For example, ethnography works in the past, that have investigated the crowdsourcing paradigm have clearly illustrated stark differences in the contexts that different crowd workers are situated in [GMHO14, MHOG14].

Through our work in this thesis, we aim to '*improve the effectiveness of the microtask crowdsourcing paradigm*'. *Effectiveness* in this context is defined by the degree to which crowd workers provide high-quality responses and requesters obtain the desired results, while optimizing the costs (task completion time, monetary reward) for all actors involved. We identify key open challenges in this realm and propose novel methods to overcome exisiting problems in each case. We also explore and reveal other factors that shape the quality of work produced by crowd workers, which have remained invisible so far. The main contributions of our work in this thesis are described below and illustrated in the Figure 1.4.

- *Advancing the Current Understanding of Task Types, Worker Behavior and Quality Control* — We first consider two pivotal aspects that influence the effectiveness of microtask crowdsourcing; task design, and crowd workers' behavior. To advance the current understanding of crowdsourced microtasks and corresponding worker behavior, we conducted an extensive study of 1,000 workers on CrowdFlower. We proposed a two-level categorization scheme for microtasks and revealed insights into the task affinity of workers, effort exerted by workers to complete tasks of various types, and their satisfaction with the monetary incentives (see Chapter 2).

  Quality control mechanisms need to accommodate a diverse pool of workers, exhibiting a wide range of behavior. A pivotal step towards fraud-proof task design is understanding the behavioral patterns of microtask workers. We analyzed the prevalent malicious activity on crowdsourcing platforms and studied the behavior exhibited by trustworthy and untrustworthy workers, particularly on crowdsourced surveys. To improve the overall quality of results, we proposed behavioral metrics that can be used to measure and counter undesirable or potentially malicious activity in crowdsourced tasks (see Chapter 3). Considering these aspects, we prescribed guidelines for the effective design of crowdsourced tasks. Leveraging the dynamics of tasks that are crowdsourced, and accounting for the behavior of workers, can help in designing better tasks.

- *Novel Mechanisms for Worker Pre-selection* — We propose two distinct and novel methods for worker pre-selection that outperform state-of-the-art approaches across different types of tasks. First, we define a data-driven worker typology. By relying on low-level behavioral traces of workers, we propose behavioral features for worker modeling and pre-selection (see Chapter 4). Our findings bear important implications for crowdsourcing systems where a worker's behavioral type is unknown prior to participation.

  Next, we reveal the diversity of individual worker competencies, making a case for competence-based pre-selection in crowdsourcing marketplaces. We show the implications of flawed self-assessments on real-world microtasks, and propose a novel worker pre-selection method that considers accuracy of worker self-assessments (see Chapter 5). Our results confirm that requesters in crowdsourcing platforms can greatly benefit by additionally considering the accuracy of worker self-assessments in the pre-screening phase.

- *Revealing Hidden Factors that Affect Crowd Work: The Cases of Task Clarity and Work Environments* — Workers of microtask crowdsourcing marketplaces strive to find a balance between the need for monetary income and the need for high reputation. Such balance is often threatened by poorly formulated tasks, as workers attempt their execution despite a sub-optimal understanding of the work to be done. We unearthed the role of *task clarity* as a characterising property of tasks in crowdsourcing, and proposed a novel model for task clarity based on the *goal* and *role* clarity constructs (see Chapter 6). We revealed that task clarity is coherently perceived by crowd workers, and is affected by the type of the task. We then proposed a set of features to capture task clarity, and used the acquired labels to train and validate a supervised machine learning model for task clarity prediction. A long-term analysis of the evolution of task clarity on Amazon's Mechanical Turk showed that clarity is not a property suitable for temporal characterization.

  An aspect that has remained largely invisible in microtask crowdsourcing is that of *work environments*; defined as the hardware and software affordances at the disposal of crowd workers which are used to complete microtasks on crowdsourcing platforms. Through multiple studies, we reveal the significant role of work environments in the shaping of crowd work (see Chapter 7). Our findings indicate that crowd workers are embedded in a variety of work environments which influence the quality of work produced. Depending on the design of UI elements in microtasks, we found that some work environments support crowd workers more than others. We introduced *ModOp*, a tool that helps to design crowdsourcing microtasks suitable for diverse crowd work environments. We empirically show that the use of *ModOp* results in reducing the cognitive load of workers, thereby improving their user experience.

1.4 Thesis Outline

We organize the remainder of this thesis as followed. In Chapter 2, we first describe the two microtask crowdsourcing platforms which play hosts to our investigation and experiments – Amazon's Mechanical Turk and CrowdFlower. We propose a two-level taxonomy to describe the landscape of crowdsourced microtasks. In Chapter 3, we present the different kinds of undesirable worker activity prevalent on crowdsourcing platforms. We propose behavioral measures and task design guidelines to inhibit malicious activity in crowdsourced surveys. Next, in Chapter 4 we explore how worker behavioral traces can be used to predict worker types according to a data-driven typology. We show that worker type based pre-selection results in a significant improvement in the quality of results without affecting the task turnover time. Chapter 5 establishes the diversity in crowd worker competencies. By drawing from self-assessment theories in psychology, we show that crowd workers often lack awareness about their true level of competence. We then operationalize worker self-assessments to propose a novel pre-selection mechanism that outperforms existing standard methods. Our findings suggest that combining worker accuracy in the pre-screening phase along with accuracy in self-assessments results in a stronger indicator of the true competence of workers. In Chapters 6 and 7, we reveal some hidden factors that shape the quality of crowd work. These factors should be considered by requesters while designing tasks and structuring workflows. In Chapter 6, we propose to model *task clarity* as a combination of *role* and *goal* clarity. We elucidate the role of task clarity through a longitudinal analysis, and show that task clarity is a characterizing property of quality. Through three distinct studies in the penultimate Chapter 7, we shed light on the role of *work environments* in shaping the quality of work. Finally, in Chapter 8, we draw conclusions and reflect on the implications of our work presented in this thesis. We highlight the important contributions that have advanced the understanding of microtask crowdsourcing, and informed workflow design to make the paradigm more effective. We conclude by discussing directions and setting precedents for future work.

To aid readers of this thesis, each chapter has been written to serve as a self-contained reflection that highlights the challenges being tackled in the chapter, the related literature in that context, the proposed approach, experimental setup and methodology, our consequent findings and their implications.

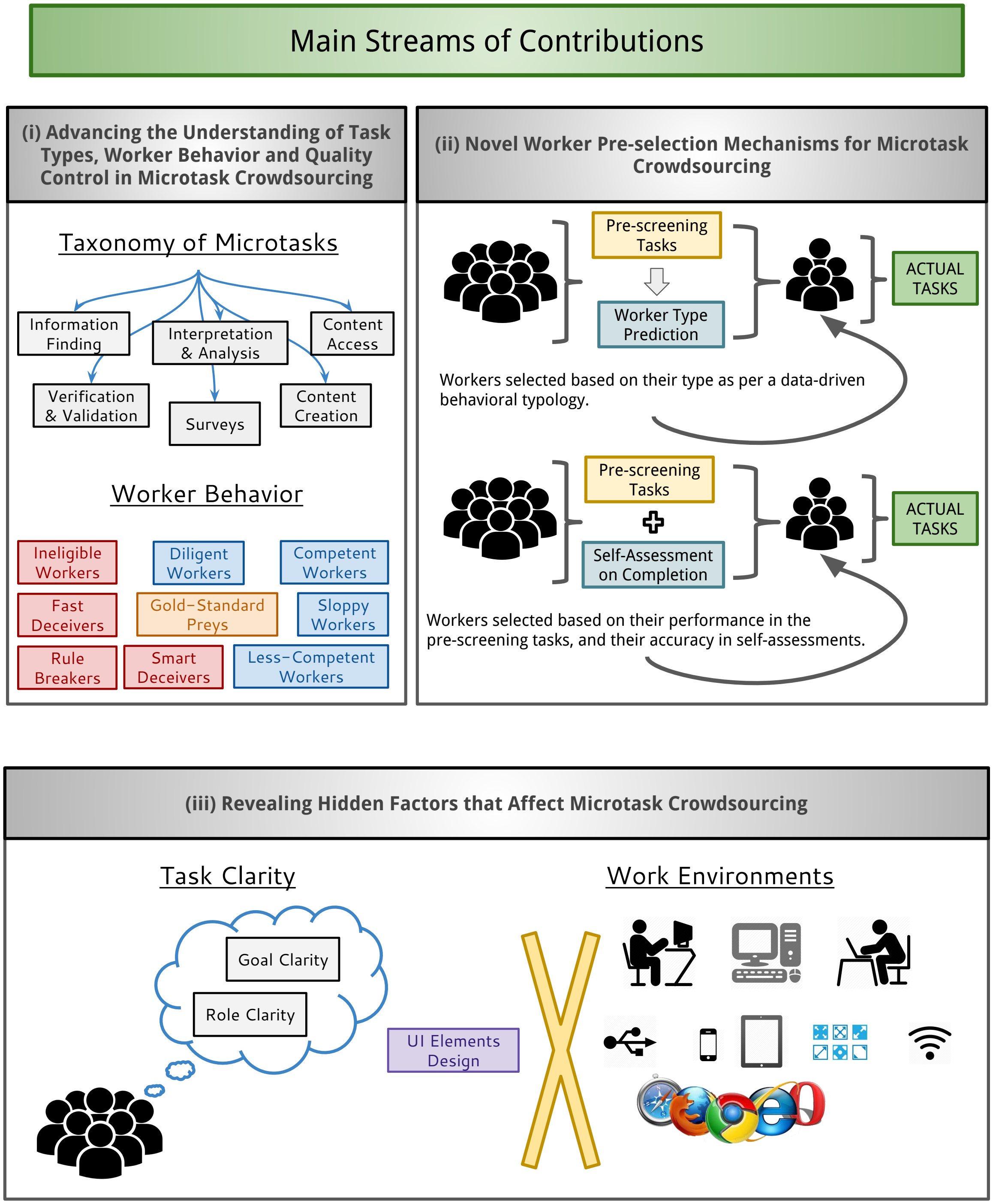

Main Streams of Contributions
(i) Advancing the Understanding of Task Types, Worker Behavior and Quality Control in Microtask Crowdsourcing
Taxonomy of Microtasks
Information Finding
Interpretation & Analysis
Content Access
Verification & Validation
Surveys
Content Creation
Worker Behavior
Ineligible Workers
Diligent Workers
Competent Workers
Fast Deceivers
Gold–Standard Preys
Sloppy Workers
Rule Breakers
Smart Deceivers
Less–Competent Workers
(ii) Novel Worker Pre-selection Mechanisms for Microtask Crowdsourcing
Pre-screening Tasks
Worker Type Prediction
ACTUAL TASKS
Workers selected based on their type as per a data-driven behavioral typology.
Pre-screening Tasks
Self-Assessment on Completion
ACTUAL TASKS
Workers selected based on their performance in the pre-screening tasks, and their accuracy in self-assessments.
(iii) Revealing Hidden Factors that Affect Microtask Crowdsourcing
Task Clarity
Goal Clarity
Role Clarity
UI Elements Design
Work Environments

Figure 1.4 Overview of the main streams of contributions of this dissertation. Contributions in the stream (i) are elaborated in chapters 2 and 3, those in stream (ii) are elucidated from chapters 4 and 5, and contributions in stream (iii) are explained in chapters 6 and 7.

2 Overview of Platforms and a Study of Task Types

> *"For me, it is far better to grasp the Universe as it really is than to persist in delusion, however satisfying and reassuring."*
>
> — *Carl Sagan*

Nowadays, a substantial number of people are turning to crowdsourcing, in order to resolve tasks that require human intervention. Understanding the dynamics of the tasks that are crowdsourced and the behaviour of workers, plays a vital role in efficient task design. In this chapter, we first introduce and describe two popular microtask crowdsourcing platforms that form the subject of our research investigations presented in the upcoming chapters. Next, we propose a two-level categorization scheme for tasks, based on an extensive study of 1,000 workers on CrowdFlower. In addition, we present insights into certain aspects of crowd behaviour; the task affinity of workers, effort exerted by workers to complete tasks of various types, and their satisfaction with the monetary incentives.

2.1 Introduction

Microtask crowdsourcing is evolving rapidly as a means to access scalable human input on demand. Over the last decade, there has been a considerable amount of work towards establishing suitable platforms and proposing frameworks for fruitful crowdsourcing. While developing a sound definition of crowdsourcing, Estelles and Guevara [EAGLdG12] suggested that microtasks are of variable complexity and modularity, and entail mutual benefit to the *worker*[1] and the *requester*[2]. Accumulating

[1] A user that performs tasks in exchange of monetary rewards on a crowdsourcing platform.

[2] A user that deploys tasks to be completed on a crowdsourcing platform, also called a *task administrator*.

small contributions through such microtasks facilitates the accomplishment of work that is not easily automatable, through rather minor contributions of each individual worker. AMT[3] and CrowdFlower[4] are good examples of popular crowdsourcing marketplaces. A typical microtask crowdsourcing platform entails two main groups of actors; *requesters* and *workers*. Requesters post tasks as per their needs on the crowdsourcing platform, with an attached monetary reward to incentivize successful completion. Workers then self-select and complete tasks they prefer to work on, from the list of tasks available to them [LCGM09]. Based on the quality of responses provided by the workers, requesters either accept or reject the responses. On successful completion of the tasks and acceptance from the requesters, workers are paid the corresponding money via the platform. This typical workflow is illustrated in the Figure 2.1. Thus, crowdsourcing platforms play a vital role in bringing together requesters who are looking for human input, and thousands of people willing to contribute and complete tasks in return for small amounts of money.

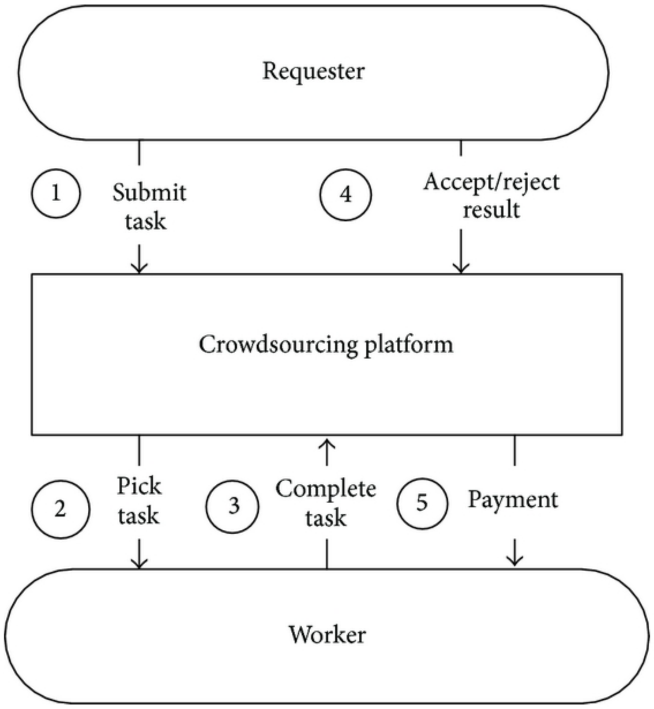


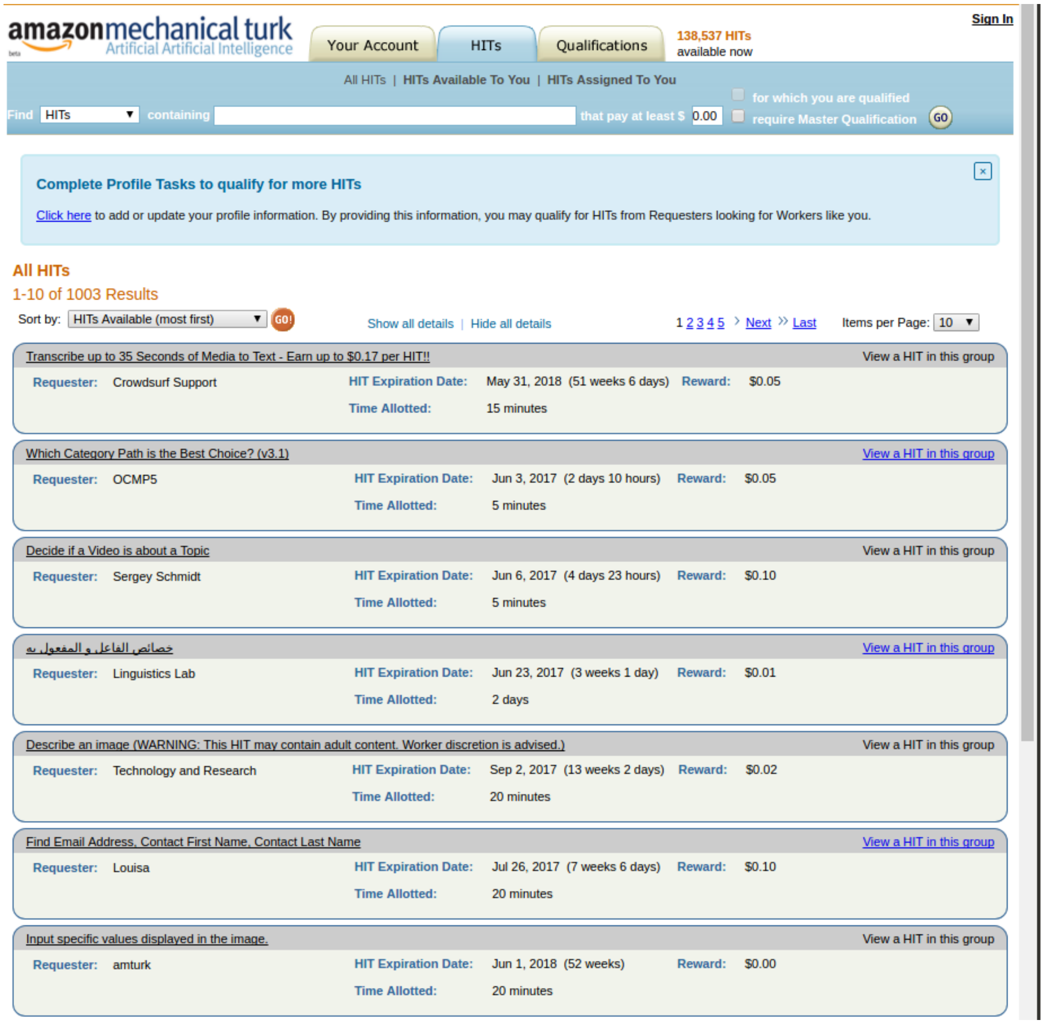
Figure 2.1 Typical workflow on a paid microtask crowdsourcing platform. *Source: [SR14]*

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

[4] http://www.crowdflower.com/

Amazon Mechanical Turk

AMT was the first and oldest microtask crowdsourcing platform of its kind, where requesters post tasks (called human intelligence tasks or HITs) and workers from around the world complete them. It was first made publicly available in November, 2005. Over the years, AMT has been the subject of a large amount of crowdsourcing research. This is largely due to the potential that a reliable on-demand workforce offers. Figure 2.2 presents the interface that workers can access with a list of available HITs. Workers from around the globe pick tasks they prefer to complete from the list of available HITs. Although there are workers based in several countries around the world, the majority of workers are currently based in the United States, followed by workers in India [Ipe10a].


amazonmechanical turk
Artificial Artificial Intelligence
beta
Your Account
HITs
Qualifications
138,537 HITs
available now
Sign In
All HITs | HITs Available To You | HITs Assigned To You
Find
HITs ▼
containing
that pay at least $
0.00
□ for which you are qualified
□ require Master Qualification
GO
×
Sort by:
HITs Available (most first) ▼
GO!
Show all details | Hide all details
1 2 3 4 5 › Next » Last
Items per Page:
10 ▼

Complete Profile Tasks to qualify for more HITs

Click here to add or update your profile information. By providing this information, you may qualify for HITs from Requesters looking for Workers like you.

All HITs

1-10 of 1003 Results

Transcribe up to 35 Seconds of Media to Text - Earn up to $0.17 per HIT!! — View a HIT in this group
Requester: Crowdsurf Support — HIT Expiration Date: May 31, 2018 (51 weeks 6 days) — Reward: $0.05
Time Allotted: 15 minutes

Which Category Path is the Best Choice? (v3.1) — View a HIT in this group
Requester: OCMP5 — HIT Expiration Date: Jun 3, 2017 (2 days 10 hours) — Reward: $0.05
Time Allotted: 5 minutes

Decide if a Video is about a Topic — View a HIT in this group
Requester: Sergey Schmidt — HIT Expiration Date: Jun 6, 2017 (4 days 23 hours) — Reward: $0.10
Time Allotted: 5 minutes

خصائص الفاعل و المفعول به — View a HIT in this group
Requester: Linguistics Lab — HIT Expiration Date: Jun 23, 2017 (3 weeks 1 day) — Reward: $0.01
Time Allotted: 2 days

Describe an image (WARNING: This HIT may contain adult content. Worker discretion is advised.) — View a HIT in this group
Requester: Technology and Research — HIT Expiration Date: Sep 2, 2017 (13 weeks 2 days) — Reward: $0.02
Time Allotted: 20 minutes

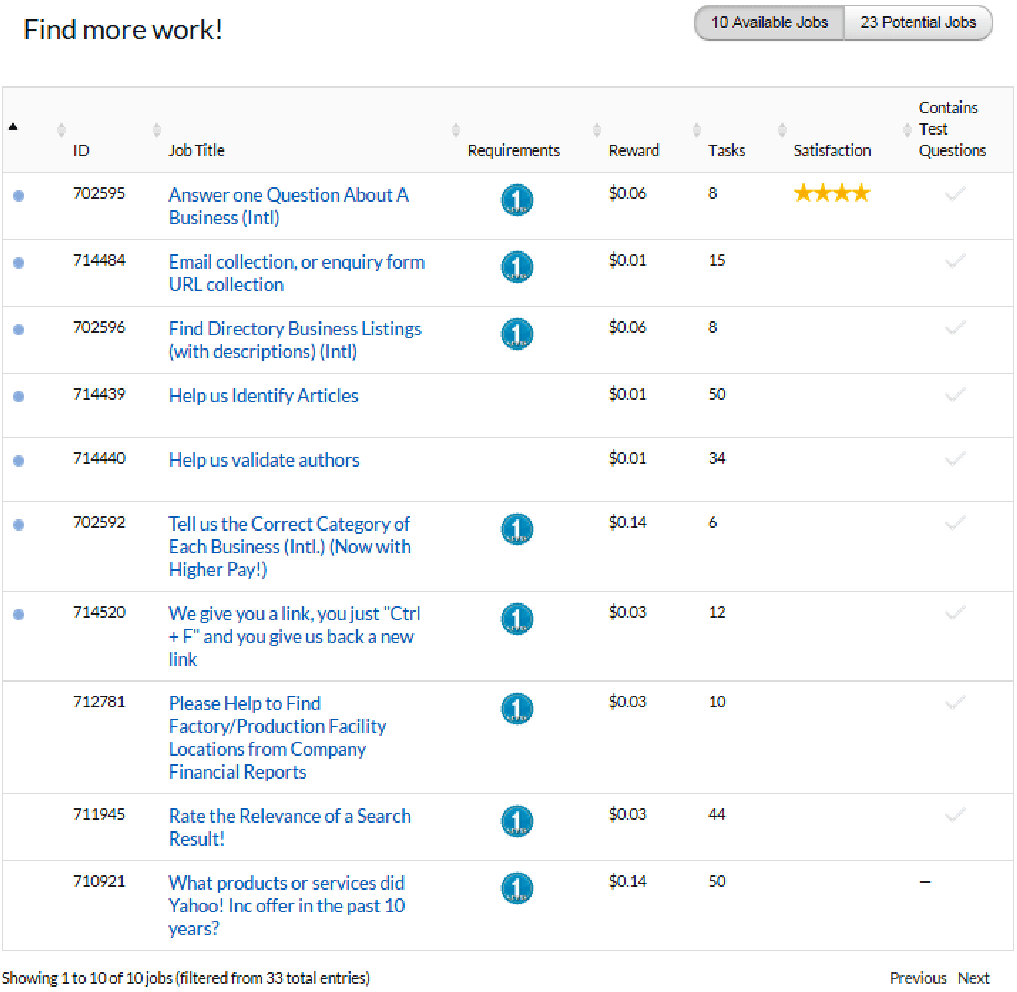
Find Email Address, Contact First Name, Contact Last Name — View a HIT in this group
Requester: Louisa — HIT Expiration Date: Jul 26, 2017 (7 weeks 6 days) — Reward: $0.10
Time Allotted: 20 minutes

Input specific values displayed in the image. — View a HIT in this group
Requester: amturk — HIT Expiration Date: Jun 1, 2018 (52 weeks) — Reward: $0.00
Time Allotted: 20 minutes

Figure 2.2 A screenshot of the Amazon Mechanical Turk interface showing requesters the HITs available for completion.

CrowdFlower

CrowdFlower is another microtask crowdsourcing platform that has gained popularity since it was founded in 2007. It provides similar functionalities as AMT. The workforce in this case, stems from partnerships with third-party *channels* such as ClickSense, Neobox, CrowdGuru, etc., that contract directly with CrowdFlower and provide requesters with a steady worker base that works around the clock to solve a variety of problems by completing available tasks (called '*jobs*' on CrowdFlower).



▲	ID	Job Title	Requirements	Reward	Tasks	Satisfaction	Contains Test Questions
●	702595	Answer one Question About A Business (Intl)	1	$0.06	8	★★★★	✓
●	714484	Email collection, or enquiry form URL collection	1	$0.01	15		✓
●	702596	Find Directory Business Listings (with descriptions) (Intl)	1	$0.06	8		✓
●	714439	Help us Identify Articles		$0.01	50		✓
●	714440	Help us validate authors		$0.01	34		✓
●	702592	Tell us the Correct Category of Each Business (Intl.) (Now with Higher Pay!)	1	$0.14	6		✓
●	714520	We give you a link, you just "Ctrl + F" and you give us back a new link	1	$0.03	12		✓
	712781	Please Help to Find Factory/Production Facility Locations from Company Financial Reports	1	$0.03	10		✓
	711945	Rate the Relevance of a Search Result!	1	$0.03	44		✓
	710921	What products or services did Yahoo! Inc offer in the past 10 years?	1	$0.14	50		–

Showing 1 to 10 of 10 jobs (filtered from 33 total entries)

Figure 2.3 A screenshot of the available jobs on CrowdFlower as shown to a worker through the Neobux channel.

Recent work that compared AMT to CrowdFlower found that CrowdFlower provided relatively better response rate, but CrowdFlower workers failed more attention-check questions and did not reproduce known effects replicated on AMT [PBSA17].

A large number of researchers have used these platforms in order to gather distributed and unbiased data, to validate results or to build ground truths. However, the literature inspecting the actors involved in the crowsourcing process is rather scarce. Only a few noteworthy works have investigated best practices, the reliability of data[MS13a], or have proposed comprehensive strategies and guidelines[WHA12]. As a consequence, without adequate knowledge of how one can effectively and efficiently exploit the wisdom of the crowd through crowdsourcing platforms, several requesters are hindered by chaotic results leading to doubtful conclusions. Thus, it is essential that requesters are, to some extent, educated in crowdsource task modeling. To facilitate a greater understanding of the dynamics between the requesters who deploy tasks and the crowd workers, we venture into determining a fine-grained goal-oriented categorization of crowdsourced tasks.

2.2 A Taxonomy of Microtasks

In previous work, Kazai et al. [KKMF11] used behavioural observations to define the types of workers in the crowd. They type-casted workers as being either *sloppy*, *spammer*, *incompetent*, *competent*, or *diligent*. By doing so, the authors expect their insights to help in designing tasks and attracting the best workers to a task. Along the same lines, Ross et al.[RIS+10] studied the demographics and usage behaviors, characterizing workers on Amazon's Mechanical Turk. Complementing such existing works, as well as in contrast, we first focus on task modeling rather than user modeling. Nevertheless, we acknowledge the consideration of both aspects as being essential for improving the effectiveness of microtask crowdsourcing.

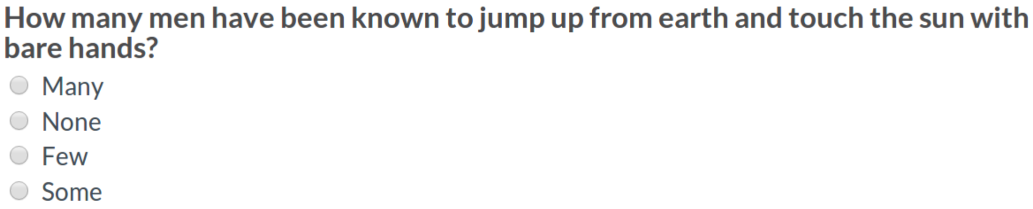
Marshall et al. profiled workers on AMT who take surveys, and examined the characteristics of surveys that may determine the data reliability [MS13a]. Similar to their work, we adopt the approach of collecting data through crowdsourced surveys in order to draw meaningful insights. Yuen et al. presented a literature survey on different aspects of crowdsourcing [YKL11]. In addition to a taxonomy of crowdsourcing research, the authors presented a brief example list of application scenarios. Their short list represents the first steps towards task modeling. However, without proper organization regarding types, goals and workflows, it is hard to reuse such information to devise strategies for task design. We aim to solve this issue by providing an articulated categorization in terms of goals and workflows.

In the realm of studying the reliability of crowd workers, and gauging their performance with respect to the incentives offered, Mason et al. investigated the relationship between financial incentives and the performance of the workers[MW10]. They found that higher monetary incentives increase the quantity but not the quality of work produced by crowd workers. A large part of their results align with our findings presented in Section 2.3. In their work, Geiger et al. proposed a taxonomic framework for crowdsourcing processes[GSS+11]. Based on 46 crowdsourcing examples they conceived a 19-class crowdsourcing process classification. As stated by the authors, they focused exclusively on an organizational perspective, thus providing valuable insights mainly for stakeholders running crowdsourcing platforms. With a different focus, our proposed categorization intends to primarily assist microtask administrators in effectively using such platforms.

Methodology

We aim to analyze tasks that are typically crowdsourced by exploiting response-based data from the workers. We deployed a survey using the CrowdFlower platform in order to gather information about typically crowdsourced jobs. To begin with, the survey consisted of questions regarding the demographics, educational and general background of the workers. Next, questions related to previous tasks that were successfully completed by the workers, were introduced. The survey consisted of a mixture of open-ended, direct, and Likert-type questions designed to capture the interest of the workers. We restricted the participation to 1,000 workers. We asked the crowd workers open-ended questions, about two of their most recent successfully completed tasks. State-of-the-art qualitative research methods [CS08], have indicated that relying on recent incidents is highly effective, since respondents answer such questions with more details and instinctive candor. We pay all the contributors from the crowd, irrespective of whether or not we discard their data for further analysis.

In addition, to keep the participants engaged we interspersed the regular questions with humour-evoking and amusing bits as shown in Figure 2.4. At the same time, these questions are also used to filter out spammers or workers with malicious intentions. We do not use other sophisticated means to curtail regular crowd worker behaviour, in order to capture a realistic composition of workers (both trustworthy and otherwise). However, we do not consider responses from discernibly malicious workers in our data analysis.



How many men have been known to jump up from earth and touch the sun with bare hands?

○ Many
○ None
○ Few
○ Some

Figure 2.4 Engaging workers and checking their alertness by using questions.

2.3 Results and Analysis

In this section we present our findings from the analysis of the data, collected through the crowd sourcing process described earlier. From the 1,000 workers that participated in the survey, we consider the responses from 490 in our analysis. The responses from 433 workers are pruned out of consideration based on their failure to pass at least one of the so called '*gold standard*' tests. A gold standard question, is one that is designed in order to prevent malicious workers from either progressing in a task, or to identify and discard such workers during analysis. For example, consider the question in Figure 2.4. Some workers appear to pick and choose from the available options at random. By using two such hidden tests, we prune out workers with seemingly ulterior intentions.

We manually curated the responses from the remaining workers, and found that 77 workers tried to cheat their way through to task completion by copy-pasting the same bits of text in response to all open-ended questions.

Of the 490 trusted workers, 76% were found to be *Male* participants while 24% were *Female*. The average age of the male and female crowd workers was similar, at 33.7 and 33.1 years respectively. We found that 88% workers cared about their reputation as crowd workers, while 12% workers claimed that they did not care about their reputation. These workers contributed to the description of 980 tasks that they successfully completed in the past. Workers claimed that in hindsight, they could have performed better in 534 of these instances, while they responded that they could not have performed better in 282 of those tasks. Workers were unsure about a possible improvement in their performance corresponding to 164 tasks.

2.3.1 How Workers Choose their Tasks

An interesting research question, which we set out to answer in this chapter, was to find out what factors influence a worker's choice in the tasks she picks to complete. Based on our survey, we gather the three most commonly stated factors that determine a worker's choice in task. An indicator could be the *monetary incentives* offered on task completion. The *interestingness* of the task itself and the *time required* to complete a task are the other factors that surfaced. However, the distribution of importance of these factors is not known. To determine this, we capture the responses from the workers for this question on a 5-point Likert-scale from *No Influence: 1* to *Strong Influence: 5*.

The aggregated results of the Likert-scale show that *monetary reward* (*4.02*) is significantly ($p < .01$) the most crucial factor for a worker while determining which task to complete. The factors *time required* (*3.76*) for the completion of a task and *topic* (*3.69*) come in next with marginal difference between them.

Apart from the factors captured on the Likert-scale, we posed additional questions to the workers regarding why they chose to complete the particular tasks that they described in the survey (the workers' two most recently completed tasks). Figure 2.5 quantifies our findings. The *ease of completion* of a task is a driving force in the task selection process of a worker. An *interesting topic*, a *high reward*, a *less time consuming* task also play a role in the choice of task of a crowd worker, albeit to a less prominent extent. It is interesting to note that a significant number of crowd workers end up completing tasks due to the *lack of other alternatives*. Additionally, we

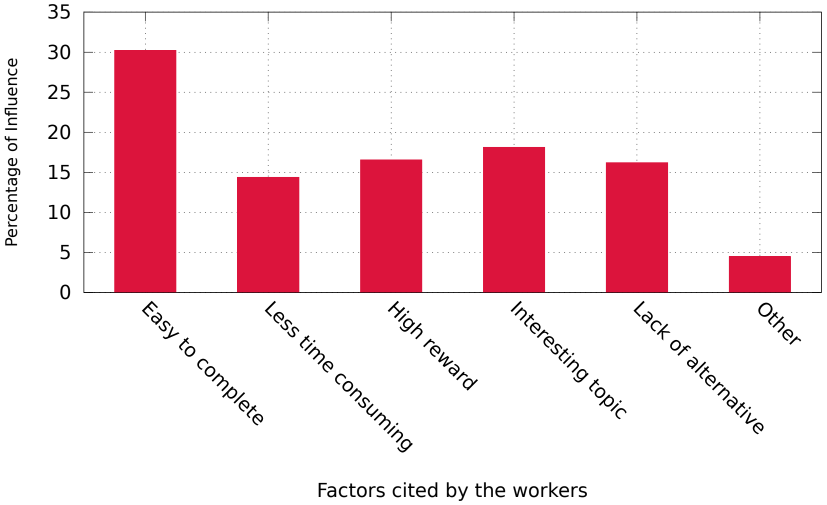


Figure 2.5 Factors that determine workers' choice of task based on their most recently completed tasks.

facilitated for an open-ended response when workers chose the *Other* option. Through this we found a few other minor reasons that workers cited for choosing to complete their previous tasks. For example, a few workers said they wanted to increase their overall profile accuracy, i.e. their reputation on the crowdsourcing platform.

Considering that the monetary reward is highly influential in the workers' choice of completing tasks, it is interesting to investigate the precise amount of rewards offered by the tasks. Figure 2.6 presents the distribution of the money earned by the workers on completion of the 980 tasks, considered in the analysis. We note that most tasks that are deployed for crowdsourcing, offer either meagre ($< 0.5$$) or small monetary rewards (between 0.5$ and 2$). This is reasonable from the point of view of the task deployers or administrators, since most of these tasks do not require a lot of effort from the crowd workers, as confirmed from our analysis (see Figure 2.7).

We clearly see that the tasks that offer bigger monetary awards ($> 3$$) are also incidentally the tasks in which the crowd workers are required to exert the most amount of effort. This suggests that the monetary rewards offered for the tasks that are typically crowdsourced, are proportional to the amount of effort that is expected for task completion by the crowd workers.

2.3.2 Task Affinity vs. Incentive

We define *task affinity* as the tendency of a crowd worker to like the task she chooses to complete. Next, we investigate how the incentive for a given task influences the task affinity of a crowd worker. From our analysis, presented in the Figure 2.7 we observe that there are two subtle kinds of behaviour exhibited by the workers in the

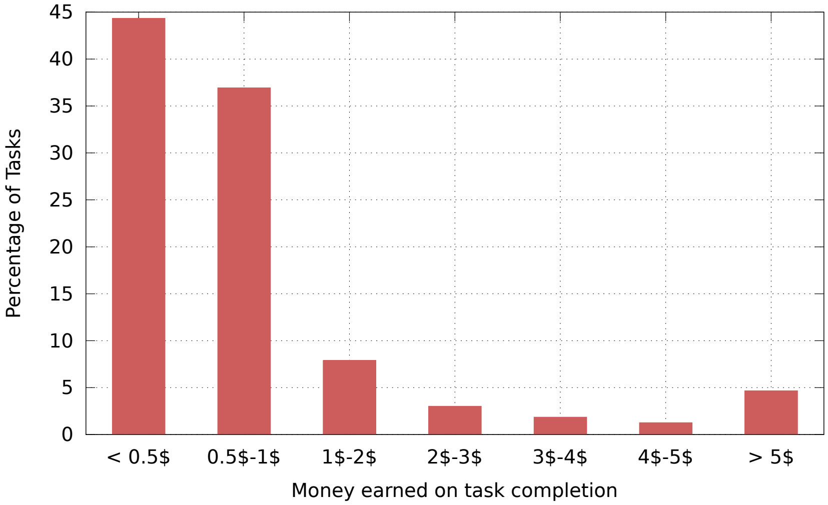


Figure 2.6 Money earned by workers on completing various tasks.

crowd. Crowd workers tend to exhibit a greater affinity to those tasks which offer higher incentives ($>$ 3$). This is understandable since the workers can earn more money by completing such tasks. On the other hand, crowd workers also depict a significant amount of affinity for tasks that offer a reasonable amount of incentive (between 1$ and 2$). This can be explained by the fact that, these tasks require significantly lesser effort from the workers.

An interesting point to note, is that although the most number of tasks deployed on crowd sourcing platforms fall under the bracket of relatively low monetary incentives, thus resulting in such tasks being completed by most workers, the workers' affinity towards these given tasks is considerably low.

We consider that a workers' approval of the monetary reward corresponding to a given task, may be subject to change on task completion. This may be attributed to the difference in the amount of effort needed for task completion when compared to the anticipated effort by a crowd worker. We capture the average satisfaction of the crowd workers with the reward they receive on task completion. Our aggregated results are presented in Figure 2.7. We observe that the satisfaction is proportional to the incentive of the reward that is offered.

2.4 Categorization of Tasks

From the responses collected through the crowdsourced survey, we manually established the following classes that describe typically crowdsourced tasks. The example task descriptions presented alongside each type of task, are extracted from the responses received from workers regarding their previous microtasks. We categorize the tasks into 6 high-level goal-oriented classes as presented next, with each class

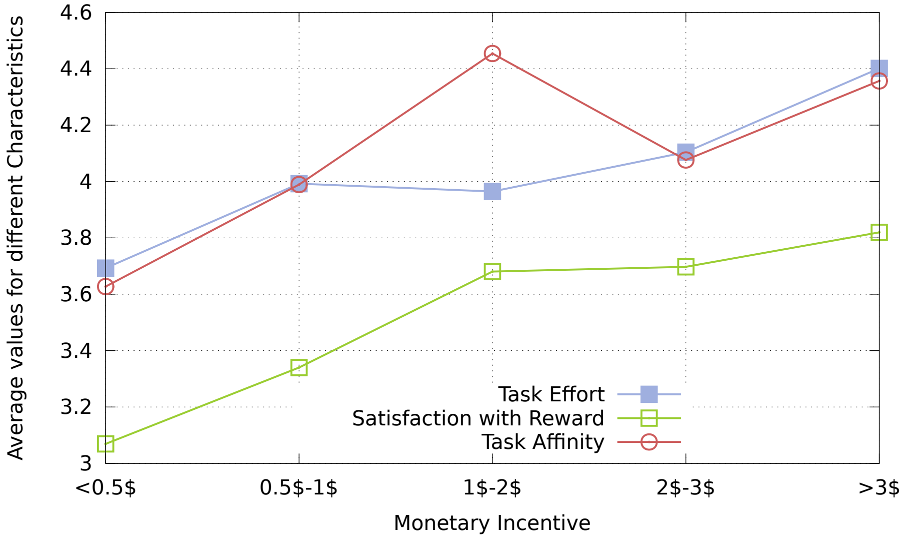


Figure 2.7 Distribution of effort required, task affinity, and satisfaction with reward of the workers with respect to varying task incentives.

containing sub-classes of other types of tasks. The high-level categorization is drawn based on the '*goals*' of a task (i.e., the overall objective of the task), while the sub-classes are based on the '*workflow*' of tasks (i.e., the steps required to be carried out to complete the task successfully).

2.4.1 Categorization Scheme

- **Information Finding**(IF)- Such tasks delegate the process of searching to satisfy one's information need, to the workers in the crowd. For example, '*Find information about a company in the UK*', or '*Find the cheapest air fare for the selected dates and destinations*'.

- **Verification and Validation**(VV)- These are tasks that require workers in the crowd to either verify certain aspects as per the given instructions, or confirm the validity of various kinds of content. For example, '*Is this a Spam Bot? : Check whether the twitter users are either real people or organisations, or merely spam twitter user profiles*', or '*Match the names of personal computers and verify corresponding information*'.

- **Interpretation and Analysis**(IA)- Such tasks rely on the wisdom of the crowd to use their interpretation skills during task completion. For example, '*Choose the most suitable category for each URL*', or '*Categorize reviews as either positive or negative*'.

- **Content Creation**(CC)- Such tasks usually require the workers to generate new content for a document or website. They include authoring product descriptions or producing question-answer pairs. For example, '*Suggest names for a new product*', or '*Translate the following content into German*'.

- **Surveys**(S)- Surveys about a multitude of aspects ranging from demographics to customer satisfaction are crowdsourced. For example, '*Mother's Day and Father's Day Survey (18-29 year olds only!)*'

- **Content Access**(CA)- These tasks require the crowd workers to simply access some content. For example, '*Click on the link and watch the video*', or '*Read the information by following the website link*'. In these tasks the workers are merely asked to consume some content by accessing it, but do nothing further.

It is important to note that, in certain cases it may be possible for a particular job to belong to more than one of the aforementioned classes. For example, a survey about the perception of a product like the new iPhone from Apple could belong to the classes of *Surveys* as well as *Sentiment Analysis*.

Apart from these high-level categorization based on the goals of the tasks, Table 2.1 presents some sub-classes of the high-level classes, which are based on the workflow of the tasks. Some sub-classes are explained below.

- Class *IF*/**Metadata Finding**- Such tasks require the users to find specific relevant information from a given data source. For example, '*Find e-mail addresses of corresponding employees from the company's websites*'.

- Class *VV*/**Content Verification**- In these tasks the crowd workers are required to verify, validate, qualify, or disqualify different aspects as dictated by the task administrators. For instance, '*Check if the following company websites describe the correct business*'.

- Class *IA*/**Categorization and Classification**- Such tasks involve the organization of entities into groups with the same features, or assigning entities to classes according to a predetermined set of principles. For example, '*Choose the most suitable category for each URL*'.

- Class *IA* or *CC*/**Media Transcription**- These tasks require the crowd workers to transcribe (put into written form) different media like images, music, video, and so forth. For example, '*See the images and find the year on which the wine bottle was manufactured*'. The tasks also include transcribing captchas. For instance, '*Type what you see in the following Captchas*'.

- Class *IA*/**Ranking**- Here the crowd workers are required to determine the most relevant entities with respect to the search query. For example, '*Search for the given terms and click on the best three results*'.

- Class *IA*/**Content Moderation**- Here the workers are required to moderate content for guideline violations, inappropriate content, spam, or others. Independent of the kind of media (text, photos, or videos), the crowd is asked to evaluate the content against a set of rules. For example, '*Moderate images for inappropriate content (sexually explicit content)*'.

- Class *IA*/**Sentiment Analysis**- Tasks that pertain to the assessment of the sentiment towards an entity or notion, fall under this category. For example, '*What do you think of the new Samsung tablet?*', or '*Identify if the tweets are positive, negative, or neutral*'.

- Class *CC*/**Data Collection and Enhancement**- Crowdsourcing is used to generate and enhance data. For instance, the crowd has been used in the past to create a dataset of colours by asking workers to annotate different hues and shades with labels[5].

- Class *S*/**Content Feedback**- In such tasks workers are asked to assess and provide feedback about products, entities, websites, and so forth. For example, '*Help us improve our website*'.

- Class *CA*/**Promoting**- In such tasks workers are asked to access and consume content. For example, '*Visit the webpage by clicking on the provided link*'.

[5] http://www.crowdflower.com/blog/2008/03/our-color-names-data-set-is-online

Table 2.1 Sub-classes of the proposed categorization for typically crowdsourced tasks.

Information Finding	**Verification** & **Validation**	**Interpretation** & **Analysis**	**Content Creation**	**Surveys**	**Content Access**
Metadata finding	Content Verification	Classification	Media Transcription	Feedback/Opinions	Testing
	Content Validation	Categorization	Data Enhancement	Demographics	Promoting
	Spam Detection	Media Transcription	Translation		
	Data matching	Ranking	Tagging		
		Data Selection			
		Sentiment Analysis			
		Content Moderation			
		Quality Assesment			

2.4.2 Tasks as Per Categorization Scheme

Based solely on the reliable data collected during the crowdsourcing process, we manually annotated each of the workers' previously completed tasks according to the categorization scheme. Figure 2.8 presents the distribution of these tasks, as per their categorization into the different proposed classes.

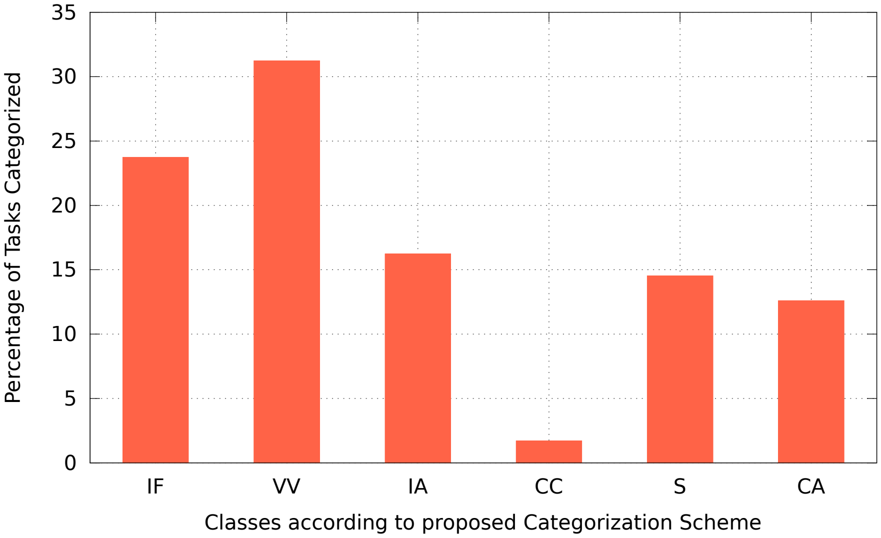


Figure 2.8 Distribution of tasks in the following classes of the proposed categorization scheme; *IF:* Information Finding, *VV:* Verification and Validation, *IA:* Interpretation and Analysis, *CC:* Content Creation. *S:* Surveys, and *CA:* Content Access.

Note that certain tasks can rightly be classified into more than one class. For example, consider the task, '*Search for spam-like comments in the following content*'. This task can be classified into the classes *Verification and Validation*, as well as *Information Finding*. This is because, the goal of such a task could be either to ensure the content is spam-free, or merely find the spam comments. Consider the task, '*Identify biographies in the following*'. This task can be classified into the class *Interpretation and Analysis* since the identification of a biography relies on the workers interpretation of the classification. At the same time, the task can be classified into *Verification and Validation*, since the goal of the task could be to validate biographies.

As a next step, we analyze the average effort that a worker needs to exert to complete a task, the task affinity, as well as the workers' satisfaction with the reward, for each of the high-level classes. The findings are presented in Figure 2.9. Understandably, tasks of the class *Content Creation* require the most amount of effort from the crowd workers, while those of the class *Content Access* require the least amount of effort. It is interesting to note that crowd workers like to work on tasks of the class

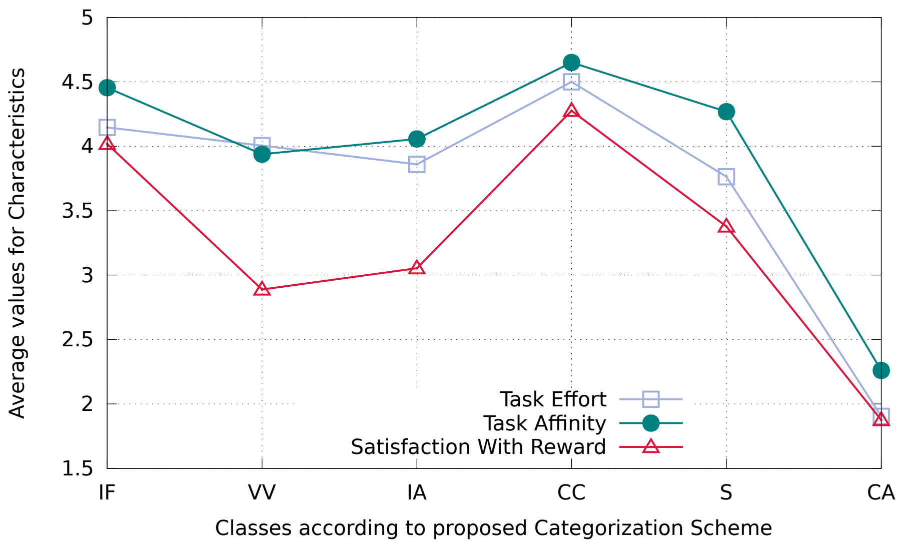


Figure 2.9 Distribution of task-related characteristics according to the proposed categorization scheme; *IF:* Information Finding, *VV:* Verification and Validation, *IA:* Interpretation and Analysis, *CC:* Content Creation. *S:* Surveys, and *CA:* Content Access.

type *Information Finding* and *Surveys* in addition to *Content Creation*. The most disparity between the effort exerted for task completion by the workers and their satisfaction with the reward, corresponds to the classes of *Verification and Validation*, and *Interpretation and Analysis*.

2.4.3 Tasks with Ulterior Motives

During our manual analysis of the data collected we identified several tasks with deceitful hidden motives. While such tasks may indicate legitimacy to some extent due to the workflow they suggest to workers, there is a clear ulterior goal of deliberately manipulating third party results. For example, improving the popularity or general sentiment of particular content. In most cases, these tasks fall in the classes *Content Access* and *Content Creation*, further being masked with an additional goal such as a *Survey*. For example, '*Search for some particular terms in Google, and click on the link of our Website*', '*Watch this video on Youtube and click like*', or '*Give a five start rating to this product*'. This is consistent with prior findings by Wang et al., who define and study such campaigns in detail [WWZ$^+$12].

We also verified that in many circumstances, these tasks are followed by a survey which contains a failure guaranteed gold standard. For example, '*What's your age?*', whereas the only correct answer is an unrealistic number. At this point, the malicious task administrator has already collected his desired data and the system prevents workers from getting their reward. As a principle, crowdsourcing platforms discourage the deployment of such tasks. We believe that the categorization scheme can help improve the identification of deceitful tasks deployed by requesters. This is a critical issue to be addressed in order to improving crowdsourcing practice.

2.5 Chapter Summary

In this chapter, we briefly introduced the main actors in the microtask crowdsourcing paradigm, and presented a meta-crowdsourcing profiling study. We found high levels of potentially unreliable workers in our study; almost 44% of the workers did not manage to correctly answer simple attention check questions. It is important to highlight that we deliberately model our crowdsourcing task to allow such behavior. These results highlight the importance of quality control mechanisms in microtask crowdsourcing and need for more effective task design strategies.

Based on the manually verified reliable responses, we collected sufficient data to characterize the behavior of workers and their preferences. Further, we thoroughly studied the types of tasks that are typically crowdsourced, and as a result, we proposed a goal-oriented *categorization scheme* for crowdsourced tasks. A fine-grained categorization of crowdsourced tasks has important implications for the user modeling of crowd workers, and recommendation of tasks. The proposed categorization of tasks, including the findings from our extensive analysis, aid in future task design and deployment. For instance, a recent longitudinal analysis of the Amazon Mechanical Turk marketplace by Difallah et al. analyzed how different types of HITs according to our proposed taxonomy evolved over time [DCD$^+$15b], serving as a vital tool to understand the dynamics of the marketplace. By drawing from our findings related to the task dependent characteristics, like task affinity, task effort, incentive required, and so forth, one can design tasks with higher success rates (i.e., maximizing the quality of the results with respect to the given reward).

3 Understanding Worker Behavior

> *"Behavior is the mirror in which everyone shows their image."*
> — *Johann Wolfgang von Goethe*

Crowdsourcing is increasingly being used as a means to tackle problems requiring human intelligence. With the ever-growing worker base that aims to complete microtasks on crowdsourcing platforms in exchange for financial gains, there is a need for stringent mechanisms to prevent exploitation of deployed tasks. Quality control mechanisms need to accommodate a diverse pool of workers, exhibiting a wide range of behavior. A pivotal step towards fraud-proof task design is understanding the behavioral patterns of microtask workers. In this chapter, we analyze the prevalent malicious activity on crowdsourcing platforms and study the behavior exhibited by trustworthy and untrustworthy workers, particularly on crowdsourced surveys. Based on our analysis of the typical malicious activity, we define and identify different types of workers in the crowd, propose a method to measure malicious activity, and finally present guidelines for the efficient design of crowdsourced surveys.

3.1 Introduction

Over the last decade, crowdsourcing has gained rapid popularity, because of the data-intensive nature of emerging tasks, requiring validation, evaluation and annotation of large volumes of data.

With the ubiquity of the internet, it became possible to distribute tasks at global scales, leading to the recent success of crowdsourcing, being later defined as an 'online, distributed problem-solving and production model [Bra08].'

In the recent past, there has been a considerable amount of work towards developing appropriate platforms and suggesting frameworks for efficient crowdsourcing. Amazon's Mechanical Turk[1], and CrowdFlower[2] are good examples of such platforms. An increasing number of research communities benefit from using crowdsourcing platforms in order to either gather distributed and unbiased data [NTDM14], to validate results, evaluate aspects, or to build ground truths [KKMF13a].

While the demand for using crowdsourcing to solve several problems is on an upward climb, there are some obstacles that hinder requesters from attaining reliable, transparent, and non-skewed results. Herein, a primary nuisance is introduced through *malicious workers*, understood by [IPW10, EdV11, GGP10] as workers with ulterior motives, who either simply sabotage a task or try to quickly attain task completion for monetary gains.

Gold-standards are the typically adopted solution to improve task performance. In general practice, gold-standards are questions where answers are known apriori to the task administrators. Thus, if a worker fails to provide the correct answer for a particular question, he is automatically flagged as an *untrustworthy worker*[3]. However, with the flourishing crowdsourcing market, we believe that malicious activities and adversarial approaches will also become more advanced and popular, overcoming common gold standards. Quality control mechanisms should thereby account for a diverse pool of workers that exhibit a wide range of behavioral patterns. Methods have been designed and used in order to tackle poor worker performance in the past [DDCM12b, DKKH12]. However, there is a need to understand the behavior of these workers and the kinds of malicious activity they bring about in crowdsourcing platforms. In this chapter, we present our work towards analyzing the behavior of malicious microtask workers, and reflect on guidelines to overcome such workers in the context of online surveys. An *online survey* is a questionnaire that can be completed over the Internet by a target audience.

We deployed a survey to 1000 workers in the crowd, and present evidence that a large number of workers are untrustworthy. This evidence shows that simple gold-standards might not be enough to provide reliable data or results. Then we conducted an analysis of both trustworthy and untrustworthy workers; we classified the behavior of the workers based on the different types of activity exhibited. To gain further insights into the prevalence of different kinds of malicious workers in the crowd, experts manually and exhaustively annotated the workers into established classes.

The main contributions of our work are listed below.

- Resulting from our analysis of workers, we present different types of malicious behavior exhibited in the crowd. This understanding of the prevalent kinds of malicious activity will be an aid in future task design.
- We suggest a novel method to measure the *maliciousness* of a worker based on

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

[2] http://www.crowdflower.com/

[3] Note that being an untrustworthy worker does not necessarily imply being a malicious worker.

the acceptability of her responses.

- We present a detailed analysis of the flow of malicious behavior of workers throughout the task, and define a *tipping point* which marks the starting point of a workers' malicious tendency.

- Finally, we propose a set of guidelines for the efficient, fraud-proof task design of *surveys*.

3.2 Related Literature

Quality and Reliability of Workers

Behrend et al. showed the suitability of crowdsourcing as an alternative data source for organizational psychology research [BSMW11]. Kittur et al. promoted the suitability of crowdsourcing user studies, while cautioning that special attention should be given to the task formulation [KCS08]. Although these works outline shortcomings of using crowdsourcing, they do not consider the impact of malicious activity that can emerge in differing ways. In our work, we show that varying types of malicious activity is prevalent in crowdsourced surveys, and propose measures to curtail such behavior.

Marshall et al. profiled Turkers who take surveys, and examined the characteristics of surveys that may determine the data reliability [MS13a]. Similar to their work, we adopt the approach of collecting data through crowdsourced surveys in order to draw meaningful insights. Our analysis quantitatively and qualitatively extends their work, and additionally provides a sustainable classification of malicious workers that sets precedents for an extension to different categories of microtasks.

Through their work, Ipeirotis et al. motivated the need for techniques that can accurately estimate the quality of workers, allowing for the rejection or blocking of low-performing workers and spammers [IPW10]. The authors presented algorithms that improve the existing techniques to enable the separation of bias and error rate of the worker. Baba et al. reported on their study of methods to automatically detect improper tasks on crowdsourcing platforms [BKK+13]. The authors reflected on the importance of controlling the quality of tasks in crowdsourcing marketplaces. Complementing these existing works, our work propels the consideration of both aspects (task design as well as worker behavior), for effective crowdsourcing.

Dow et al. presented a feedback system for improving the quality of work in the crowd [DKB+11]. Oleson et al. present a method to achieve quality control for crowdsourcing, by providing training feedback to workers while relying on programmatic creation of gold data [OSL+11]. However, for gold-based quality assurance, task administrators need to understand the behavior of malicious workers and anticipate the likely types of worker errors with respect to different types of tasks. Understanding the behavior of workers, is therefore an important objective of this chapter.

In the realm of studying the reliability and performance of crowd workers with respect to the incentives offered, Mason et al. investigated the relationship between financial incentives and the performance of the workers [MW10]. They found that higher monetary incentives increase the quantity of workers but not the quality of work. A large part of their results align with our findings presented in the following sections.

Worker Traits, Tasks Design and Metrics

Researchers in the field have acknowledged the importance and need for techniques to deal with inattentive workers, scammers, incompetent and malicious workers.

Ross et al. studied the demographics and usage behaviors characterizing workers on Amazon's Mechanical Turk [RIS$^+$10]. Kazai et al. defined types of workers in the crowd by type-casting workers as either *sloppy*, *spammer*, *incompetent*, *competent*, or *diligent* [KKMF11]. By doing so, the authors expect their insights to help in designing tasks and attracting the best workers to a task. The authors use worker-performance to define these types, while we delve into the behavioral patterns of workers.

Wang et al. presented a detailed study of *crowdturfing systems*, which are dedicated to organizing workers to perform malicious tasks [WWZ$^+$12]. While the authors of this paper investigated systems solely dedicated to malicious activities, in our work, we explore and analyze the prevalence of malicious workers and activities on regular crowdsourcing platforms. In their work, Eickhoff et al. aimed to identify measures that one can take in order to make crowdsourced tasks resilient to fraudulent attempts [EdV11]. The authors concluded that understanding worker behavior better is pivotal for reliability metrics. Understanding malicious workers, is in fact the main goal of this chapter.

Difallah et al. reviewed existing techniques used to detect malicious workers and spammers and described the limitations of these techniques [DDCM12b]. Buchholz and Latorre proposed metrics for the post-hoc exclusion of workers from results [BL11]. In another relevant work by Eickhoff et al., the authors proposed to design and formulate microtasks such that they are less attractive for cheaters [EdV13]. In order to do so, the authors evaluated factors such as the type of microtask, the interface used, the composition of the crowd, and the size of the microtask. While our work presented in this chapter complements the prior work done by Eickhoff et al. it is significantly different, in that we investigate the behavioral patterns of trustworthy and untrustworthy workers, and suggest remedies to detect and inhibit their prominence based on the specific type of behavior. Notably, we introduce novel metrics such as *maliciousness* of a worker, to quantify the behavioral patterns thus observed.

Yuen et al. present a literature survey on different aspects of crowdsourcing

[YKL11]. In addition to a taxonomy of crowdsourcing research, the authors present a humble example list of application scenarios. Their short list represents the first steps towards task modeling. However, without proper organization regarding types, goals and work-flows, it is hard to reuse such information to devise strategies for task design. As a step forward, in earlier work Gadiraju et al. proposed a comprehensive and exhaustive taxonomy for the different types of microtasks [GKD14c]. By studying the various kinds of behavior exhibited by trustworthy and untrustworthy workers in the crowd, in this work, we present a closer and detailed understanding of workers that will aid in developing anti-adversarial techniques.

3.3 Background

We build on the previous work done by Gadiraju et al. [GKD14c], where the authors analyzed the nature of crowdsourced tasks. Firstly, the rationale behind the choice of workers to complete a job and the nature of the jobs themselves were studied. *Monetary reward* was found to be the most crucial factor that motivates workers across different task types, in their choice to complete a task. Additionally, *ease of completion* of a task is a driving force in the task selection process of a worker. An interesting topic, a high reward, a less time consuming task also play a role in the choice of task of a worker in the crowd, albeit to a less prominent extent.

Secondly, a generic umbrella classification of microtasks, which is conceptualized based on the final goal of the tasks was proposed. This goal-oriented taxonomy splits the tasks into six high-level categories:

- *Information Finding*: tasks that require workers to simply find pieces of information by following instructions.
- *Verification and Validation*: tasks that require workers to verify certain aspects as per the given instructions.
- *Interpretation and Analysis*: tasks that require workers to provide information that is subject to their individual interpretation.
- *Content Creation*: tasks that require workers to generate new content.
- *Surveys*: tasks that require workers to answer several questions based on their opinion and background.
- *Content Access*: tasks that require workers to simply access some online content.

This top classification encompasses different kinds of microtasks that vary according to specific goals, however, at this level the classification is considered to be exhaustive. From the analysis of Gadiraju et al. [GKD14c], we learnt that microtask workers are dictated by their top priorities; to maximize monetary gain and minimize effort. In particular, the indifference of workers towards their reputation leads to many microtask workers becoming malicious. It is clear that many workers attempt to exert a minimum amount of effort to receive their reward. Unfortunately, in many cases the minimum effort is not enough for a task administrator to accumulate good or even acceptable results. What is equally clear, is that a task administrator must therefore try to prevent alternatives that allow workers to receive their rewards without providing valid results, i.e. prevent *cheating*.

Based on this prior knowledge of the workers' preferences, and the taxonomy of microtasks, in this chapter we analyze the malicious behavior of workers in a specific type of microtasks: *Surveys*. We specifically choose to study this category, since surveys present the most difficult challenges with respect to ensuring accurate responses from workers. This is due to the inherently subjective nature of most surveys. Thus, *gold standards* cannot be applied easily. For example, in an '*Information Finding*' task, the task administrator might typically be able to ensure the validity of workers' responses by employing questions for which the answers are priorly known (gold standard). Thus, verifying the character of the worker. On the other hand, in a simple demographic survey, which is subject to receive multiple valid responses for a single question from the target audience, such practice is infeasible.

In this work, we aim to address research questions (RQs) by using the following definitions.

Definition 1. *Malicious workers* are workers deemed to have ulterior motives that deviate from the instructions and expectations as defined a priori by the microtask administrator.

Definition 2. *Untrustworthy workers* are workers who provide wrong answers in response to one or more simple and straightforward attention-check or gold standard questions.

RQ #1: Do untrustworthy workers adopt different methods to complete tasks, and exhibit different kinds of behavior?

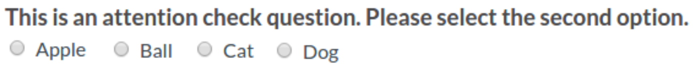
RQ #2: How can task administrators benefit from the prior knowledge of plausible worker behavior?

RQ #3: Can behavioral patterns of malicious workers in the crowd be identified and quantified?

3.4 Data Analysis

In this section, we first plot general results of the crowdsourced task. Later, we classify the behavior of trustworthy and untrustworthy workers in the crowd. We identify 432 untrustworthy workers by using test questions similar to the one depicted in Figure

3.1, who fail to pass at least one of two simple questions. These untrustworthy workers are then studied further in comparison to trustworthy workers, to determine plausible malicious traits.



This is an attention check question. Please select the second option.

○ Apple ○ Ball ○ Cat ○ Dog

Figure 3.1 Attention check question to identify untrustworthy workers.

3.4.1 Where Are the Workers From?

The Crowdflower platform forwards tasks to several different third-party crowdsourcing platforms, called '*channels*'[4]. In order to achieve coverage and results that are representative of the general crowdsourcing market, we do not impose any restrictions with respect to the channels. 50% of the workers who participated in the task used the 'Neodev' channel, while almost 25% of the workers used 'Clixsense'.

Since our survey was deployed in English, the first restriction we enforce via the platform is the language of the worker. However, it is difficult to accurately tell whether a worker is proficient in a given language. A simple workaround provided by the platforms exploits the location of the workers. Although imperfect, it is a reasonable assumption that a person located in an English speaking country, e.g. United States or Australia, is proficient in English (at least to an extent to understand and respond correctly to the questions in the task). Figure 3.2 shows the country distribution of the workers who participated in our task. We can observe that India leads by a large margin, followed by the USA, and Pakistan.

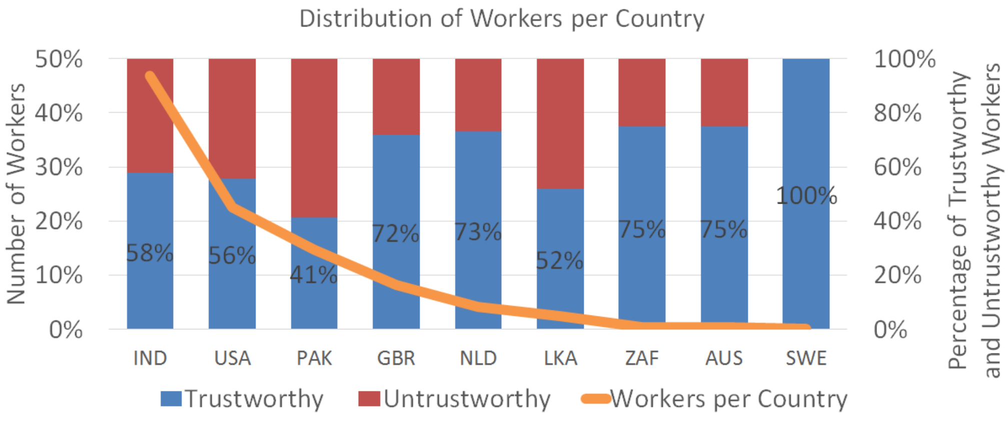


Figure 3.2 Distribution of the workers per country (left axis), and the distribution of trustworthy and untrustworthy workers (right axis).

[4] http://www.crowdflower.com/labor-channels

These numbers are anticipated since crowdsourcing is renown for widely employing workers from developing countries. In Figure 3.2, we divide the workers into two groups, trustworthy and untrustworthy, solely based on their responses to the attention check questions. In terms of percentage, we see that Pakistan, Sri Lanka, USA, and India lead in the number of workers who did not pass the attention checks.

Several hypotheses could be raised from these results. However further analysis of the influence of demographics, political, and economic factors are out of the scope of our work here. We are interested in analyzing and uncovering the universal user behavior that leads us to a coherent understanding of malicious activities. This can therefore provide us the required competence to restrict such malicious activity.

3.4.2 Analyzing Malicious behavior in the Crowd

Prior research has shown that by devising typologies, we can provide a better structure to organize knowledge and study the relationships between disorderly concepts [GV95]. We analyze the implicit behavioral patterns of malicious workers by the means of their responses. Based on aspects such as (i) the eligibility of a worker to participate in a task, (ii) whether responses from a worker conform to the preset rules, or (iii) whether responses fully satisfy the requirements expected by the administrator, we determine the following types of behavioral patterns.

- **Ineligible Workers** (*IE*). Every microtask that is deployed on crowdsourcing platforms presents the workers in the crowd with a task description and a set of instructions that the workers must follow, for successful task completion. Those workers who do not conform to the priorly stated pre-requisites, belong to this category. Such workers may or may not provide valid responses, but their responses cannot be used by the task administrator since they do not satisfy the pre-requisites. For example, consider a pre-requisite in our survey, '*Please attempt this microtask ONLY IF you have successfully completed 5 microtasks previously*'. We observed that some workers responded to questions regarding their previous tasks with, '*this is my first task*', clearly violating the pre-requisite.

- **Fast Deceivers** (*FD*). Malicious workers tend to exhibit a behavior that is strongly indicative of the intention to earn easy and quick money, by exploiting microtasks. In their attempt to maximize their benefits in minimum time, such workers supply ill-fitting responses that may take advantage of a lack of response-validators. These workers belong to the class of *fast deceivers*. For example, workers who copy-paste the same response for different questions. In our survey, some workers copy-pasted the title of our survey, '*What's your task?*', in response to several unrelated open-ended questions. Some others simply entered gibberish such as '*adasd*', '*fygv fxc xdgj*', and so forth.

- **Rule Breakers** (*RB*). Another kind of behavior prevalent among malicious workers is their lack of conformation to clear instructions with respect to each response. Data collected as a result of such behavior has little value. For instance, consider the question from our survey, '*Please identify at least 5 keywords that represent this task*'. In response, some workers provided fewer keywords. In such cases, the resulting response may not be useful to the extent intended by the task administrator.

- **Smart Deceivers** (*SD*). Some eligible workers that are malicious, try to deceive the task administrators by carefully conforming to the given rules. Such workers mask their real objective by simply not violating or triggering implicit validators. For example, consider the instruction, '*Transcribe the words in the corresponding image and separate the words with commas*'. Here, workers that intentionally enter unrelated words, but conform to the instructions by separating the words with commas, may neutralize possible validators and achieve successful task completion. While this type of workers behave to an extent like *fast deceivers*, the striking difference lies in the additional attempt of *smart deceivers* to hide their real goal and bypass any automatic validating mechanisms in place. In our survey, some workers provided irrelevant keywords such as '*yes, no, please*', '*one, two, three*', and so forth to represent their preferred task-types. Some of these workers take special care to avoid triggering attention-check or gold standard type questions.

- **Gold Standard Preys** (*GSP*). Some workers who abide by the instructions and provide valid responses, surprisingly fail to surpass the gold standard questions. They exhibit non-malicious behavior, only to be tripped by one or more of the gold standard test questions. This may be attributed to the inattentiveness of such workers.

568 workers passed the gold standard questions (*trustworthy*) and 432 workers failed to pass at least one of the two test questions (*untrustworthy*). On analyzing each response from the workers, we found that only 335 of the trustworthy workers gave perfect responses (*elite workers*). A panel of 5 experts were presented the responses of each worker from the remaining 665 non-elite workers (233 trustworthy and 432 untrustworthy workers), and they manually classified the workers into the different classes, according to the class behavioral patterns described earlier. The inter-rater agreement between the experts during the classification of workers as per Krippendorffs Alpha is 0.94. Based on majority voting and the agreement between the experts, we finalize the worker classification without discrepancies. 73 untrustworthy workers and 93 trustworthy workers were classified into 2 different classes, while the rest were classified into unique classes. Note that a worker can depict different kinds of behavior and thereby belong to multiple classes. Figure 3.3 presents the experts' classification of these workers as per the different types of behavioral patterns.

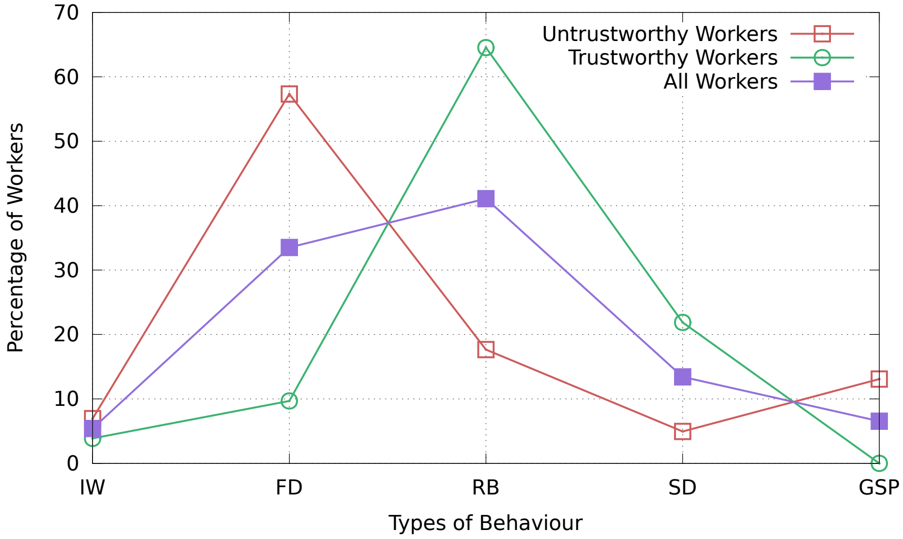


Figure 3.3 Distribution of non-elite workers as per their behavior.

More than 70% of all 665 workers classified are either *rule breakers* or *fast deceivers*. Nearly 60% of untrustworthy workers are *fast deceivers*, who intend to bypass response validators in order to earn monetary rewards easily. This is consistent with the findings of Kaufmann et al., wherein the authors establish that the number of workers who are mainly attracted by monetary rewards represent a significant share of the crowd [KSV11a]. About 65% of all non-elite trustworthy workers are *rule breakers*, who do not conform to the instructions laid out by the task administrators and thereby provide partially correct or limitedly useful responses.

The third most prevalent kind of untrustworthy workers are *smart deceivers*. Around 13% of all the classified workers take cautious steps in order to deceive task administrators and achieve task completion. These are malicious workers that tend to slip through most of the existing automated standards to prevent malicious activity, since they take special care to deceive the task administrators and receive the rewards at stake. This is made evident by the fact that over 20% of the non-elite trustworthy workers are *smart deceivers*, who give poor responses despite passing the gold standard questions.

Over 6% of all workers, seem to have failed the gold standard due to a lack of alertness (*gold standard preys*). This implies that a portion of workers' responses can be useful although the workers are deemed to be untrustworthy. Therefore, methods to identify and detect gold standard preys can benefit in maximizing the value of responses. This can be achieved either in a post-processing manner, or on the fly, at a relatively small additional cost.

Around 2.5% of workers attempt and complete tasks despite being clearly ineligible to take part (*ineligible workers*), as dictated by the pre-requisites. In our survey, such workers responded in languages other than in English, or in some cases claimed to have not completed any tasks before, thereby violating clearly stated pre-requisites.

3.4.3 Measuring Maliciousness of Workers

Next, we aim to measure the *maliciousness* of workers, as indicated by the acceptability of their individual responses. Note that this notion of *maliciousness* solely refers to the observable behavior exhibited by workers and is not intended to cast the persons as being generally malicious.

Definition 3. The *acceptability* of a response can be assessed based on the extent to which a response meets the priorly stated expectations.

For example, consider the question, '*Enter the names of any 5 colors (separated by commas).*' A fully acceptable response to this question would be one which contains the names of 5 colors separated by commas (awarded a score of '1'). An *unacceptable* response on the other hand, is one which does not meet the requirements at all (awarded a score of '0'). So, in case of the same example, a response which does not contain names of colors would be completely unacceptable.

An important aspect to consider when measuring the maliciousness of a worker is interpreting the responses of the worker accurately. This means that we cannot reliably include subjective responses from the workers in such an analysis. For instance, consider a question with multiple check-box options; any combination of responses to such a question may be acceptable. This means that in order to perform a reliable analysis, we have to consider only those responses with unambiguous corresponding acceptability. Therefore, we measure the maliciousness of workers by exploiting the acceptability of their responses to open-ended questions.

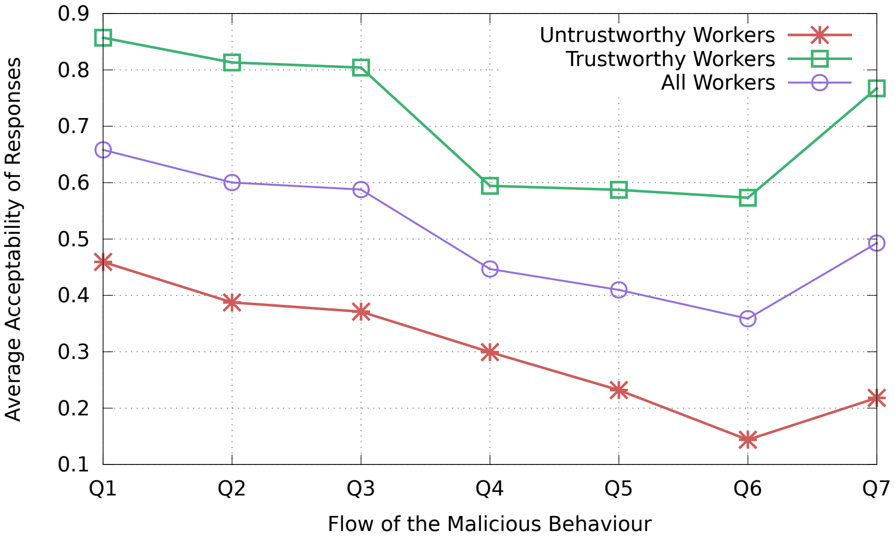


Figure 3.4 Average acceptability of responses from workers for each open-ended question Q.

Experts manually annotated the responses from each worker for every open-ended question as either *acceptable* or *unacceptable*. The agreement between the experts was found to be 0.89 as per Krippendorf's Alpha. Figure 3.4 presents the average acceptability of workers' responses with respect to each open-ended question. Note that the questions *Q1*, and *Q4* ask workers to share the titles of their previously completed tasks on crowdsourcing platforms. *Q2* and *Q5* correspond to the description of these tasks, while *Q3* and *Q6* correspond to keywords representing these tasks. In the last open-ended question (*Q7*), the workers are asked to provide keywords pertaining to tasks that they prefer.

Since we do not randomize the order of the questions for the different workers, we do not draw insights about the trend in acceptability through the course of the survey. However, we clearly observe that the acceptability of responses of the malicious workers reduces with the increase in required input from the workers (studied in literature as *task effort* [GKD14c]). It is easier for the workers to pass off a *title* as acceptable, than doing the same with either the *description* or *keywords* describing the task. On the whole, our findings indicate that the acceptability of individual responses of malicious workers decreases with an increase in the effort required to provide suitable responses.

Based on the acceptability of each response from a worker, we can compute the *average acceptability* (**A**) of a given worker pertaining to a task. In order to do so, we score each *acceptable response* with 1, and each *unacceptable response* with 0. Finally, we compute the *maliciousness* (**M**) of a worker using the following formula.

$$M_{worker} = 1 - (1/n\sum_{i=1}^{n} A_{r_i})$$

where,

n is the number of responses from the worker which are assessed, and A_{r_i} is the acceptability of response r_i.

$M_{worker} = 0$ indicates a completely non-malicious worker, while a worker is said to exhibit complete maliciousness if $M_{worker} = 1$. Figure 3.5 presents our findings regarding the distribution of workers with respect to the degree of their maliciousness, segmented by trustworthiness. In addition, the figure also depicts the corresponding average task completion time of the workers.

We can see that 50% of the untrustworthy workers exhibit a very strong maliciousness degree (greater than 0.8) while most of trustworthy workers (56%) have very low maliciousness. Nearly 20% of the untrustworthy workers exhibit a maliciousness degree between 0.4 and 0.6, while almost 15% indicate a high degree of maliciousness between 0.6 and 0.8. In addition, we observe that the average task completion time of untrustworthy workers decreases with the increasing maliciousness. The same is observed for the trustworthy workers, where the group with highest maliciousness has the lowest average times. We find that for untrustworthy workers, the maliciousness and average task completion time show a high correlation of 0.51, as measured using

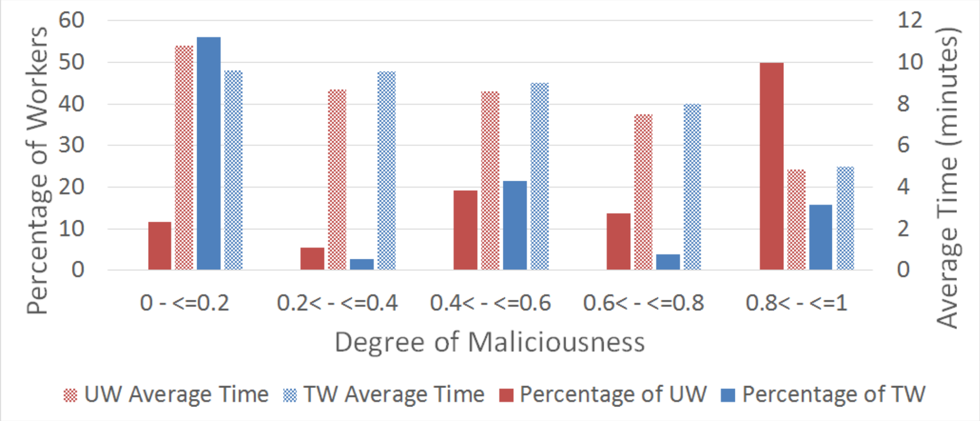


Figure 3.5 Degree of maliciousness of trustworthy (TW) and untrustworthy workers (UW) and their average task completion time.

Pearson Correlation. For trustworthy workers this correlation is moderate at 0.37.

3.4.4 The Tipping Point

In our study of the trustworthy and untrustworthy workers, we find that several workers provide acceptable responses to begin with, before depicting malicious behavior. We thereby investigate this tendency of workers to trail off into malicious behavior, and present our findings here.

Definition 4. We define the first point at which a worker begins to exhibit malicious behavior after having provided an acceptable response, as the *tipping point*. The tipping point can be determined in terms of the number of responses at which the worker exhibits the first sign of malicious activity. In our analysis, we consider the open-ended questions. Note that we do not consider the workers who begin with providing unacceptable responses (we find 233 such untrustworthy workers, and 81 such trustworthy workers).

Figure 3.6 presents our findings. We can see that over 30% of all workers have a tipping point at their second response (R-2). This is largely due to the finding that almost 40% of all untrustworthy workers have a tipping point at R-2. Another 30% of all workers have a tipping point at their fourth response (R-4). Trustworthy workers largely contribute to this case. Nearly 60% of the non-elite trustworthy workers have a tipping point at R-4. On further analysis, we observe that these workers are mostly *rule breakers* who provide poor responses after the first set of questions about the previous tasks. Just below 25% workers depict tipping points at R-3, while under 5% of workers have a tipping point at R-5, R-6, and R-7. This shows that a significant number of malicious workers (especially untrustworthy workers) exhibit early signs of malicious activity, while a smaller percentage depict signs of malicious activity at a later stage.

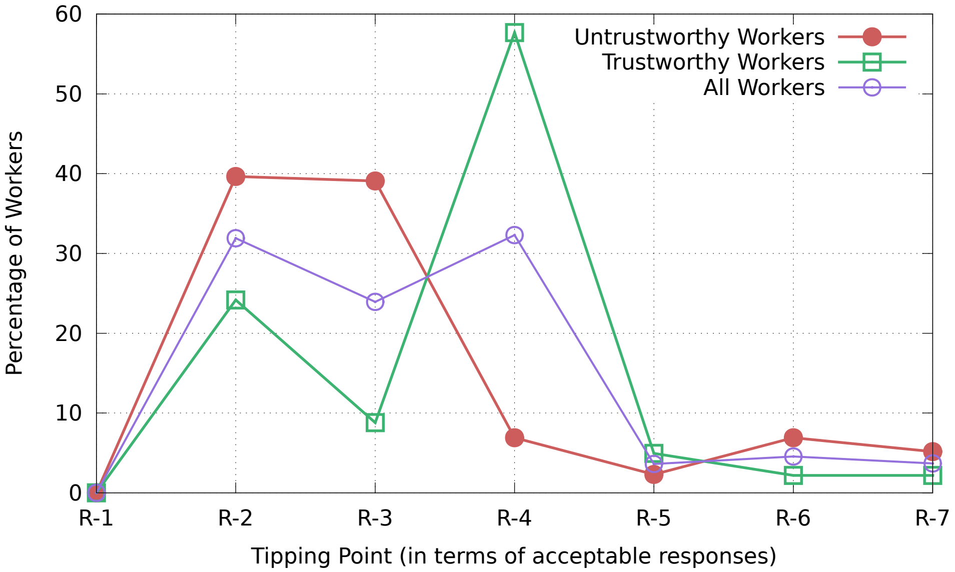


Figure 3.6 Distribution of Tipping Point of trustworthy and untrustworthy workers.

3.4.5 Worker Maliciousness vs Tipping Point

We investigate the relationship between the maliciousness *(M)* of untrustworthy workers *(UW)*, trustworthy workers *(TW)* and their corresponding tipping points. We hypothesize that a worker with a greater maliciousness would have an earlier tipping point. Based on the analysis, we present our findings in Table 3.1.

Table 3.1 Relationship between the Maliciousness and Tipping Point of untrustworthy and trustworthy workers (percentage of workers having tipping point @R).

Maliciousness	**UW**	**TW**
$0 < M \leq 0.2$	40.9% @ R-7 31.8% @ R-6	28.5% @ R-7 28.5% @ R-5
$0.2 < M \leq 0.4$	43.47% @ R-3 21.73% @ R-6	30% @ R-5 30% @ R-3
$0.4 < M \leq 0.6$	66.19% @ R-3 25.35% @ R-2	88% @ R-4 5.1% @ R-3
$0.6 < M \leq 0.8$	71.05% @ R-2 28.95% @ R-3	60% @ R-3 40% @ R-2
$0.8 < M \leq 1$	100% @ R-2	100% @ R-2

We find that a majority of untrustworthy workers (40.9%) and trustworthy workers (28.5%) having a $M \leq 0.2$, have a tipping point at R-7. In case of the untrustworthy workers having $0.2 > M \leq 0.4$, 43.47% of workers have a tipping point at R-3, while 21.73% have a tipping point at R-6. In all the cases where $M > 0.4$, a great majority of workers have a tipping point at either R-3 or R-2. We observe a clear trend, which implies that the greater the maliciousness of a worker, the earlier is the 'tip' towards unacceptability. From this we learn that, a worker who provides poor responses in the beginning should be dealt with stricter measures, since there is a greater probability that the worker is malicious.

3.4.6 Worker Behavior Beyond the Tipping Point

We analyzed the behavior of workers beyond their tipping points in order to verify whether the tipping point is a true indicator of further malicious activity from workers. Table 3.2 presents the amount of workers who depict malicious activity after their corresponding tipping points. We observe that over 95% of trustworthy and untrustworthy workers that have a tipping point at R-2, go on to provide at least one more unacceptable response.

Table 3.2 Percentage of workers that depict malicious activity after their corresponding Tipping Points.

Workers (in %)	R-2	R-3	R-4	R-5	R-6
Trustworthy	95.45	93.75	100	55.56	25
Untrustworthy	98.55	100	69.23	75	41.67

All trustworthy workers having a tipping point at R-4 and all untrustworthy workers having a tipping point at R-3, go on to provide at least one more unacceptable response. From Table 3.2, we learn that the tipping point is a good indicator of forthcoming malicious activity within the task.

3.4.7 Task Completion Time vs Worker Maliciousness

We also investigate the time that workers take in order to complete the task. In order to draw a comparison across the different types of behavior exhibited by workers, with respect to the time that they take for task completion, we use the average task completion time for each type of worker behavior. Apart from this, we also compare the maliciousness exhibited by each group of workers constituting the different types of behavior. We find that the *average task completion time* and the *average maliciousness* of untrustworthy workers show a high Pearson Correlation of 0.514.

Figure 3.7 presents our findings with respect to the analysis described here. We observe that *fast deceivers* exhibit the most amount of maliciousness on average. Interestingly, they also take the least amount of time to complete the task. This is coherent with the type of behavior they exhibit, which is providing bad responses and

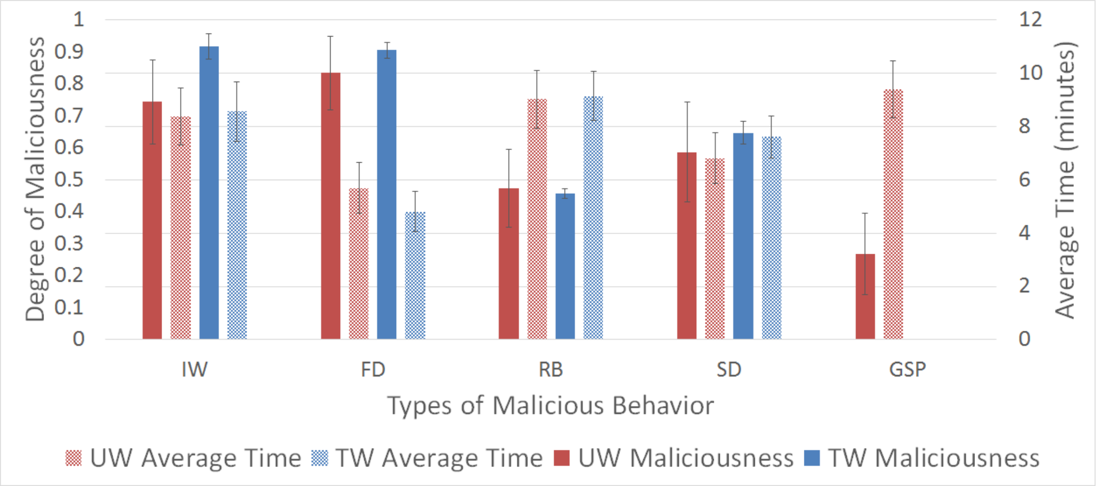


Figure 3.7 Comparison of Worker Maliciousness and Average Response Time of the different classes of malicious workers.

achieve quick task completion (usually by copy-pasting same responses for multiple questions, or entering gibberish responses). On the other hand, we observe that *smart deceivers* also exhibit high malicious content, but they take more time to complete the task. We reason that this is due to the fact that *smart deceivers* take more precautions in order to bypass possible validators. *Gold standard preys* depict the least amount of maliciousness amongst all the types of untrustworthy workers. They also depict the highest average task completion time, indicative of a lower maliciousness. The *rule breakers* depict a high average task completion time, and a moderate maliciousness. This is attributed to their behavioral pattern; wherein the workers do not provide responses that meet the priorly stated requirements. *Ineligible workers* who complete the task, also depict a high maliciousness.

3.4.8 Caveats and Validity Threats

It is important to note that in this work, when we refer to 'maliciousness', we infer this based on the responses provided by a worker. There is no way to learn about the real intentions of a worker behind each response, based merely on the response itself.

While studying the major challenges that stand in the way of efficient crowdsourcing paradigms, Kittur et al. say that workers who are new and have relatively low expertise, as well as task administrators who do not provide clear instructions contribute to poor responses [KNB+13]. In order to ensure that we did not introduce unwanted bias due to the inexperience of workers (that could result in spiking the number of malicious workers), we ensured that all the questions in the crowdsourced survey were straightforward and easy to answer, even if a worker has little experience. Moreover, clear and thorough instructions were provided in the survey to aid the workers in completing the task.

We acknowledge that trustworthy workers may provide poor responses due to fatigue or boredom, as discussed in previous works. However, by varying the format of questions (open-ended, multiple choice check-boxes, Likert-type), limiting questions of the same type (two sets of 3 open-ended questions about previous tasks), and engaging the crowd with humor evoking attention-check questions, we attempt to curtail such bias.

The degree of *acceptability* corresponding to a worker's response is a metric that can be used at the discretion of the task administrator. In our case, we have computed the *acceptability* of a worker by awarding scores of '1' or '0' to each response, depending on whether a response is acceptable or unacceptable respectively. However, if a clear distinction with respect to the extent of usefulness of a response can be made, then a task administrator can use a continuous value between the closed interval of [0,1] in order to represent the acceptability of a response.

Since we do not randomize the order in which questions are answered by workers, we do not venture into analyzing the relationship between the type of question and the *tipping point*. We aim to extend this work with such an analysis in future. By doing so, we can empirically propose ideal lengths of tasks featuring different types of questions. Finally, more trials on different platforms, using varying design types would be ideal to further reinforce our findings.

3.5 Discussion

Our experimental setup and findings are based on the task type, '*survey*'. A survey-type task inherently begets a general population of the crowd, without restricting participation due to the open design. Thus, the various kinds of trustworthy and untrustworthy workers presented in our work are representative of the general crowd. Having said that, the distribution of different kinds of untrustworthy workers depends on the type of task. This is due to the fact that a particular type of task may or may not be breached by some kinds of malicious workers, depending on the nature of the task and the gold standards being used.

Our experimental results showed that there was no significant correlation between the channels that the workers used for task completion and the behavioral patterns observed.

In our study of workers, we detect different types of worker behavior as described earlier. A key observation is that gold standard test questions alone remain insufficient to curtail malicious activity. We find that trustworthy workers who pass test questions can still provide ill-fitting responses (as in case of *rule breakers*), or deceive the task administrators (as in case of *smart deceivers*). By understanding the various kinds of behavior prevalent in the crowd, administrators can design tasks much more effectively. Being aware of the different ways in which malicious workers attempt to cheat their way towards task completion, can help in developing mechanisms to counter such activity. For instance, we find that tasks of the 'survey' type are most prone to activity of the kind exhibited by *fast deceivers* and *rule breakers*. This urges the need for stringent response validation especially for open-ended questions, to curtail possible attempts to cheat by *fast deceivers*, and *rule breakers* who provide sub-optimal responses.

The responses from *Gold Standard Preys* are valid and acceptable, though they are tripped by the gold standard questions, owing to a possible lack of attentiveness. By detecting such workers and consuming their responses instead of discarding them, task administrators can enhance the value of responses received from the pool of workers in a task. This essentially means that, by detecting gold standard preys the value of responses can be maximized without increasing the costs for task completion.

The measurement of *maliciousness (M)* of the workers, as presented earlier can be extended to different types of tasks, since the method relies on determining the acceptability of the individual responses from a worker in the context of the task. Depending on his needs, a task administrator can choose to discard responses from workers based on their maliciousness, thus using M as a sliding window for filtering responses.

3.5.1 Task Design Guidelines

We propose the following guidelines in order to design tasks of the '*survey*' type efficiently. By adhering to these key guidelines, we claim that the malicious activity prevalent in tasks of this type can be curtailed to a significant extent.

- The *tipping point* can be used to identify workers who 'tip early in the job. By excluding such workers, the quality of the produced results can be improved.
- In order to restrict the participation of *ineligible workers*, task administrators could employ a commonly used pre-screening method.
- Stringent validators should be used in order to ensure that workers cannot bypass open-ended questions by copy-pasting identical or irrelevant material as responses. This is an important guideline to enforce for survey-type tasks, since open-ended questions are popular in surveys and the majority of malicious workers are *fast deceivers*.
- *Rule breakers* can be curtailed by ensuring that basic response-validators are employed, so that workers cannot pass off inaccurate responses, or nearly fair responses. Even trustworthy workers tend to tip through the course of a task, providing poor or partially accurate responses. This demands for methods to monitor the progress of workers. Such validators can enforce workers to meet the exact requirements of the task and prevent ill-fitting responses.

- Additional methods and careful steps are required to prevent malicious activity by *smart deceivers*. Since such workers take care to avoid being flagged, they present the most difficulties in detecting and containing. Only a small number of workers make the additional effort to deceive task administrators in surveys. Yet, these workers can be restricted by using psychometric approaches such as repeating or rephrasing the same question(s) periodically and cross-checking whether the respondent provides the same response.

- Surveys garner a fair number of *gold standard preys*. Therefore, a post-processing step should be accommodated in order to identify such workers and consider their acceptable responses if needed.

3.6 Chapter Summary

The ubiquity of the Internet, allows to distribute *crowdsourcing* tasks that require human intelligence at an increasingly large scale. This field has been gaining rapid popularity, not least because of the data-intensive nature of emerging tasks, requiring validation, evaluation and annotation of large volumes of data. While certain tasks require human intelligence, humans can exhibit maliciousness that can disrupt accurate and efficient utilization of crowdsourcing platforms. In our work, we aim to understand this phenomenon.

We have studied the behavior of malicious workers in the crowd by showcasing the task type of *Surveys*. Based on our analysis, we have identified different types of malicious behavior (**RQ #1**), which go beyond existing works and are better-justified through our data. An understanding of these aspects helps us to efficiently design tasks that can counter malicious activity, thereby benefiting task administrators as well as ensuring adequate utilization of the crowdsourcing platforms (**RQ #2**). By conducting an extensive analysis, we have introduced the novel concepts of measuring potential 'maliciousness' of workers in order to quantify their behavioral traits, and 'tipping point' to further understand worker behavior (**RQ #3**). Our contributions also include a set of guidelines for requesters to efficiently design crowdsourced surveys by limiting malicious activity.

4

Behavioral Traces for Worker Modeling and Pre-selection

> *"A worker may be the hammer's master, but the hammer still prevails. A tool knows exactly how it is meant to be handled, while the user of the tool can only have an approximate idea."*
>
> — *Milan Kundera*

The suitability of crowdsourcing to solve a variety of problems has been investigated widely. Yet, there is still a lack of understanding about the distinct behavior and performance of workers within microtasks. In this chapter, we first introduce a fine-grained data-driven worker typology based on different dimensions and derived from behavioral traces of workers. Next, we propose and evaluate novel models of crowd worker behavior and show the benefits of behavior-based worker pre-selection using machine learning models. We also study the effect of task complexity on worker behavior. Finally, we evaluate our novel typology-based worker pre-selection method in image transcription and information finding tasks involving crowd workers completing 1,800 HITs. Our proposed method for worker pre-selection leads to a higher quality of results when compared to the standard practice of using qualification or pre-screening tests. For image transcription tasks our method resulted in an accuracy increase of nearly 7% over the baseline and of almost 10% in information finding tasks, without a significant difference in task completion time. Our findings have important implications for crowdsourcing systems where a worker's behavioral type is unknown prior to participation in a task. Finally, we reflect on leveraging the automatic detection of worker types to identify and aid those workers who require further training to improve their performance. By providing a powerful mechanism to detect worker types, we make a case for promoting fairness, trust and transparency.

4.1 Introduction

A primary challenge in microtask crowdsourcing is quality assurance [KNB+13]. Various aspects can effect the quality of data collected [IPW10], ranging from poor HIT (Human Intelligence Task) design to the presence of malicious activity [GKDD15b]. To improve crowdsourced data quality, early work has focused on aggregating multiple answers from different workers in the crowd by going beyond the simple majority vote [DDCM12a, SL13, VGK+14a]. Other works have focused on modeling worker skills and interests to assign available HITs to them [DDCM13, BBC+13]. Other proposed techniques include task design approaches such as the use of gamification [FSVKS15a] and collaboration [RZS15].

Rzeszotarski and Kittur [RK11], proposed to track worker activity to distinguish between *good* and *bad* workers according to their performance. Recently, Dang et al. built a framework called *mmm*Turkey, by leveraging this concept of tracking worker activity [DHL16]. Rzeszotarski et al. showed several benefits of their approach when compared to other quality control mechanisms due to aspects such as effort, skill and behavior that can be interpreted through a worker's activity, and eventually help in predicting the quality of work [RK11, RK12]. While it is certainly useful to predict good versus bad quality of work, we argue that further benefits can be revealed by understanding worker activity at a finer level of granularity. For example, the knowledge that even *good* workers perform and operate in different ways to accomplish tasks, leads to the question of whether such differences can have practical implications. With the rise in adoption of crowdsourcing solutions that leverage human input through microtask marketplaces, new requirements have emerged. Often it is not sufficient to predict the quality of work alone when there are additional constraints on costs (in terms of time and money). Moreover, a better understanding of how good workers differ in complex crowdsourcing tasks can lead to further benefits like improved HIT design or HIT assignment models.

In this chapter, we aim to understand and identify the different types of workers in the crowd by focusing on *worker behavior*. Our objective is to advance the current understanding of different types of workers present in a crowdsourcing platform and leverage this for worker pre-selection, given a task to be completed. To do so, we collect activity tracking data from workers completing 1,800 HITs with varying length, type, and difficulty. We refine the existing understanding of worker types and extend it to multi-dimensional definitions within a *worker typology*. We experimentally show that it is possible to automatically classify workers into granular classes based on supervised machine learning models that use behavioural traces of workers completing HITs. Leveraging such worker type classification, we can improve the quality of crowdsourced tasks by pre-selecting workers for a given task. Our pre-selection method based on worker types yields an improvement of up to 10% compared with standard worker pre-selection techniques.

The main contributions of this work are presented below:

- We propose a granular *worker typology* based on multiple dimensions (*behavior*, *motivation*, *performance*) and derived from the workers' low-level behavioral traces.
- We define several behavioral features used to model worker behavior in crowdsourcing platforms, independent of different task features (task type, length and difficulty).
- We reveal the value of typecasting workers beyond *good* and *bad*, by highlighting the benefits of distinguishing *good* workers according to the typology.
- We evaluate the use of automatic worker type classification for the problem of worker pre-selection, outperforming standard pre-selection methods based on qualification tests.
- Finally, we leverage our findings and reflect on promoting notions of fairness, trust and transparency in the marketplace.

4.2 Related Work

Modeling Crowd Workers. Work on understanding and modeling worker behavior includes [KKMF11], where authors propose worker types based on outcomes of behavior, such as time spent on the task and quality of the work. They define four classes of workers: diligent (workers completing the task carefully), competent (efficient and effective workers), incompetent (workers with a poor understanding of the task), and spammers (malicious workers).

In [KKMF13b], it was observed that varying task design properties (task difficulty and reward) has an impact on the type of crowd which completes the task and workers' performance based on their interest and perceived challenge. Bored workers underperformed and workers who found the task difficult obtained lower accuracy. Worker models have been built by indexing Facebook pages that workers liked, to assign HITs to those workers whose skills and interests best fit the task topic [DDCM13]. A worker taxonomy focusing on different types of poorly performing workers is described in [VDV12] where diligent workers are compared to sloppy workers (i.e., honest but providing low quality labels), random spammers with an inter-worker agreement rate close to random, and uniform spammers who repeat the same answer over the task. In [VDV12], methods to automatically detect such workers are based on the comparison with other worker answers. Recent work focused on the understanding of malicious behavior exhibited by workers in the crowd [GKDD15b] where authors observed different malicious techniques used by workers to complete HITs with the sole purpose of obtaining monetary rewards, without providing a quality response.

These prior works primarily focus on the outcomes of work completed to typecast workers. We advance the understanding of worker types by integrating the different dimensions considered in lone typologies in each case of previous work. The result is a holistic perspective of behavior, performance and motivation for each category in the proposed worker typology with a higher granularity of worker behavior.

Quality Control. Classic approaches to detect low quality work compare worker responses against a gold standard dataset [WIP11]. As discussed earlier, prior work looked at worker tracking data with the purpose of distinguishing between high and low performing workers [RK11]. Additionally, in [RK12] the authors present visual analytics tools that allow requesters to observe worker performance and identify low performers to be filtered out. In [KKMF12] authors look at worker demographics and personality traits as indicators of work quality. Our work is complementary to these prior works. By relying on a more granular understanding of worker types, we afford pre-selection of desired workers. We extract behavioral features and propose a supervised machine learning model, that automatically detects worker types, thus going beyond the good/bad binary classification problem.

Tracking User Activity on the Web. Tracking user activities on the Web is a common approach to study user engagement [LOYT14]. Web systems traditionally collect logs of activities at the lowest level of individual clicks. Mining user activity logs is a very popular technique applied to a variety of problems, such as online search engine result pages [HWD11].

In [GA08] authors leverage mouse cursor patterns to infer the query intent of users of a Web search engine. In [ALV14] authors look at mouse tracking to understand the consumption of news articles showing how cursor activity correlates with user experience. Similarly, we use worker activity signals to understand their behavior in crowdsourcing platforms.

In the context of crowdsourcing, there has been little use of user tracking techniques. Examples include [CTIB15] where authors logged browser window focus changes to understand interruptions. In [FLS$^+$15] authors used mouse tracking to generate heatmaps over HITs to see which part workers focused on. More recently, [KZ16] looked at behavioral data to compare experts and crowd workers on relevance judgments HITs.

In this chapter we go beyond this basic use of tracking data by identifying patterns and classifying workers into a predefined set of types, that allows us to pre-select the preferred type of workers for a given task.

4.3 Research Goals and Setup

We aim to address the following research questions:

RQ#1 How can requesters benefit from the knowledge of worker types at a fine granularity?

RQ#2 Can worker behavioral traces be used to classify workers automatically into distinct types?

RQ#3 What is the impact of *task complexity* and *task type* on the behavior of crowd workers?

RQ#4 How effective is behavior-based pre-selection of crowd workers?

4.3.1 Modeling Task Complexity

One can model *task complexity* [YRDB16] from a worker's point of view, where worker competence for example, could play a role in determining how complex a given task is. This is logically sound, since one worker can find a given task to be difficult while another can find the same task to be simple. However, including inherent worker traits in task complexity modeling would make it subjective. To define task complexity from a purely objective standpoint, we consider the characteristics of the task alone. Herein, we model task complexity as a function of (i) the objective difficulty-level of the task and (ii) the length of the task.

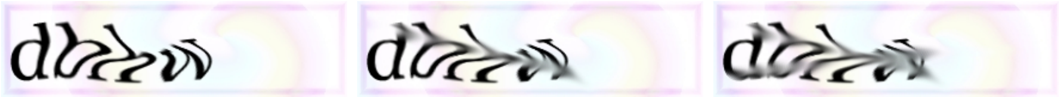
4.3.2 Methodology

To address the research questions stated earlier, we consider the task types of *Content Creation* and *Information Finding* [GKD14d]. A recent study on the dynamics of crowdsourced tasks on Amazon's Mechanical Turk (AMT) showed that content creation tasks have been the most popularly crowdsourced tasks over the last 5 years, while information finding tasks have depicted the most growth over the last 3 years [DCD$^+$15a].

Microtask Design - Content Creation

Due to its popularity, we choose to use *image transcription* as the content creation task in our experiments. For this purpose, we use a dataset of captchas[1] where crowd workers are asked to decipher characters from a given image [GFK15b]. To cater to varying task complexity and observe consequent behavior of workers participating in the tasks, we consider tasks with lengths of 20, 30, and 40 units respectively. In each unit a worker is asked to transcribe a captcha. Apart from this, to model the difficulty-level aspect of task complexity, we use the objective notion of smudging the captchas with `no-stroke`, `one-stroke`, and `two-strokes` to indicate a progressively increasing difficulty-level of tasks (as shown in Figure 4.1). Thus, we aim to replicate the objective reality of image transcription tasks where some images can be easier to transcribe than others.

[1] http://www.captcha.net/

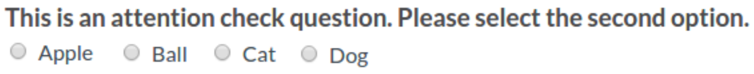


(a) Difficulty-Level I (`no-stroke`)

(b) Difficulty-Level II (`one-stroke`)

(c) Difficulty-Level III (`two-strokes`)

Figure 4.1 Progressive difficulty-levels in the content creation task of transcribing captcha images.



This is an attention check question. Please select the second option.

○ Apple ○ Ball ○ Cat ○ Dog

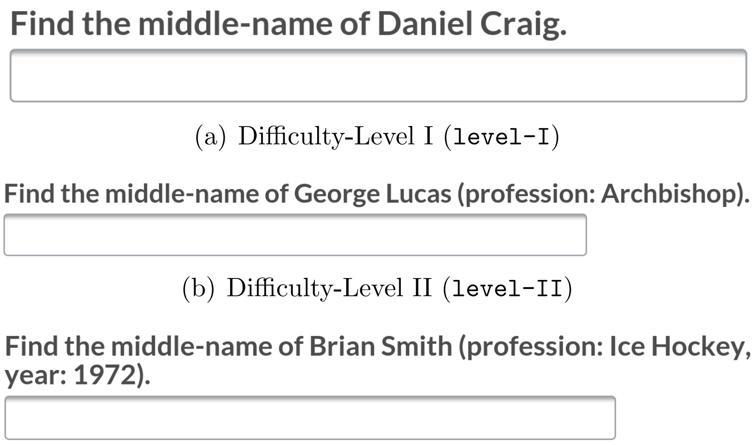
Figure 4.2 Question to assess *trustworthiness* of workers.

To deduce the *trustworthiness* of a worker as demonstrated in [GKDD15b], we intersperse multiple choice questions between the image transcription units at a regular interval of 25% of total units. We explicitly ask workers to pick a given option, (as shown in Figure 4.2), and due to the fact that this is a change in question format (from transcribing an image in a text field to answering a multiple choice question) we believe that it is not possible for a trustworthy worker to miss the direct and clear instruction. We consider workers who answer one or more of these questions incorrectly to be *untrustworthy*.

Microtask Design - Information Finding

For the information finding type, we adopt the task of finding the middle-names of famous people. To investigate the effect of varying task complexity on worker behavior, we consider tasks with length of 10, 20, and 30 units (since this type of task requires more time for completion in comparison to the content creation task of image transcription). In each unit, a worker is asked to find the middle-name of a given person. We model the task difficulty objectively into 3 levels, wherein workers need to consider an additional aspect in each progressively difficult level as shown in Figure 4.3.

In `level-I`, workers are presented with unique names of famous persons, such that the middle-names can be found using a simple Google or Wikipedia search. In `level-II` workers are additionally provided with the profession of the given person. We manually selected the names such that there are at least two different individuals with the given names in `level-II`, and the distinguishing factor that the workers need to rely upon is their profession. In `level-III` workers are presented names of persons, their profession, and a year during which the persons were active in the given profession. There are multiple distinct individuals with the given names, associated with the same profession in `level-III`. The workers are required to identify the accurate middle-name by relying on the year in which the person was active in



Find the middle-name of Daniel Craig.

(a) Difficulty-Level I (`level-I`)

Find the middle-name of George Lucas (profession: Archbishop).

(b) Difficulty-Level II (`level-II`)

Find the middle-name of Brian Smith (profession: Ice Hockey, year: 1972).

(c) Difficulty-Level III (`level-III`)

Figure 4.3 Progressive difficulty-levels in the Information Finding task of finding the middle-names of famous persons.

the profession. We use the same method adopted in the content creation tasks to determine the *trustworthiness* of workers in these information finding tasks.

4.3.3 Experimental Setup

We deployed 9 tasks of the content creation type, with varying combinations of length (20, 30, 40 units) and objective difficulty-levels (no-stroke, one-stroke, two-strokes) on CrowdFlower[2]. Similarly we deployed a further 9 tasks of the information finding type, with varying combinations of length (10, 20, 30 units) and objective difficulty (level-I, level-II, level-III). For each of these 18 tasks, we gathered responses from 100 distinct workers resulting in a total of 1,800 HITs. We deployed tasks of the same type and difficulty-level concurrently, in order to avoid potential learning biases. Workers were paid in accordance to the task complexity of a given task (10, 20, 30 USD cents per unit).

Mousetracking Implementation: We implemented mousetracking using Javascript and the JQuery library, and logged user activity data ranging from mouse movements to keypresses. We took measures to distinguish between workers that use a mouse and those who use a touchpad. We also distinguish between worker mannerisms with respect to scrolling behavior; use of scrollbar as opposed to the mousewheel. In this way, we gathered worker activity data from each of the experimental tasks deployed on CrowdFlower. Apart from this data, we use a Javascript implementation[3] of *browser fingerprinting* [Eck10] in order to identify workers that participate in tasks multiple times (*'repeaters'*) by virtue of using different `worker-ids`. We take mea-

[2] http://www.crowdflower.com/

[3] http://github.com/Valve/fingerprintjs

sures to avoid privacy intrusion of workers by hashing various browser characteristics such as the user agent, cookies settings, screen resolution, and so forth, results in a 64-bit browser fingerprint. We do not retain any worker-specific browser traits other than the resulting fingerprint to identify repeaters.

4.4 Categorization of Workers

4.4.1 Modeling Worker Behavior

Crowd worker behavior is influenced by several aspects, some of which are inherent to the worker (such as trustworthiness of a worker) and others that are induced by the nature of tasks (such as *task complexity* [YRDB16]). Workers can be categorized based on the quality of their work. Some categories proposed by [GKDD15b] and [KKMF11] include elite workers, competent workers, less-competent workers, and so forth. As described by [EHdVS12], *money-driven* workers are motivated by the monetary incentives, while *entertainment-driven* workers mainly seek diversion but readily accept the monetary rewards as additional extrinsic motivation.

We present a worker typology by building on these prior works in an inductive and data-driven fashion prescribed by [BLL04]. We explicitly gathered information from workers regarding their motivation for participation. We combine behavior, motivation, and performance rather than looking at each aspect individually to type-cast workers. We manually inspected workers' responses to the 1,800 HITs and built rubrics around their task completion time, trustworthiness, and performance to assign appropriate labels. The rubrics were such that worker types could be assigned without clashes between the classes. Three authors of this chapter acted as experts and designed a coding frame according to which we could decide which category in the typology a worker belonged to. In case, the characteristics exhibited by workers did not fit any existing category, a new one was created. After resolving disagreements on the coding frame every worker was labeled with a category. We followed the guidelines suggested by [Str87, BLL04] while conducting the open-coding of behavioral data, collected over the 1,800 HITs run on CrowdFlower, consequently leading to the following categories. Note that worker types describe session-level behavior of the workers rather than properties of a person.

– **Diligent Workers (DW).** These crowd workers may be *money-driven* or *entertainment-driven*. They make sure to provide high quality responses and spend a long time to ensure good responses.

– **Competent Workers (CW).** These crowd workers may be *money-driven* or *entertainment-driven*. They possess skills necessary to complete tasks in a quick and effective manner, producing high quality responses.

– **Fast Deceivers (FD).** These crowd workers are *money-driven*, and attempt to complete a given task in the fastest possible way to attain the rewards offered. Due to this, *fast deceivers* provide poor responses by copy-pasting content and taking advantage of loopholes in the task design (such as weak or missing validators).

– **Smart Deceivers (SD).** These crowd workers are *money-driven* and aware of potential validators and checks that task requesters may be using to flag workers (such as minimum time spent on a question). They provide poor responses without violating validators, and thereby exert less effort to attain the incentives.

– **Rule Breakers (RB).** These crowd workers may be *money-driven* or *entertainment-driven*. They provide mediocre responses that fall short of the expectations of a requester (eg., providing 3 keywords where 5 are required).

– **Less-competent Wokers (LW).** These crowd workers may be *money-driven* or *entertainment-driven*. They appear to have a genuine intent to complete a given task successfully by spending ample time on it, but lack the necessary skills to provide high quality responses.

– **Sloppy Workers (SW).** These crowd workers may be *money-driven* or *entertainment-driven*. They complete tasks quickly and perform with an average or below average accuracy. Sloppy workers [KKMF11] appear to err due to their speed within the task.

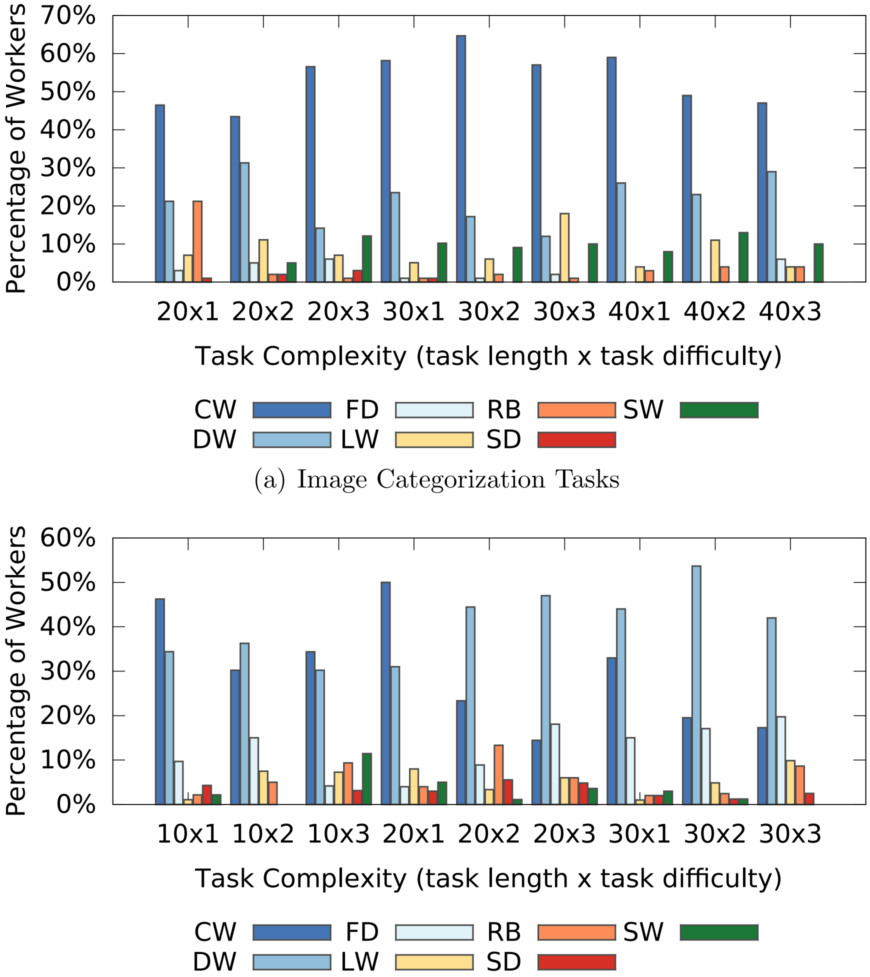
Based on the responses of individual workers in each of the 18 different tasks, 3 authors of this work acted as experts and manually categorized workers into different classes of the worker typology presented earlier. In the 9 content creation tasks of image transcription, the overall inter-rater agreement on the expert annotations was found to be 80.1% according to percent agreement, while that in case of the 9 information finding tasks of finding middle-names was found to be 89.1%. Following a phase of discussion between the experts, the instances with disagreements were resolved in order to ensure accurate categorization of workers.

4.4.2 Worker Types in CC and IF Tasks

Figure 4.4(a) presents the distribution of different worker types based on manual annotations in the *content creation* (CC) tasks of image transcription with varying task complexity.

We note that in tasks with the length of 20 units, the percentage of sloppy workers (SW) and fast deceivers (FD) increases with an increase in task difficulty, while that of rule breakers decreases. In the tasks with length 30 units, we observe an increase in the number of less-competent workers (LW) with an increase in difficulty level. This indicates that as the complexity of a task increases, the competence or skill of a worker plays a more decisive role in the worker's performance. In tasks with a length of 30 and 40 units we note a high fraction of sloppy workers (SW) on average.

Figure 4.5 presents the average accuracy of different types of workers and their task completion time in each of the CC tasks with varying task complexity. Across all the tasks, we observe that competent workers (CW) and diligent workers (DW)



(a) Image Categorization Tasks

(b) Information Finding Tasks

Figure 4.4 Distribution of worker types in the (a) content creation (CC) tasks and (b) information finding (IF) tasks, with varying task complexity. The different worker types presented here are as follows. `CW`: Competent Workers, `DW`: Diligent Workers, `FD`: Fast Deceivers, `LW`: Less-competent Workers, `RB`: Rule Breakers, `SD`: Smart Deceivers, `SW`: Sloppy Workers.

exhibit the highest levels of accuracy. However, CW take significantly lesser time than DW to complete the tasks (*p<.001*). We also note that with increasing task complexity, DW take more time to complete. As shown in Tab. 4.1, less-competent workers (LW) also take a long time for task completion, but exhibit a much lower accuracy. Fast deceivers (FD) and smart deceivers (SD) exhibit lowest accuracies and task completion times on average across all tasks, indicating their reward-focused intentions. Rule breakers (RB) perform with a low accuracy across all the CC tasks, indicative of their behavior resulting in partial responses.

Table 4.1 Overall average accuracy and task completion time of different types of workers in CC tasks.

Worker Type	Avg. Acc (in %)	Avg. Time (in mins)
CW	83.79	5.01
DW	82.94	11.37
FD	7.81	1.89
LW	62.30	10.34
RB	30.23	4.49
SD	21.57	3.94
SW	59.14	4.48

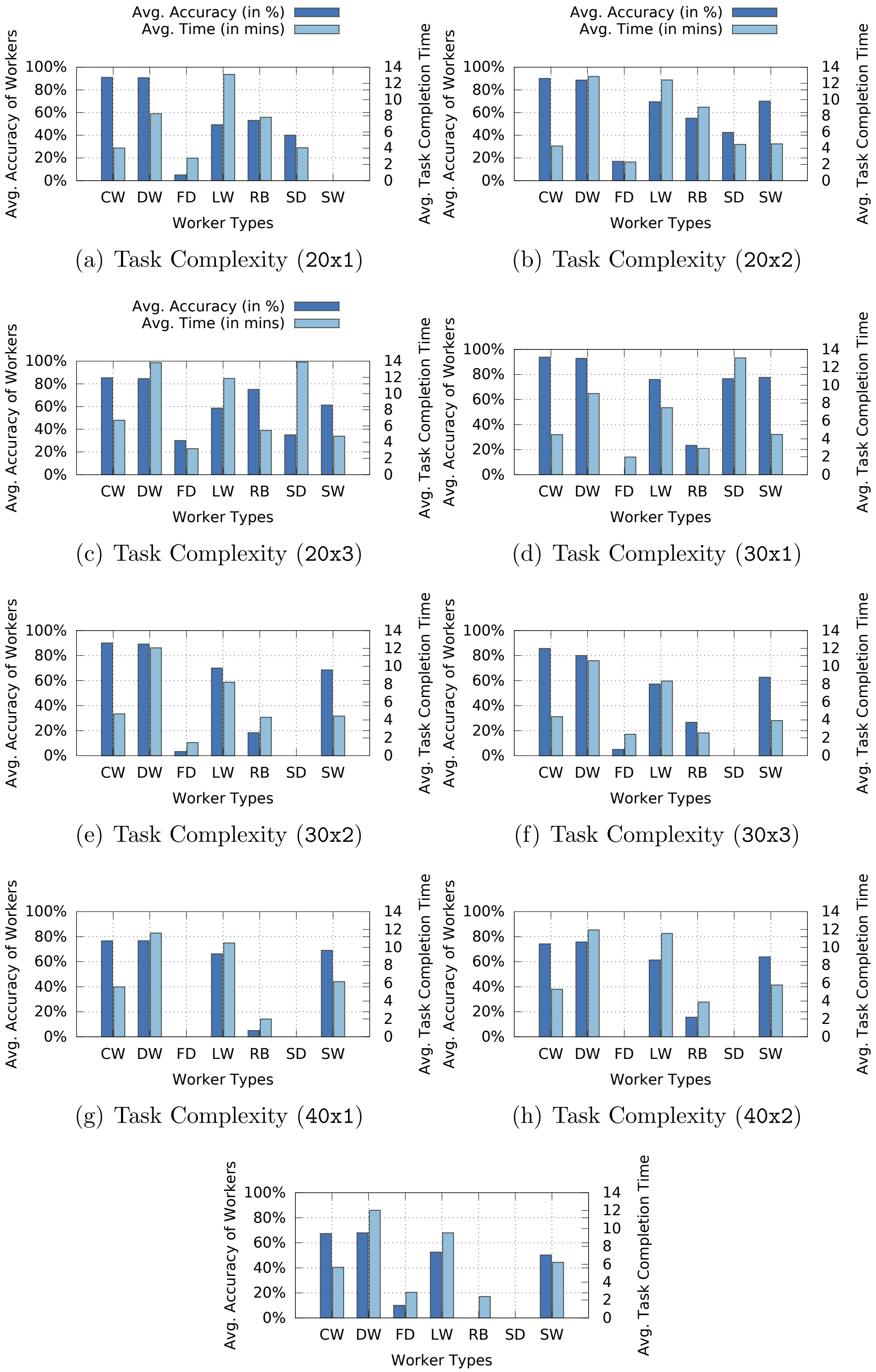
Figure 4.4(b) presents the distribution of different worker types based on manual annotations in the *information finding* (IF) tasks of finding middle-names with varying task complexity. We can see that with an increasing task complexity there is a decrease in the number of CW and increase in the number of DW. This indicates that complex tasks can go beyond the competence of workers and therefore workers tend to require more time to complete the task in order to perform accurately. We also note an increase in the number of FD with increasing task complexity.

Figure 4.6 presents the average accuracy and task completion time of different types of workers in the IF tasks with varying task complexity. Once again we notice that CW and DW exhibit the highest accuracies across the different tasks, with CW taking significantly lesser time to complete the tasks (*p<.001*). Table 4.2 presents the overall accuracy of different types of workers and their corresponding task completion times in the IF tasks. We observe that on average DW fractionally outperform CW (but this difference is not statistically significant). FD and SD exhibit the lowest accuracies and task completion times due to their behavior. LW spend a considerable amount of time on the tasks but fail to attain a high level of accuracy.

4.5 Automatic Categorization

To use the proposed worker typology in practice, we trained supervised machine learning models to automatically classify worker types based on behavioral traits.

Features Indicating Behavioral Traces



(a) Task Complexity (20x1)

(b) Task Complexity (20x2)

(c) Task Complexity (20x3)

(d) Task Complexity (30x1)

(e) Task Complexity (30x2)

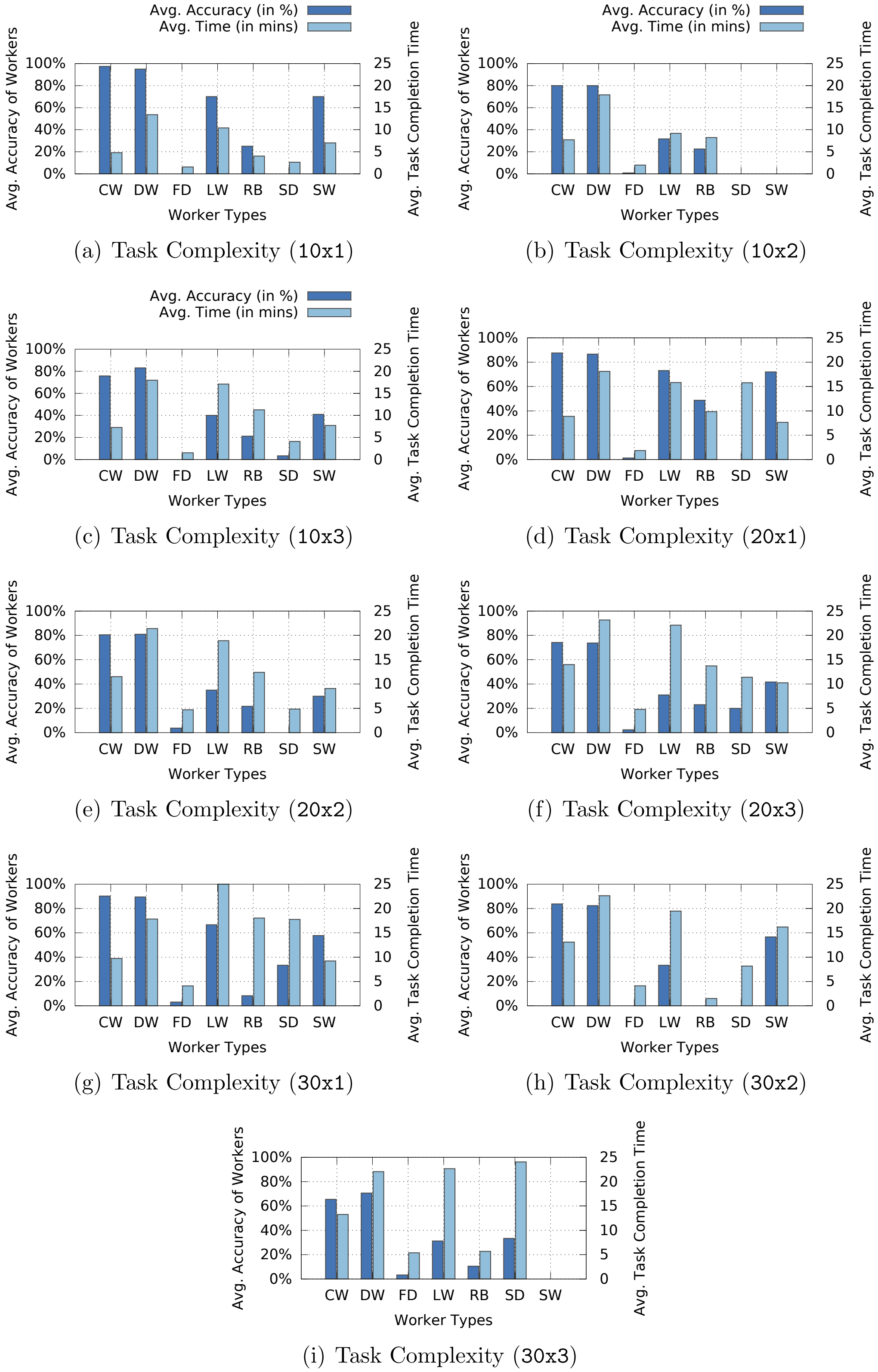
(f) Task Complexity (30x3)

(g) Task Complexity (40x1)

(h) Task Complexity (40x2)

(i) Task Complexity (40x3)

Figure 4.5 Average accuracy (scaled on the y-axis) and task completion time (scaled on the y2-axis) of different types of workers in **image transcription tasks** with varying task complexity.



(a) Task Complexity (10x1)

(b) Task Complexity (10x2)

(c) Task Complexity (10x3)

(d) Task Complexity (20x1)

(e) Task Complexity (20x2)

(f) Task Complexity (20x3)

(g) Task Complexity (30x1)

(h) Task Complexity (30x2)

(i) Task Complexity (30x3)

Figure 4.6 Average accuracy (scaled on the y-axis) and task completion time (scaled on the y2-axis) of different types of workers in **information finding tasks** with varying task complexity.

Table 4.2 Overall average accuracy and task completion time of different types of workers in IF tasks.

Worker Type	Avg. Acc (in %)	Avg. Time (in mins)
CW	73.62	18.39
DW	75.51	26.79
FD	1.75	3.03
LW	43.28	20.27
RB	18.51	11.57
SD	10	9.57
SW	41	6.90

We study the mousetracking data (including keypresses) generated by crowd workers in 1,800 HITs through the 9 content creation and 9 information finding tasks, in order to determine features that can help in the prediction of a worker type. Some of the important features are presented below. A complete list of features used can be found here[4].

- `time`: The task completion time of a worker.
- `tBeforeLClick`: The time taken by a crowd worker before responding to the multiple choice demographic questions in the tasks.
- `tBeforeInput`: The time taken by a crowd worker before entering a transcription in the content creation task or a middle-name in the information finding task.
- `tabSwitchFreq`: No. of times that a worker switches the tab while working on a particular task.
- `windowToggleFreq`: No. of times that a worker toggles between the current and last-viewed window while working on a particular task.
- `openNewTabFreq`: No. of times that a worker opens a new tab while working on a particular task.
- `closeCurrentTabFreq`: No. of times that a worker closes the current tab while working on a task.
- `windowFocusBlurFreq`: No. of times that the window related to the task goes in and out of focus until task completion by the crowd worker.
- `scrollUp/DownFreq`: No. of times that a worker scrolls up or down while working in a task respectively.
- `transitionBetweenUnits`: No. of times a worker moves the cursor from one unit to another in the task.
- `totalMouseMoves`: The total no. of times that a worker moves the cursor within the task.

By exploiting the expert annotated HITs and the features defined based on worker behavioural traces described earlier, we train and test a random forest classifier to predict worker types at the end of a completed task. We distinguish models for tasks with and without 'gold questions' (i.e., questions with known answers used to check

[4] Shortened URL - https://goo.gl/jjv0gp

for work quality). We study the effectiveness of our supervised models to predict worker type in CC and IF tasks with varying task complexity. Tables 4.3 and 4.4 present Accuracy and F-Measure (to account for unbalanced classes) of our supervised worker type classifiers evaluated using 10-fold cross validation over IF and CC tasks.

Table 4.3 Supervised worker type classification evaluation for IF tasks with varying task complexity.

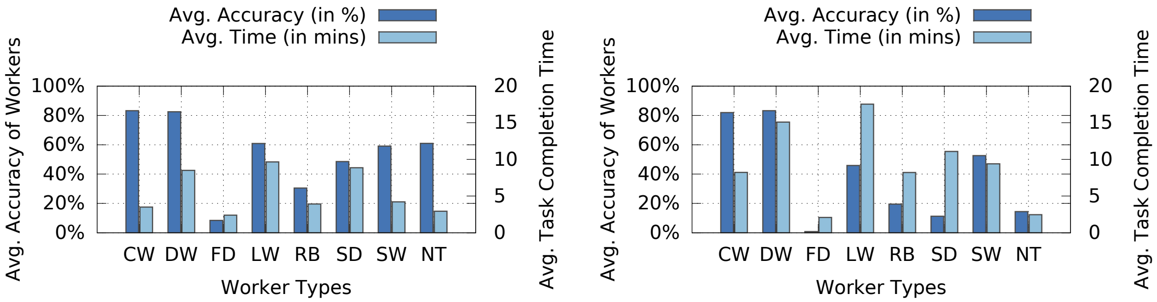
	with Gold Questions		w/out Gold Questions	
HIT Length	**Accuracy**	**F-Measure**	**Accuracy**	**F-Measure**
10	77.3	0.748	73.6	0.679
20	74	0.701	74	0.691
30	**81.4**	**0.786**	**79.8**	**0.763**
HIT Difficulty	**Accuracy**	**F-Measure**	**Accuracy**	**F-Measure**
Level-I	**82.3**	**0.779**	**80.5**	**0.754**
Level-II	79.4	0.77	74.6	0.718
Level-III	72.3	0.691	64.2	0.587

Table 4.4 Supervised worker type classification evaluation for CC tasks with varying task complexity.

	with Gold Questions		w/out Gold Questions	
HIT Length	**Accuracy**	**F-Measure**	**Accuracy**	**F-Measure**
20	69.02	0.671	58.6	0.532
30	**84.5**	**0.828**	75.6	0.712
40	80.3	0.768	**78.7**	**0.729**
HIT Difficulty	**Accuracy**	**F-Measure**	**Accuracy**	**F-Measure**
Level-I	74.7	0.714	**70**	**0.643**
Level-II	**77.5**	**0.746**	67.4	0.611
Level-III	72.5	0.696	64.5	0.59

We can observe that it is easier to predict worker types when gold questions are available in the task. We also observe higher accuracy of automatic worker type classification for IF in comparison to CC tasks. Moreover, as *longer* tasks typically provide more behavioral signals, they lead to better automatic classification of workers in our typology. A similar conclusion can be drawn for *less difficult* tasks where worker types can be better distinguished. Due to the imbalance in the different worker types, we also ran undersampling and oversampling experiments, that yielded similar results.

Additional results from the supervised classification evaluation showed that the easiest worker types to be predicted are CW (91% accuracy) and DW (87% accuracy) for CC tasks and DW (88.7% accuracy) and FD (86.6% accuracy) for IF tasks. Most confused worker types by our models are SW classified as CW for CC tasks and CW classified as DW for IF tasks. Feature selection by Information Gain shows



(a) Image Transcription Tasks

(b) Information Finding Tasks

Figure 4.7 Average accuracy (scaled on the y-axis) and task completion time (scaled on the y2-axis) of the first 5 judgments received from different worker types in the (a) image transcription and (b) information finding tasks. `CW`: Competent Workers, `DW`: Diligent Workers, `FD`: Fast Deceivers, `LW`: Less-competent Workers, `RB`: Rule Breakers, `SD`: Smart Deceivers, `SW`: Sloppy Workers, `NT` (No Type): First 5 judgments without considering worker type.

that the most predictive features to automatically predict the worker type are mouse movement, windows focus frequency, the task completion time, the score, and tipping point[5] computed from gold questions (when available).

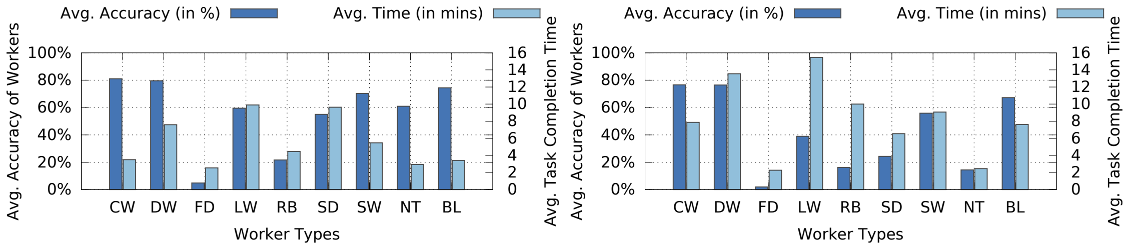
4.6 Evaluation and Implications

4.6.1 Benefits of Worker Type Information

In this section we investigate the potential benefit of automatically classifying workers as per the granular typology introduced in this chapter. We analyze the average accuracy of the first 5 workers of each type who submit their responses (where the worker type is considered according to the expert annotations). In typical crowdsourcing tasks where redundancy is required, 5 judgments has been considered the norm [VDV12]. By comparing this to the classic setting where worker type is unknown (`No Type`), i.e., the first 5 responses overall without considering worker types, we can measure the weight of worker type information.

Figure 4.7 depicts the benefit of having prior knowledge of worker types. We see that in both image transcription and information finding tasks `CW` and `DW` outperform the `No Type` setting. Moreover, in case of `CW` a high level of accuracy is observed with a fairly low task completion time. This makes the competent workers (CW) preferable when compared to diligent workers (DW) who tend to take more time. We also note that the average performance of `CW` (*M=83.24, SD=8.08*) across all

[5] First point at which a worker provides an incorrect response after having provided at least one correct response [GKDD15b].



(a) Image Transcription Tasks

(b) Information Finding Tasks

Figure 4.8 Average accuracy (scaled on the y-axis) and task completion time (scaled on the y2-axis) of the first 5 judgments received from different automatically predicted worker types in the (a) image transcription and (b) information finding tasks. The different worker types presented here are as follows. `CW`: Competent Workers, `DW`: Diligent Workers, `FD`: Fast Deceivers, `LW`: Less-competent Workers, `RB`: Rule Breakers, `SD`: Smart Deceivers, `SW`: Sloppy Workers, `NT` (No Type): First 5 judgments without considering worker type, `BL` (Baseline): First 5 judgments from workers who passed the standard pre-selection test.

image transcription tasks (Fig. 4.7(a)) is significantly better than `No Type` (*M=60.9, SD=17.62*) with *t(8)=4.5, p<.001*. Note that the other worker types apart from `CW` and `DW` can be considered detrimental, and automatically detecting these workers is an effective way to separate them from the worker pool.

Similarly, in case of the information finding tasks (Fig. 4.7(b)), we note that the average performance of `CW` (*M=81.96, SD=8.33*) is significantly better than `No Type` (*M=14.7, SD=22.1*) with *t(8)=5.04, p<.001*. We allude the poor performance of `No Type` in case of the information finding tasks to the inherent task complexity of the tasks. Since these tasks require relatively more time for completion the first responses tend to be submitted by workers who complete tasks very quickly (and with low accuracy, for e.g., `FD` or `SW`). In a typical crowdsourced task, requesters finalize units when a certain number of judgments are received. Thus, we observe an adverse effect on the quality of responses in the absence of pre-selection.

4.6.2 Results: Automatic Worker Classification

Here, we assess the impact of worker type predictions made by the proposed ML models described earlier. Once again we consider the first 5 judgments submitted by workers of each type (worker type as predicted by the classifer). We compare our proposed worker type based pre-selection method with the standard approach of using qualification tests which we refer to as the `Baseline`. In the `Baseline` method, we consider the first 5 responses from each worker to be a part of the qualification test. Only workers who achieve an accuracy of $\geq 3/5$ in the qualification test are considered to have passed the test. This follows our aim to replicate a realistic pre-screening scenario[6]. To compare the `Baseline` method with our proposed approach of worker type based pre-selection, we consider the first 5 judgments submitted by workers who passed the qualification test.

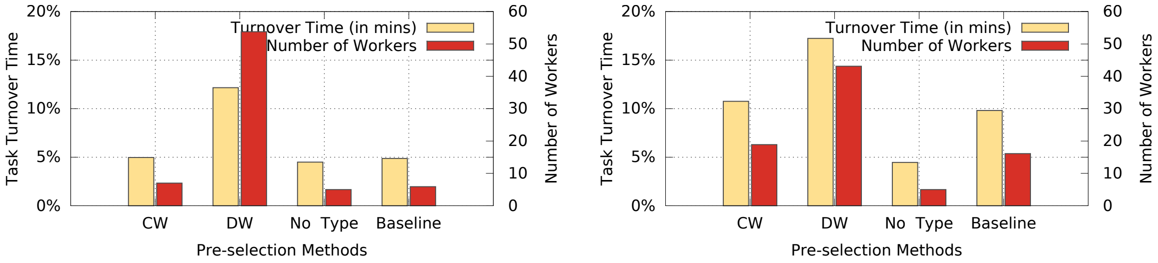
Figure 4.8 presents the results of our evaluation for the two task types. In case of the image transcription tasks (Fig. 4.8(a)) we note that on average across all tasks, `CW` (*M=81.03, SD=8.52*) significantly outperform workers in the `No Type` setting (*M=60.9, SD=18.69*) with *t(8)=5.04, p<.0005*. Interestingly, the task completion time (in mins) of `CW` (*M=3.5, SD=0.85*) is slightly more than that of `No Type` (*M=2.93, SD=0.48*) with *t(8)=1.86, p<.05*. `CW` also perform significantly better than the `Baseline` method (*M=74.41, SD=14.06*) with *t(8)=1.86, p<.05*. The differences in task completion time between `CW` and the `Baseline` method were not statistically significant, indicating that worker type based pre-selection of `CW` can outperform existing pre-selection methods in terms of quality without a negative impact on the task completion time.

For the information finding tasks (Fig. 4.8(b)), we note that on average across all tasks `CW` (*M=76.59, SD=11.34*) significantly outperform workers in the `No Type` setting (*M=14.44, SD=23.6*) with *t(8)=5.04, p<.0005*. In addition, we also observe that `CW` significantly outperform workers that are pre-selected using the `Baseline` method (*M=67.26, SD=14.92*) with *t(8)=1.86, p<.05*. The task completion time (in mins) of `CW` (*M=7.87, SD=3.56*) is not significantly different from that of the `Baseline` method (*M=7.62, SD=3.45*).

4.6.3 Task Turnover Time

The amount of time required to acquire the full set of judgments from crowd workers, thereby completing and finalizing a task considering pre-defined criteria (such as qualification tests or pre-selection) is called the *task turnover time*. We additionally evaluated the task turnover time of the different image transcription and information finding tasks when using the proposed typology-based worker pre-selection in comparison to the `Baseline` and `No Type` methods. Figure 4.9 presents our findings on average across all the tasks of the (a) image transcription tasks and (b) information finding tasks. We note that the average turnover time of the image transcription tasks where `CW` are pre-selected (*M=4.97, SD=1.47*) is negligibly longer than in case of the baseline method (*M=4.98 , SD=1.51*), with no statistically significant difference. These observations also hold for the information finding tasks where we did not find a significant difference between the turnover times corresponding to using the `CW` (*M=10.76 , SD=4.96*) and `Baseline` (*M=9.8, SD=4.17*) methods. Although we see that the `No Type` method results in a significantly lower turnover time when compared to `CW` with *p<.05*, as described earlier the accuracy of results when type information is not considered for pre-selection is relatively much lower. We note that the number of workers that were required before the task turnover was not signifi-

[6] CrowdFlower suggests a min. accuracy of 70% by default.



(a) Image Transcription Tasks

(b) Information Finding Tasks

Figure 4.9 Task turnover time and the number of workers required for task turnover on average across all (a) image transcription and (b) information finding tasks. `CW`: Competent Workers, `DW`: Diligent Workers, `No Type`: First 5 judgments without considering worker type, `Baseline`: First 5 judgments from workers who passed the standard pre-selection test.

cantly different between `CW` and `Baseline` methods across the different tasks in our experiments.

We also present the turnover times and the number of workers required for task turnover when `DW` are pre-selected for the sake of comparison. `DW` pre-selection results in significantly higher turnover times and requires more workers for task turnover (*p<.001*). In tasks without time constraints, requesters can consider pre-selecting `DW` in addition to `CW` due to their high result accuracy.

4.6.4 Implications on Fairness, Trust and Transparency

Prior work has discussed that requesters should consider the context in which workers are embedded while contributing work in online labour markets [MHOG14, GG16]. The work environments may not always be appropriate, and the devices that workers use to complete tasks may not be ergonomically suitable. Recent work has brought to light the influence of task clarity on the quality of work that is produced [GYB17]. Supporting such previous works that reflect on the wide landscape of quality in crowdsourced microtasks, our results show clear benefits in automatically typecasting workers in the pre-selection phase. However, employing such mechanisms should not alienate or discriminate against less-competent workers (LW). On the contrary, such workers should be supported in a manner that allows them to learn and transform into more effective and capable contributors [DKKH12]. Power asymmetry between workers and requesters in crowdsourcing marketplaces has been acknowledged as an issue, and addressed by recent works [IS13, GMG^{+}16]. Thus, it is important to consider other factors that promote fairness and transparency in the marketplace. Aiding, helping and training workers to learn and improve their performance in microtasks [GFK15b, GD17b] can have a positive impact on the mutual trust between workers and task requesters. Our results suggest that there is a need to support workers so that they become more effective and efficient (especially those who complete tasks while exerting genuine effort, such as the less-competent workers). One way to achieve this is to provide constructive feedback to workers who do not pass the pre-selection phase.

We can also rely on the low-level behavioral data of workers to mine behavioral patterns that lead to good performance. This can then be effectively communicated to workers who do not pass the pre-screening phase, with an aim to help them improve. For instance, based on the mousetracking data of crowd workers within the 18 image transcription and information finding tasks, we observed a variety of behavioral patterns as described below.

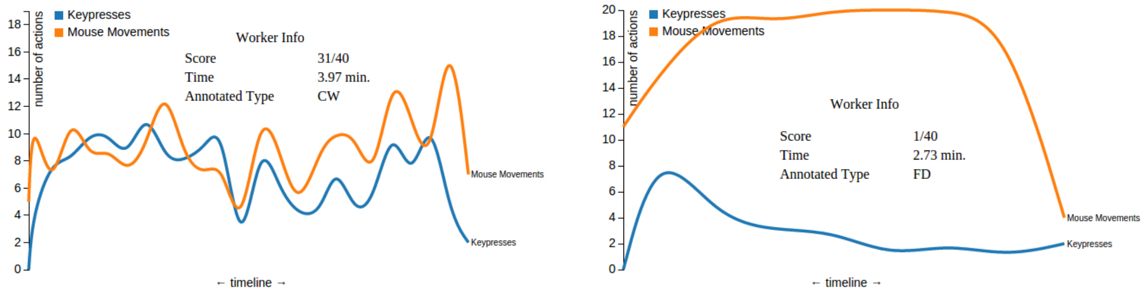
– **Multitaskers**: These workers participate in multiple tasks simultaneously to maximize their earnings. We deduced this behavior through keypresses that some workers used indicating opening and closing new tabs, switching between tabs and window toggles. This holds for the image transcription tasks since workers do not need to use multiple tabs in order to provide responses. However, in the information finding tasks workers are expected to search for the required information and therefore such keypresses do not necessarily indicate potential multitasking behavior.

– **Divers and Feelers**: Some workers read the instructions and start working on the task by providing responses immediately. Such workers are called *divers*. Others tend to scroll through the task before beginning to provide responses. These workers are called *feelers*.

– **Wanderers**: Some workers move back and forth through the task either to check their previous responses or read the instructions again. These workers appear to wander around in the task window and are called *wanderers*.

– **Copy-Pasters and Typers**: In the IF tasks of finding middle-names of people, we found that some workers copy-pasted the middle-names (*copy-pasters*) while others typed them into the response text fields (*typers*). We also observed the copy-pasting behavior in the image transcription tasks, wherein fast deceivers in some cases copy-pasted the same response for all units in the task.

To serve as illustrative example, Figure 4.10 presents screen captures of a comparative visualization between the mouse movements and keypresses of a competent worker (CW) and a fast deceiver (FD) in the image transcription task (40x3). We note that crowd workers exhibit different combinations of these behavioral patterns, and such behavior can lead to a worker belonging to any of the classes in the worker typology described earlier. Presenting workers who fail to pass the pre-screening phase with feedback based on their behavioral patterns, as well as the reason why they failed to pass the screening can help them reflect and improve their performance [Tar02]. In this way we can promote the following – (i) fairness with respect to how all genuine workers are treated in terms of opportunity, (ii) trust between task requesters and workers (since genuine workers who are less-competent can be assured that they will be helped to improve), and (iii) transparency in how workers can pass



(a) A competent worker (CW) in the image transcription task 40x3.

(b) A fast deceiver (FD) in the image transcription task 40x3.

Figure 4.10 Example screen captures of visualized behavioral patterns (*mousemovements* and *keypresses*) of a competent worker (CW) in comparison to a fast deceiver (FD) in the image transcription task with task complexity `40x3`.

the pre-screening phase and qualify for further work.

4.7 Discussion and Caveats

Over the last 3 years there has been a surge in the number of new task requesters on the Amazon MTurk platform (over 1,000 new requesters per month) [DCD+15a]. Tasks designed by less experienced requesters can be easy targets for fast deceivers (FD) and smart deceivers (SD) alike. In this chapter, we have shown that FD and SD take the least amount of time to provide responses despite of task complexity. There are two adverse effects of this behavior; (i) FD and SD can access a lot of available work that is susceptible to their behavior in the marketplace due to their quick task completion times. (ii) Due to the fact that responses provided by FD and SD cannot be easily distinguished from genuine workers on the fly, requesters accept the validity of their responses, thereby depriving other more suitable workers from participating in the task. Requesters thus face the dual-curse of getting sub-optimal returns for their investment in terms of response quality, and would require to deploy the tasks once again on discovering poor quality through a post-hoc analysis. In this context, automated pre-selection of workers based on their behavior, as proposed in this chapter can help requesters in improving their costs-benefits ratio while assuring the reliability and speed of produced results.

We also investigated the effect of task complexity on worker behavior. From our experiments, we found that with increasing task complexity the fraction of underperforming workers increases. In complex tasks it is therefore all the more important to pre-select workers who are capable of performing accurately as exhibited by competent and diligent workers (`CW`, `DW`).

The importance of distinguishing between `CW` and `DW` is realized when requesters need to account for cost-bound constraints (time, money). In such cases `CW` are more desirable. Although workers of other task types are found to be detrimental, detecting each type of workers can facilitate personalized feedback and training that can improve the overall effectiveness of crowd work in the long run. Thus, we argue in favor of the typology-based prediction and pre-selection of workers, more so in tasks with high complexity due to the clear benefits in quality. At the same time, we shed light on the opportunity to identify less-competent workers `LW` and help them improve their performance.

4.8 Chapter Summary

We found that crowd worker behavioral traces can be leveraged to classify workers in a fine-grained worker typology that can be used for better worker pre-selection (**#RQ1, #RQ2**).

We modeled task complexity and studied the impact of task complexity on worker behavior. Based on our experiments and results we have shown that pre-selection based on worker types significantly improves the quality of the results produced, especially in tasks with high complexity (**#RQ3, #RQ4**). We showed that our proposed approach significantly outperforms existing methods for pre-selection. Since our approach is based on gathering behavioral signals from a worker during the pre-screening phase, no prior information about a woker is required. This has important implications on structuring workflow.

We highlight clear benefits of distinguishing beyond *good* and *bad* workers in image transcription and information finding tasks. This is not just useful for requesters to attain better and faster results from crowdsourcing platforms but also goes in the direction of a worker analytics dashboard where crowd workers can be helped to understand their performances and improve over time.

5 Worker Self-Assessments for Competence-based Pre-Selection

> *"I think self-awareness is probably the most important thing towards being a champion."*
> — *Billie Jean King*

Paid crowdsourcing platforms have evolved into remarkable marketplaces where requesters can tap into human intelligence to serve a multitude of purposes, and the workforce can benefit through monetary returns for investing their efforts. In this work, we focus on individual crowd worker competencies. By drawing from self-assessment theories in psychology, we show that crowd workers often lack awareness about their true level of competence. Due to this, although workers intend to maintain a high reputation, they tend to participate in tasks that are beyond their competence. We reveal the diversity of individual worker competencies, and make a case for competence-based pre-selection in crowdsourcing marketplaces. We show the implications of flawed self-assessments on real-world microtasks, and propose a novel worker pre-selection method that considers accuracy of worker self-assessments. We evaluated our method in a sentiment analysis task and observed an improvement in the accuracy by over 15%, when compared to traditional performance-based worker pre-selection. Similarly, our proposed method resulted in an improvement in accuracy of nearly 6% in an image validation task. Our results show that requesters in crowdsourcing platforms can benefit by considering worker self-assessments in addition to their performance for pre-selection.

5.1 Introduction

Researchers and practitioners have actively been both studying and exploiting the crowdsourcing paradigm over the last decade. A recent report regarding the state of *crowdsourcing* in the year 2015 has shed light on the remarkable adoption of crowdsourced solutions to solve a multitude of problems in various industries[1].

Typically in a paid microtask crowdsourcing system, a worker accesses the tasks available and chooses which task(s) to complete. The factors that influence a worker's choice in task selection have been studied in detail in previous works [KSV11b, GKD14d]. The self-centric and subjective nature of task selection on a large crowdsourcing platform (such as Amazon's Mechanical Turk[2] or CrowdFlower[3]) is apparent, i.e., it is up to the crowd workers to select a task according to their interests, preference, or expertise. The increasing popularity of crowdsourcing microtasks along with the range of platforms facilitating such efforts, can lead to an overload of choices for a crowd worker. As pointed out by Barry Schwartz in his influential psychology and social theory works, an overload of choices often tends to have detrimental effects on the decision making process of people [SW04, Sch04]. The large variety of choices in the tasks that are available for an experienced crowd worker [CHMA10] makes it difficult for one to select an appropriate task to complete; workers struggle to find tasks that are most suitable for them.

Prominent marketplaces like Amazon's Mechanical Turk (AMT) or CrowdFlower, that serve as intermediaries to numerous other crowdsourcing channels, gather and accumulate large numbers of diverse tasks. The effort required to search for suitable tasks (in terms of a workers' competencies or interests), or in some cases a lack of alternatives [GKD14d], leads to workers settling for less suitable tasks. The quality of the work thus produced eventually decreases. This is supported by the findings of [CHMA10], where the authors found that workers most often choose tasks from the first page of the 'recently posted tasks', or the first two pages of 'tasks with most available instances'. More recently, a study of the dynamics of microtasks on AMT by Difallah et al. showed that freshly published tasks have almost ten times higher attractiveness for workers as compared to old tasks [DCD$^+$15a]. While some workers settle to work on tasks that are not optimally suited to them, some more capable workers may be deprived of an opportunity to work on the tasks they are ideally suited for, due to limitations on the number of participants or individual contributions. Workers often participate in tasks which are beyond their competence and skills, despite their inherent attempt to maintain their reputation. Thus, the overall effectiveness of the crowdsourcing paradigm decreases.

In order to solve the problem of unsuitable workers participating in tasks, preselection of workers is the popularly adopted solution [OSL$^+$11]. Such pre-screening

[1] http://bit.do/eyeka-crowdsourcing-trend-report

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

[3] http://www.crowdflower.com/

methods are generally based on the performance of workers on prototypical tasks. If a worker passes a prototypical task or a qualification test, then she can proceed to participate in the actual task. This means that the performance of a worker in a prototypical task is assumed to be an indicator of the competence of a worker. In this work, we draw from self-assessment theories in psychology and organizational behavior in order to show that crowd workers often lack an awareness regarding their competence. We build on these theories which suggest that true competence goes hand-in-hand with the *awareness* of competence, or the lack of it [Dun11, DHS04]. In contrast to exisiting methods, we show that by using worker self-assessments as an indicator of competence alongside performance in the pre-screening phase, one can facilitate pre-selection leading to better results in paid crowdsourcing microtasks.

The main contributions of our work stem from (a) investigating whether flawed self-assessments (based on the Dunning-Kruger effect, described in the following section) are prevalent in crowd workers within the microtask crowdsourcing paradigm, and (b) studying the use of self-assessments for worker pre-selection in crowdsourced microtasks. Our contributions are listed below.

- By establishing that some crowd workers fall prey to flawed self-assessments, we show that not all workers are aware of their true competence.
- We show that a worker's estimate of her competence in a task is affected by the objective difficulty-level of the task.
- We show that by using rapidly-prototyped self-assessments within the pre-selection process, requesters can ensure that relatively more competent crowd workers participate in their tasks.
- We evaluated our proposed method on a real-world sentiment analysis task and an image validation task, and found an improvement in the quality of results by over 15% and 6% respectively when compared to the existing state-of-the-art pre-selection method.

5.2 Background and Research Questions

5.2.1 Dunning-Kruger Effect

The Dunning-Kruger effect is a cognitive bias that entails inflated self-assessment and illusionary superiority amongst incompetent individuals [Dun11]. The authors proposed that incompetence in a particular domain reduces the metacognitive ability of individuals to realize it. Skills that encompass competence in a particular domain are often the same skills that are necessary to evaluate competence in that domain. For example, consider the ability to solve a Math problem; the skills required to solve the problem are the same skills that are necessary in order to assess whether the

Math problem has been accurately solved. The authors attribute this bias to the metacognitive inability of incompetent individuals. On the other hand, competent individuals tend to underestimate their relative competence due to falsely assuming that tasks that they find easy are also easy for others. The authors thereby show that incompetent individuals cognitively miscalibrate by erroneously assessing oneselves, while competent individuals miscalibrate by erroneously assessing others. In their studies, the authors investigate the self-assessment of individuals over 4 quartiles of their performance distribution. We compute the quartiles such that the top-quartile consists of individuals whose performance score falls in the top-25% of all scores, and the bottom-quartile consists of individuals whose performance score falls in the bottom-25% of all scores, as shown in Figure 5.1.

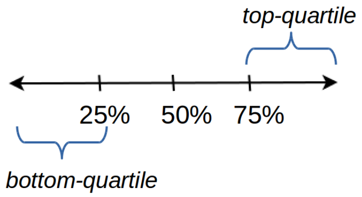


Figure 5.1 The Dunning-Kruger Performance Quartiles.

5.2.2 The Domain of Microtask Crowdsourcing : Motivation

Kruger and Dunning consolidated their findings through 4 studies that addressed a total of 350 Cornell University undergraduate students [KD99]. In our work, we investigate whether the Dunning-Kruger effect can be observed in the paid microtask crowdsourcing paradigm. The characteristic features of paid microtask crowdsourcing are very different in comparison to the controlled environment where undergraduate students were studied. Firstly, there is a large diversity in the demographics of crowd workers [RIS$^+$10, KKMF12]. Secondly, crowd workers have varying motivations to participate in microtask completion, resulting in a wide range of behavior [GKD14d, KSV11b]. Thirdly, while the authors rewarded students with credit points for participating in their studies, we provide monetary incentives to crowd workers. It is noteworthy that our study addresses a considerably larger magnitude of participants (over 2,000 crowd workers). Finally, task difficulty for workers can vary across different tasks. In this chapter, we will use the following terms to refer to crowd workers with different skills.

Definition 1. *Competent workers* are those crowd workers whose performance in a task lies within the *top*-quartile.

Definition 2. *Least-competent workers* are crowd workers whose performance in a task lies within the *bottom*-quartile.

5.2.3 Research Questions and Methodology

We address the following research questions in this chapter.

RQ#1. Can the Dunning-Kruger effect bear implications on the quality of crowdsourced work?

RQ#2. How are crowd worker self-assessments affected by the inherent level of difficulty in a given task?

RQ#3. Can accurate self-assessments of a crowd worker contribute to realize a stronger indicator of the worker's competence, when compared to performance alone in the pre-screening phase of a given task?

Based on the Dunning-Kruger effect, we adapt the following hypotheses (I, II, and III) to fit the crowdsouring paradigm. We presume that by investigating these hypotheses, we can establish the existence and extent of the Dunning-Kruger effect among crowd workers in paid microtask crowdsourcing platforms.

Hypothesis I. *Least-competent crowd workers overestimate their performance with respect to the competent workers, relative to certain objective criteria.* An example of objective criteria in this context is score in a given test.

Hypothesis II. *Least-competent crowd workers are less capable of identifying competence in themselves or other workers, in comparison to competent workers.*

Hypothesis III. *Least-competent crowd workers are less capable of identifying competence in themselves given the responses of the rest of the crowd, in comparison to competent workers.*

To validate the hypotheses we carry out two studies; in *Study-I* we assess whether crowd workers are aware of their competence, drawing comparison between competent and least-competent workers (addressing Hypothesis I, II). In *Study-II* we investigate whether knowledge about responses of other workers has an effect on the performance of competent and least-competent workers (addressing Hypothesis III).

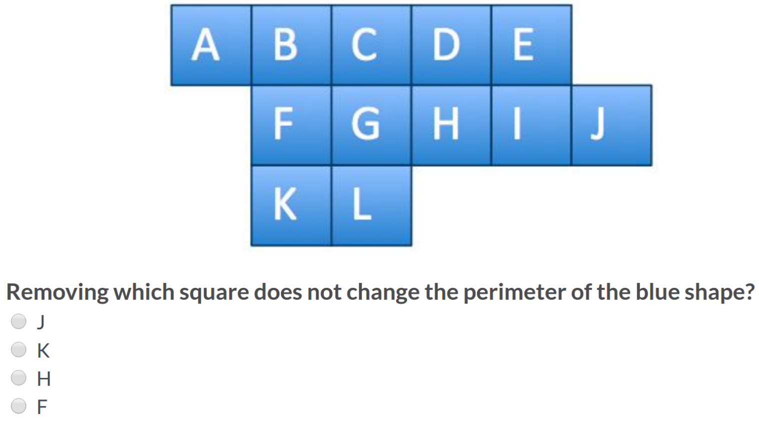
In studies *III, IV* we evaluate whether considering self-assessments of crowd workers can result in realizing a stronger indicator of their true competence. We propose the pre-selection of workers based on their performance and self-assessments, as opposed to traditional pre-selection based on performance alone. In *Study-III* we consider the task of sentiment analysis, and in *Study-IV* we consider an image validation task, since they are popular examples of real-world crowdsourcing microtasks.

5.3 Study I : Self-Assessment of Crowd Workers

Aiming to gather responses from crowd workers and investigate the pre-stated hypotheses (I, II), and to analyze the diversity in competence among crowd workers, we consider the domain of logical reasoning (as in [KD99]).

5.3.1 Microtask Design

The task begins with some basic background and demographic questions. It is then followed by 15 questions in the domain of *logical reasoning*. We used logical reasoning questions from $A + Click$[4], where the questions are based on the Common Core Standards[5]. The Common Core is a set of academic standards in Mathematics and English. These learning goals indicate what a student should know and be able to do at the end of each grade.



Removing which square does not change the perimeter of the blue shape?

○ J
○ K
○ H
○ F

Figure 5.2 An example logical reasoning question from $A + Click$ that was administered to crowd workers in the task corresponding to Grade 5.

To assess the varying competencies among crowd workers and the effect of task difficulty, we deployed 8 tasks on CrowdFlower[6] that are designed similarly except for the difficulty level of the logical reasoning questions. We used graded questions from $A + Click$ to administer logical reasoning questions from the level of Grade 5 to Grade 12. An example is presented in Figure 5.2. Initial empirical tests showed that crowd workers tend to achieve nearly perfect accuracy in logical reasoning tasks that correspond to grades lower than 5. We thereby do not scrutinize grades below 5 further. To separate *trustworthy* workers (TW)[7] from *untrustworthy* workers (UW)[8], we intersperse attention check questions recommended by [MS13b, GKDD15a] as shown in Figure 5.3.

At the end of the logical reasoning questions, workers are requested to answer questions in relation to their performance, corresponding to the following aspects.

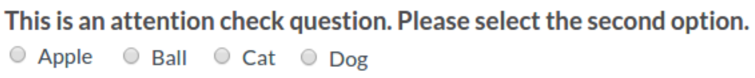
[4] http://www.aplusclick.com/

[5] http://www.corestandards.org/

[6] http://www.crowdflower.com/

[7] Workers who correctly answer all 3 attention check questions embedded in the task.

[8] Workers who incorrectly answer at least 1 of the 3 attention check questions embedded in the task.



This is an attention check question. Please select the second option.

○ Apple ○ Ball ○ Cat ○ Dog

Figure 5.3 Attention check questions to identify *untrustworthy* workers.

- **Perceived test score.** Number of questions that the workers believe to have answered correctly. The corresponding question was phrased as follows – *How many questions do you think you answered correctly?* (answer range: 0-15).
- **Perceived test score of others.** Number of questions on average, that workers think others participating in the task will have answered correctly. The corresponding question was phrased as follows – *On average, how many questions do you think the other workers completing this task will answer correctly?* (answer range: 0-15).
- **Perceived ability.** The expected percentile ranking of the workers. The corresponding question was phrased as follows – *At what percentile ranking (1-100) do you expect to be, with respect to all the workers who will perform this task? '1' indicates the very bottom, '50' indicates exactly average, and '100' indicates the very top* (answer range: 1-100).

Finally, in order to analyze aspects pertaining to real-world tasks, workers were asked to provide as many tags as possible for two different pictures. Tagging images is a popular type of crowdsourced task. Prior research has shown that having verifiable questions such as tags is a recommended way to design tasks and assess crowdsourced results [KCS08].

The order in which different questions were asked did not have an impact on any of the results reported in our work. We thereby do not mention it further. We paid each worker according to a fixed hourly wage of 7.5 USD. In each of the 8 tasks, corresponding to the 8 different graded levels of competencies, we gathered 250 responses from independent workers, resulting in a total of 2,000 crowd workers overall.

5.3.2 Trustworthiness of Workers

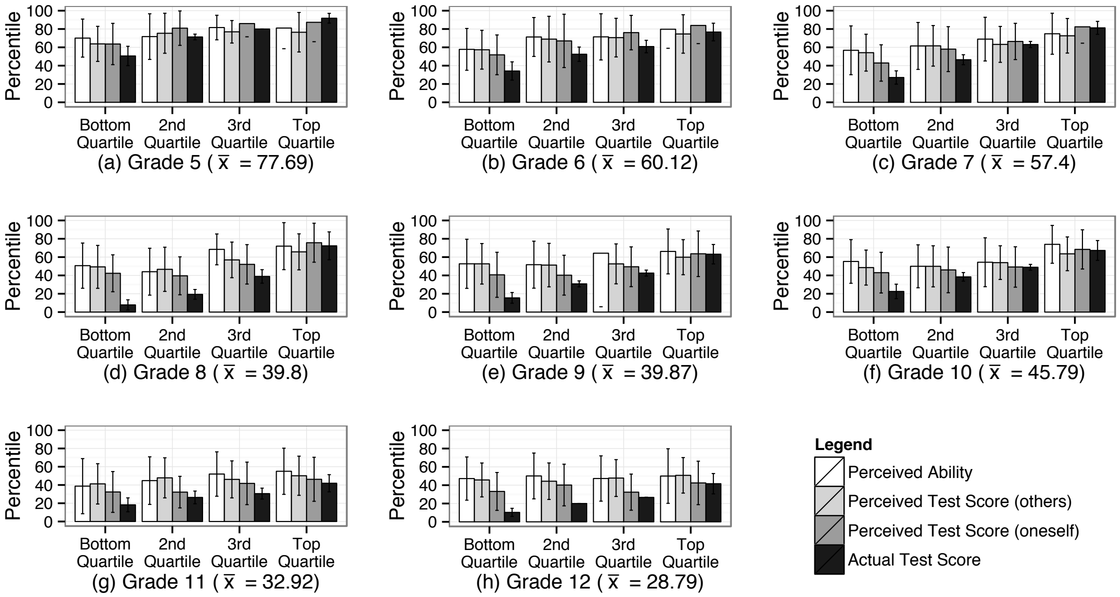
From the responses gathered through the 8 tasks, we first separated trustworthy workers (TW) from untrustworthy workers (UW). Table 5.1 shows the number of TW out of the 250 workers that participated in total, in each grade. On average each grade has around 216 TW participants.

To establish a correlation between the country of origin and the performance of a worker, several experiments that consider aspects such as the time of task deployment, batch size, channels used, and so forth are needed. Addressing the implications of

Grade	G5	G6	G7	G8	G9	G10	G11	G12	Avg. G5-G12
#TW	228	216	226	207	207	215	214	219	216.5
TW (in %)	91.2	86.4	90.4	82.8	82.8	86	85.6	87.6	86.6

Table 5.1 Distribution of Trustworthy Workers (TW) across the graded microtasks.

cultural differences in task performance [MS12] is beyond the scope of this work. Note that we do not consider the UW in the rest of our study and analysis.



(a) Grade 5 (x̄ = 77.69)

(b) Grade 6 (x̄ = 60.12)

(c) Grade 7 (x̄ = 57.4)

(d) Grade 8 (x̄ = 39.8)

(e) Grade 9 (x̄ = 39.87)

(f) Grade 10 (x̄ = 45.79)

(g) Grade 11 (x̄ = 32.92)

(h) Grade 12 (x̄ = 28.79)

Figure 5.4 Perceived test scores and perceived ability of workers across the graded microtasks. Quartiles are presented on the x-axis, percentile on the y-axis, and '$\bar{x}$' is the mean performance of workers in the corresponding grade.

5.4 Results: Self-Assessment of Crowd Workers

5.4.1 Perceived Test Score and Ability of Oneself

We analyzed the responses of each worker for the questions pertaining to *perceived test score* and *perceived ability*. Our findings are presented in the Figure 5.4. We observe that through all the grades (G5-G12), the least-competent workers (i.e. bottom-quartile workers) significantly overestimate their ability and raw test scores. We find that Figure 5.4(a) represents a perfect scenario of the dual fallacy resulting in the self-assessments observed by [KD99]. The least-competent workers overestimate their ability (by nearly 20 percentile points) and performance (by around 13 percentile points). Hence, we observe that least-competent crowd workers cognitively miscalibrate by erraneously assessing themselves, while competent crowd workers miscalibrate by erraneously assessing others (they underestimate their ability by 10 percentile points and performance by nearly 4 percentile points).

With the increase in grade levels (from G5 through G12), we note that least-competent workers depict an increase in the degree of overestimation in the assessment of their ability and performance (perceived test score). A novel finding through our work pertains to that of the competent workers. We note that with an increase in grade levels, competent workers also tend to gradually shift towards overestimation of their ability and performance. We attribute this to the increasing grade levels which potentially go beyond their competence at some point. However, it is clear that least-competent crowd workers indeed overestimate their ability and performance by several percentile points (*M=30.18, SD=9.53*) in comparison to the competent crowd workers (*M=-3.73, SD=7.79*) across all grades (*t(13)=4.22, p<.001*). We found a very large effect size; *Cohen's d = 3*. Thus, we found support for **Hypothesis-I**.

5.4.2 Perceived Test Score of Others

From the plots in Figure 5.4, we clearly observe that least-competent crowd workers greatly miscalibrate their assessment of others in terms of the raw test score (i.e., the number of questions answered correctly by others on average). For instance, consider the Figure 5.4(a) corresponding to G5. Here, least-competent crowd workers placed the average performance of other workers in the 63rd percentile, while the actual mean performance was 77. Competent crowd workers on the other hand fractionally overestimated the average performance of others and placed it in the 78th percentile.

Interestingly however, we note that with the increasing grade levels from G6 through G12, both competent and least-competent workers overestimate the average performance of other workers. Moreover, we find that the degree of *miscalibration* (i.e., the difference between the actual score of a worker and the worker's perceived test score) is more prominent with respect to competent workers. While the actual mean performance was in the 39th percentile on average across all grades, the least-competent workers overestimate the performance of others by around 14 percentile points, and the competent workers overestimate the performance of others by around 25 percentile points. We believe that due to the increasing difficulty inherent to progressive grade levels, competent workers tend to further miscalibrate their relative competence and least-competent workers start recalibrating their relative competence in the accurate direction. Due to the fact that competent workers tend to wrongly believe that their peers are of relatively good competence, they overestimate the performance of others to a greater extent in the higher grades. We thereby found that across all grades competent workers overestimate the performance of others by more percentile points (*M=17.16, SD=7.71*) than incompetent workers (*M=9.41, SD=5.65*), *t(13)=3.01, p<.005*, with a large effect size; *Cohen's d = 1.15*. Thus, we did not find full support for **Hypothesis-II**, stating that 'least-competent crowd workers are less capable of identifying competence in themselves or other workers in comparison to competent workers'.

5.5 Study II : Self-Assessment in the Presence of Others' Answers

To assess whether least-competent workers are capable of identifying their true level of competence given the performance of the rest of the crowd (hypothesis III), we deployed a second set of tasks on CrowdFlower.

5.5.1 Microtask Design

Since we aim to draw a comparison between the competent and least-competent workers alone, we contacted the top and bottom-quartile workers from our priorly completed graded tasks (in Study I) via e-mail and requested them to participate in the subsequent task for each corresponding grade. Over 70% of the top and 60% bottom-quartile workers participated in these tasks over two weeks from deployment. To make valid comparisons across the different grades, we considered the first 60% in each of the top and bottom-quartile workers that participated. These tasks were identical to the initial 8 graded tasks that were deployed (including the incentives offered), with one exception. In this case, we show the overall answer distributions (see Figure 5.5) provided by all workers in the initial round of tasks (in a bar graph) alongside each question in the set of 15 logical reasoning questions.

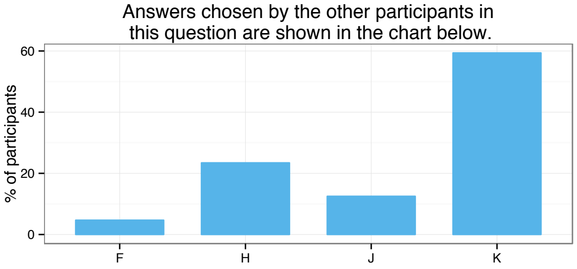
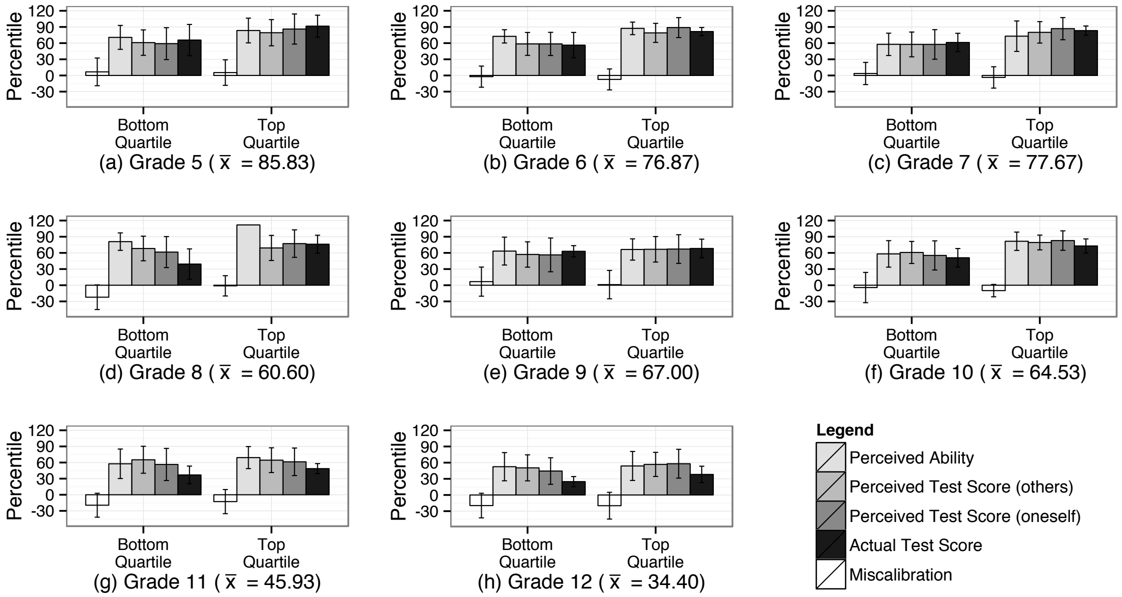


Figure 5.5 The overall answer distribution corresponding to a sample question from Grade 5 (see Figure 5.2), that is shown alongside the question in *Study-II*.



(a) Grade 5 (x̄ = 85.83)

(b) Grade 6 (x̄ = 76.87)

(c) Grade 7 (x̄ = 77.67)

(d) Grade 8 (x̄ = 60.60)

(e) Grade 9 (x̄ = 67.00)

(f) Grade 10 (x̄ = 64.53)

(g) Grade 11 (x̄ = 45.93)

(h) Grade 12 (x̄ = 34.40)

Figure 5.6 Perceived test scores and perceived ability of *competent* and *least-competent* workers across the graded microtasks (G5–G12), in the presence of overall answer distributions for each question. Quartiles are represented on the x-axis, and the y-axis represents percentages of the corresponding attribute. '$\bar{x}$' is the mean performance of workers in the grade.

5.5.2 Results (Study II)

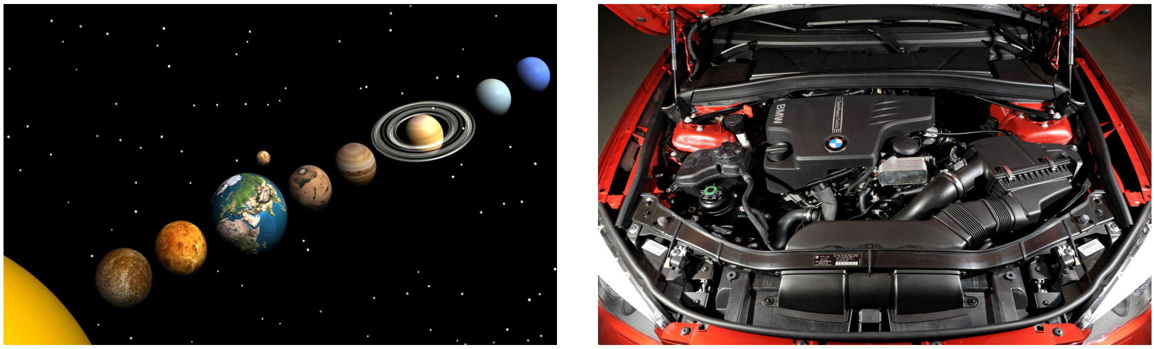
Figure 5.6 presents our findings. We observe that the overall mean performance for each grade improves in comparison to the first set of tasks. This is expected since the workers participate in the same task for the second time. In addition, the workers are aided by the distribution of answers for each question, since they can go with the majority in case they are unsure about certain answers. Our observations are further validated by the completion time of workers in the top and bottom quartiles. In case of top-quartile workers, showing the distribution of answers for each question results in significant reduction in completion time (*M=4.14, SD=1.88*), *t(14)=4.14, p<.001, Cohen's d = 1.34* with a reduction of 4.1 minutes on average across all grades. However, this is not the case for the bottom-quartile workers (*M=1.39, SD=2.21*), where the difference in completion time is not significant, and the reduction in completion time is only 1.39 minutes on average.

In grades G6 and G10, the competent workers depicted a greater degree of miscalibrated self-assessment when compared to the least-competent workers. Thus we note that the *miscalibration* (i.e., the difference between the actual score of a worker and the worker's perceived test score) of least-competent workers is more pronounced, though it is inconsistent. Therefore, we find only partial support for **Hypothesis III**.

5.6 Implications on Real-world Microtasks

Through our findings from Study I and Study II, we can conclude that the Dunning-Kruger effect can be observed in the crowd, subject to the task difficulty at hand. To understand what this difference in competence between top-quartile workers and bottom-quartile workers means in terms of their performance in a real-world microtask, we investigate the tagging task that workers completed at the end of each of the graded tasks in Study I.

To observe the implications of worker competence on a traditional crowdsourcing task like *tagging* (a popular example of content creation tasks [GKD14d]), we analyzed the tags received from least-competent (bottom-quartile) and competent (top-quartile) workers in Study I. Workers were asked to provide as many tags as possible, corresponding to two pictures presented as shown in Figure 5.7. We first processed the responses from crowd workers, so as to ignore meaningless phrases and gibberish tags. We evaluate tags with respect to *quality* (i.e., the reliability of a tag[9]) and *quantity* (i.e., the number of tags).



(a) Picture 1 (Pic#1) – the solar system.

(b) Picture 2 (Pic#2) – the engine of a car.

Figure 5.7 Pictures corresponding to the tagging task that workers were asked to complete at the end of *Study-I*.

Figure 5.8 shows that the total number of tags and unique tags provided by workers decreased gradually with the increase in grade level (adjusted for worker distribution across grades, see Table 5.1). This implies that due to the increasing difficulty with progressive grades, workers exert relatively less effort in providing tags. This is in accordance with findings from prior works that have explored the effect of one microtask on another, and between those with varying difficulty levels [NR16, CIT16]. Corresponding to Pic#1, there were a total of 1,267 tags with 195 unique tags for Grade 5, when compared to 860 tags with 162 unique tags for Grade 12. In case of Pic#2, there were a total of 784 tags with 252 unique tags for Grade 5. This decreased to a total of 692 tags with 197 unique tags for Grade 12.

[9] A tag that is mentioned by at least 10 distinct workers is defined as a *reliable tag*.

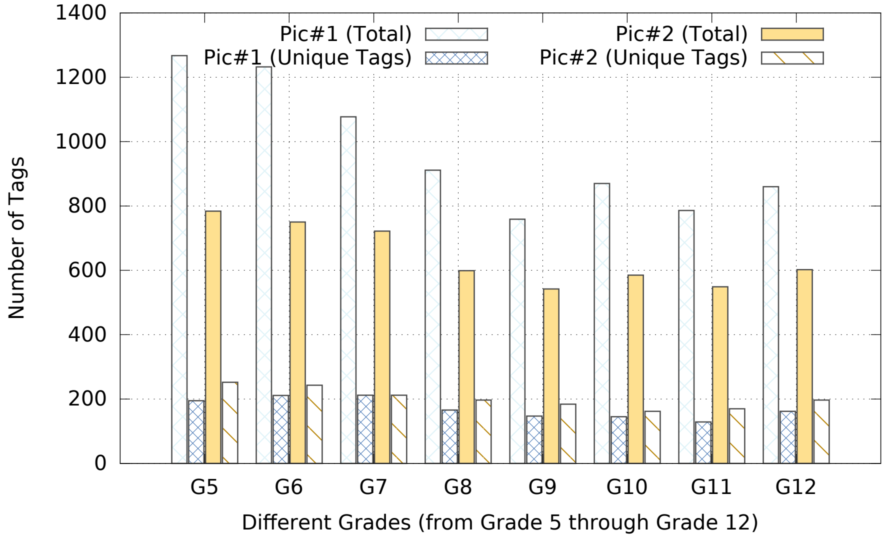


Figure 5.8 Distribution of tags contributed by workers in each of the grades (G5-G12).

We did not find a significant difference in the *quantity* of *reliable tags* across the different grades (G5-G12). On average, for each grade there were around 18 reliable tags corresponding to Pic#1, and nearly 7 corresponding to Pic#2. Figure 5.9 presents the distribution of tags from all grades with respect to the performance quartile. We found that competent workers provided more distinct reliable tags (31 for Pic#1, 25 for Pic#2) than least-competent workers (18 for Pic#1, 8 for Pic#2). These differences in number of reliable tags produced by the competent workers (*M=18.75, SD=3.62*) and least-competent workers (*M=3.75, SD=2.23*) across the grades are found to be statistically significant, $t(11)=4.43$, $p < 0.01$. Our findings suggest that competent (top-quartile) workers provide more reliable tags, with a higher diversity, when compared to least-competent (bottom-quartile) workers.

Operationalizing Worker Self-Assessments

From our findings in Study I, Study II it is evident that not all crowd workers are adept at making accurate self-assessments; competent workers are relatively better at doing so. This is further reinforced by our findings in the tagging task, where we observe that top-quartile workers produce tags with both a higher quality as well and quantity. Based on this understanding, we propose that it can be beneficial to operationalize worker self-assessments as an indicator of worker competence and therefore performance. To do so, we propose to use accuracy of worker self-assessments in the pre-screening tasks in addition to their actual performance in the pre-screening tasks to select workers. Thus, the only additional requirement in our proposed method is a self-assessment question at the end of the pre-screening tasks, making it straight-

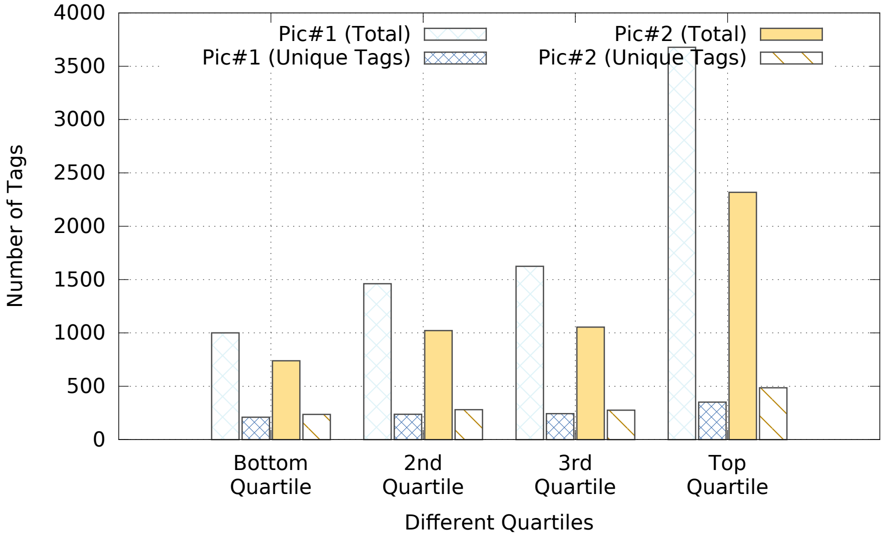


Figure 5.9 Distribution of tags from all grades (G5-G12) across the quartiles.

forward to implement. Figure 5.10 illustrates the traditional pre-screening method in comparison to our proposed self-assessment based pre-screening approach.

5.7 Study III: Evaluation in Sentiment Analysis Task

From our findings in Study I and II we note that some crowd workers (bottom-quartile) exhibit inflated self-assessments. We also found that the top-quartile workers produce significantly better quality of work, as observed in the abridged tagging task of Study I. In Study III, we seek to answer whether we can operationalize the ability of workers to accurately self-assess their performance in a real-world microtask, in order to pre-select a more suitable crowd with respect to the task. Can worker self-assessments be used as a means to provide a stronger indicator of worker competence (#**RQ3**)?

We evaluated our proposed method of using worker self-assessments as a basis for pre-screening crowd workers, as opposed to traditional pre-screening that is purely based on the performance of workers. We considered a popular crowdsourcing task; *sentiment analysis* [GKD14d]. In this task composed of 30 units, crowd workers are asked to read a tweet in each unit and classify the projected sentiment as either `positive`, `negative` or `neutral`. For this purpose we use the dataset introduced by [GFK15b], that consists of expert-classified tweets, thereby providing our ground truth.

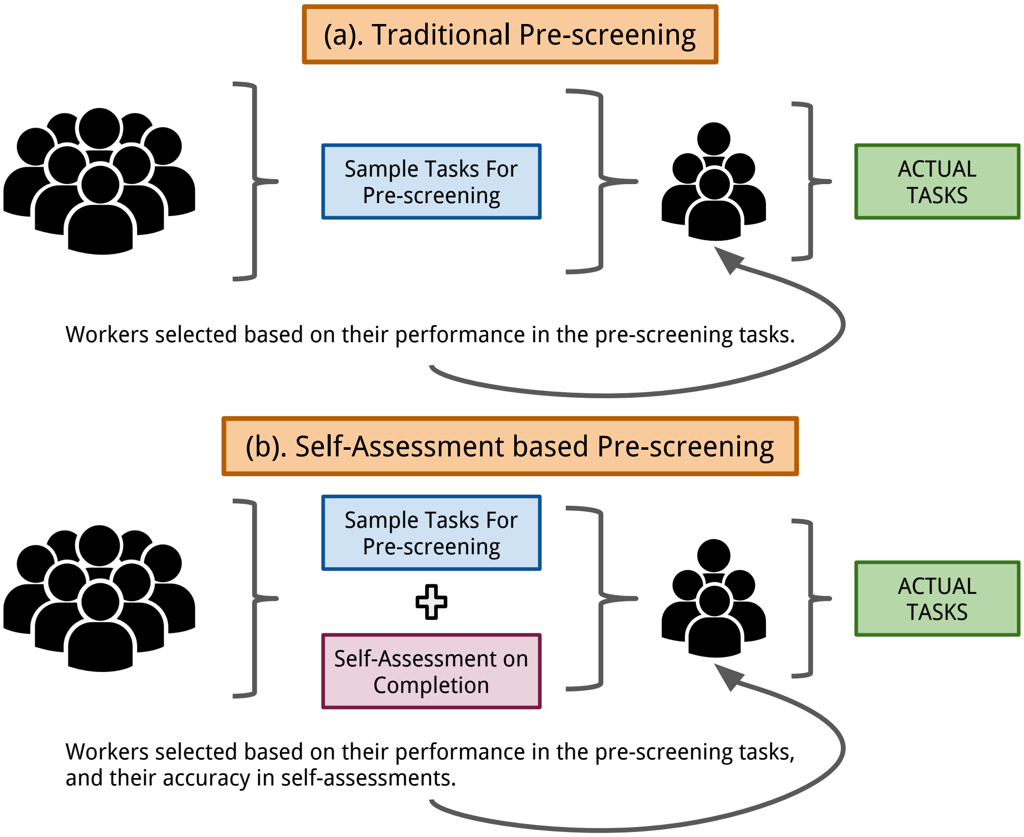

(a). Traditional Pre-screening
Sample Tasks For Pre-screening
ACTUAL TASKS
Workers selected based on their performance in the pre-screening tasks.
(b). Self-Assessment based Pre-screening
Sample Tasks For Pre-screening
Self-Assessment on Completion
ACTUAL TASKS
Workers selected based on their performance in the pre-screening tasks, and their accuracy in self-assessments.

Figure 5.10 Comparison between (a) the traditional pre-screening method based on worker performance in pre-screening tasks, and (b) the self-assessment based pre-screening method which considers worker performance in the pre-screening tasks as well as their accuracy in self-assessments.

5.7.1 Method I : Self-Assessment Based Pre-screening

We prototyped a 5-unit task for the sentiment analysis, consisting of tweets different from those in the actual 30 units considered for the evaluation task. On completing these 5 units, workers are asked the question, '*How many questions do you think you answered correctly?*'. We consider a worker to have passed this screening task, if the worker accurately predicts her score while the actual score is > 3, or if the worker miscalibrates her prediction by one point while her actual score is > 3 (i.e., $miscalibration = 0$ or 1). The intuition behind using a threshold of '3' is due to our aim to replicate a realistic pre-selection scenario. CrowdFlower suggests a minimum accuracy of 70% by default[10] for the traditional pre-screening method (which is actual score > 3 in our case). We deployed this task on CrowdFlower and gathered responses

[10] CrowdFlower's guide to test questions and quality control on: https://success.crowdflower.com/hc/en-us

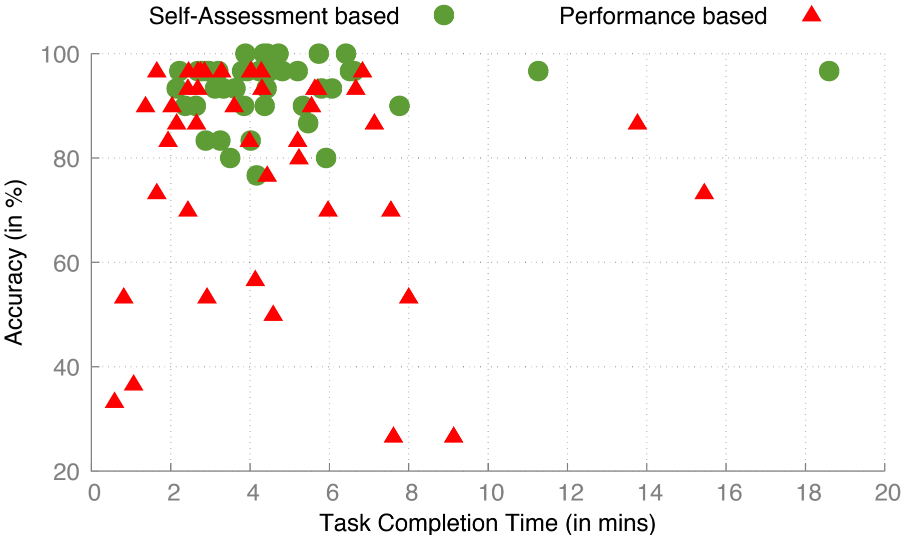


Figure 5.11 Performance of workers acquired by the proposed *Self-Assessment based Pre-Screening* and by traditional *Performance based Pre-Screening*

from 300 workers by offering a compensation of 2 USD cents. We found that only 110 out of 300 workers passed the threshold of actual score $> 3/5$. Of these 70 workers passed the self-assessment accuracy criteria and thereby passed the pre-screening. Next, we deployed the actual evaluation task consisting of 30 units to these 70 workers alone[11] by using their e-mail IDs. We offered a reward of 5 USD cents to workers. Within a span of 1 week, 50 of the 70 workers completed the task.

5.7.2 Method II : Traditional Pre-Screening

One week later, we deployed an identical task consisting of the same 30 units on CrowdFlower. There was no overlap in the pool of workers across the two tasks. Hence, the observed results are not due to ordering effects. We used the same 5 units in the traditional pre-screening process as in the case presented above, and only those workers who answered > 3 units correctly were allowed to participate in the actual task. We gathered responses from 50 distinct workers, and these workers were also paid a compensation of 5 USD cents (to match the incentive offered and number of collected judgments in the self-assessment based pre-screening method.

5.7.3 Results

We evaluated the two different methods based on the following two aspects: accuracy of the pre-selected workers in the tasks following the screening, and their task completion time. We found that the self-assessment based pre-screening method (green dots in Figure 5.11) resulted in workers who performed with an accuracy of nearly 94% on average, with an inter-annotator agreement of 0.95 (computed by pairwise percent agreement (PPA)). The traditional pre-screening method (presented in Figure 5.11 in the red color) resulted in workers who performed with an average accuracy of around 78%, with an inter-annotator agreement of 0.83 (computed by PPA).

We found that the difference in the resulting worker performances between using the self-assessment based pre-screening method (M=27.95, SD=1.79) and the traditional pre-screening method (M=23.63, SD=6.23) was statistically significant *t(95)=3.40, p < 0.01*, with a large effect size; *Cohen's d = .94*. We did not find a significant difference in the task completion time of workers resulting from the two different methods of pre-screening.

It is important to note that in the self-assessment based pre-screening method, the average actual scores of workers on the qualification test was 4.4/5 and that of workers in the traditional pre-screening method was 4.3/5, without a significant difference. This shows that the observed improvement is due to the consideration of worker self-assessments, and not simply a result of selecting workers who performed better in the pre-screening phase. We highlight that there may be a confound in having workers wait, then self-select to return and complete the actual evaluation task in the self-assessment based pre-screening method. Such workers may be more diligent than workers in the traditional pre-screening method, who immediately began the actual evaluation task. However, due to the number of workers in the pool, the significant differences and the large effect size observed, we believe this does not risk the overall result and does not pose a threat to its validity.

From these results, we observe that pre-screening crowd workers based on their self-assessments provides a better reflection of their actual competence, leading to an improved quality of results. We note an improvement of over 15% in accuracy and 12% in agreement between workers by using self-assessment based pre-screening of workers in a sentiment analysis task. Thus, we can conclude that operationalizing self-assessments of workers in a given task in conjuction to their performance in the task, can serve as a stronger indicator of worker competence than relying on worker performance alone.

[11] CrowdFlower provides support for this via the *internal workforce*, https://success.crowdflower.com/hc/en-us/articles/202703355-Contributors-CrowdFlower-s-Internal-Channel.

5.8 Study IV: Evaluation in Verification and Validation Task

In Study III, we operationalized worker self-assessments in a sentiment analysis task and improved the pre-selection of crowd workers. In this Study IV, similar to the sentiment analysis task described in the previous section, we considered an additional real-word task of image validation. Our aim is to verify whether our proposed approach would yield similarly improved results in another type of task, due to the effectiveness of our proposed worker pre-selection method.

In this task composed of 13 units in total, crowd workers were asked to analyze the pictures in online automobile ads to spot mismatched information. To publish an online ad, sellers need to textually describe the state of the vehicle (damaged or not) and its mileage. Sellers commonly omit damage-related information from the description or claim a lower mileage in order to achieve a better placement in the search results (see Figure 5.12). In many cases this information is evident in the pictures. While this cannot be easily detected by automated algorithms, it is a rather simple task for humans.

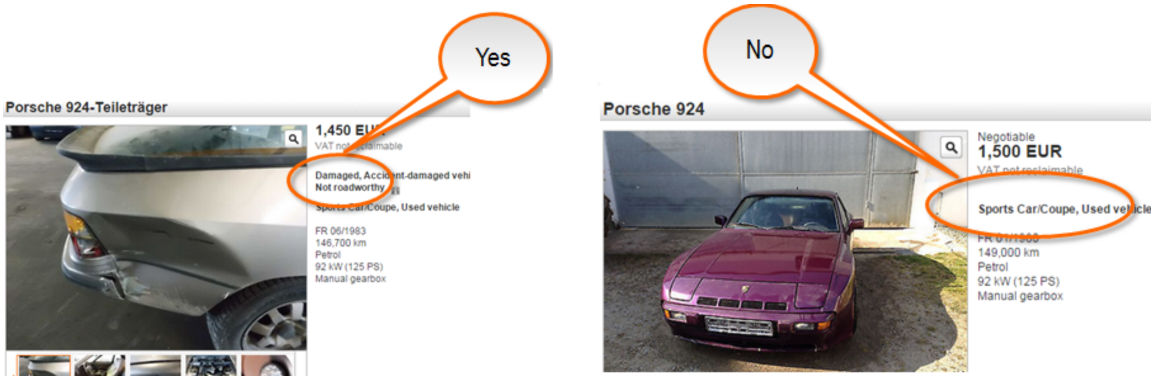
5.8.1 Task Setup

We manually found and annotated a total of 13 vehicle ads[12] which served as groundtruth for the task. Each ad corresponds to one unit where workers are asked to answer three multiple choice questions: (i) Is the car marked as damaged? (ii) Can you identify that the car has a visible damage or functional problems based on the pictures? (iii) Is the mileage information consistent with the picture? We took care to find distinct ads that produced an even distribution of the options corresponding to each question. The units were randomized and after answering 3 units (total of 9 questions), workers were asked to assess their performance on the 9 questions. With an aim to compare self-assessment based pre-screening with performance based pre-screening, all workers were allowed to continue onto 10 more units. Each worker was rewarded with 5 USD cents on successful task completion. We deployed this task on CrowdFlower and collected responses from 100 distinct workers.

5.8.2 Results

Traditional Pre-screening: Similiar to the previous sentiment analysis task, the traditional pre-screening method is characterized by a performance threshold of 70% in the pre-screening phase. Thus, we filtered out workers (36 in total) who did not achieve a minimum of 70% accuracy in the first 3 units (9 questions). In the 10 units that followed, comprising the actual task, this group of workers (*N=64*) achieved an

[12] We used publicly available ads from the online marketplace http://www.mobile.de/



Porsche 924-Teileträger

1,450 EU

VAT no claimable

Damaged, Accident-damaged vehi
Not roadworthy

Sports Car/Coupe, Used vehicle

FR 06/1983
146,700 km
Petrol
92 kW (125 PS)
Manual gearbox

Porsche 924

Negotiable
1,500 EUR
VAT not reclaimable

Sports Car/Coupe, Used vehicle

149,000 km
Petrol
92 kW (125 PS)
Manual gearbox

(a) Seller declared visible damage in the description of the advertisement.

(b) Seller omitted visible damage-related details from the description of the advertisement.

Figure 5.12 Example automobile ads from the online marketplace `mobile.de` that either (a) declare damages in the vehicle description, or (b) omit damage-related information.

average accuracy of 84.05% *(M=84.05, SD=10.35)*, with an inter-annotator agreement of 0.81 using pairwise percent agreement (PPA).

Self-Assessment Based Pre-screening: In case of the proposed self-assessment based pre-screening approach, we consider the accuracy of worker self-assessments in addition to the 70% accuracy threshold in the pre-screening phase. Here again, we tolerate an error of 1 point in the workers self-assessments (i.e., $miscalibration = 0$ or 1). Workers who passed this pre-screening phase (*N=49*), performed with an accuracy of 89.6% (*(M=89.6, SD=6.6*) in 10 units that followed, comprising the actual task. In this case, the inter-annotator agreement was found to be 0.9 (PPA).

To summarize, we found that 64 of the 100 workers passed 70% accuracy threshold. Of these, 49 workers passed the self-assessment accuracy criteria and thereby passed the pre-screening. The self-assessment based pre-screening approach resulted in an improvement in accuracy of nearly 6%, and an increase in the inter-annotator agreement between workers by 8% in comparison to the traditional pre-screening method. The difference in worker accuracy between the traditional and the self-assessment based pre-screening methods was found to be statistically significant with a moderately large effect size; *t(112)=2.60, p<.01*, Hedge's *g = .62*. Once again, we noted that the difference in performance in the pre-screening phase (3 units, 9 questions) across the two groups of workers was not statistically significant, indicating that the improvement in the accuracy of workers using our proposed approach is due to the consideration of accuracy of workers' self-assessments. We also did not find a significant difference in the task completion time of workers selected using the different methods.

5.9 Discussion, Caveats and Limitations

5.9.1 Self-Assessments for Competence-based Pre-Selection

Through our experimental findings and evaluation, we observe that using worker self-assessments for competence-based pre-selection can provide a stronger indicator of worker competence and potential performance to requesters. Since workers need to answer only one additional question with regard to estimating their performance, the time overhead in comparison to performance-based pre-selection is negligible. Due to the same reason, rapidly protyping a self-assessment based pre-selection phase requires relatively the same effort from a requester's point of view. Moreover, since there is an improvement in the quality of the results produced, requesters can improve their costs-benefit ratio with respect to a given task. Requesters can also adjust the passing threshold in the pre-selection process to suit their needs. However, this approach may entail a loss in workforce due to more effective pre-selection and thereby increase the overall task completion time. In Study III, around 37% of the 300 workers passed the traditional pre-screening method, while around 24% of the workers passed the self-assessment based pre-screening method. Similarly in Study IV, 64% of the workers passed the traditional pre-screening method and 49% of the workers passed the self-assessment based pre-screening method. On average across the two studies, we note a loss in workforce of less than 14% resulting from the self-assessment based pre-screening method in comparison to the traditional method. Due to the abundance of crowd workers and in the interest of significantly improved results, we believe our proposed approach will lead to meaningful trade-offs. It is important to note that other quality control measures can be easily used in addition to the self-assessment based pre-selection method to further improve the quality of the crowdsourced work.

From our results in Study III and Study IV, we note that the proposed approach yields better results in comparison to traditional pre-screening methods across the two different types of tasks considered. We note that the self-assessments based pre-screening method results in a relatively larger improvement in the sentiment analysis task (considered in Study III) than in the image validation task (considered in Study IV). While this reflects on the generalizability of the proposed approach across task types, the results also indicate that the method can be effective to varying degrees. We reason that this difference is due to the inherent difficulty levels of the task types considered. Further experiments are required to gauge the impact of the proposed approach under the interaction of different task types and task difficulty.

Pre-selection of workers according to our proposed self-assessments based pre-screening approach can mean that workers in such stystems may not get to work on tasks that go beyond their competence. This can be a limitation since challenging tasks can be more interesting and play a developmental role for workers. The resulting potential power imbalance between workers and requesters in terms of using self-assessments for pre-selection, can be overcome by using the self-assessments of workers to also raise their self-awareness, thereby playing a constructive role in supporting the growth [Bou13] of workers and developing crowd work. We will explore the use of self-assessments to increase worker self-awareness in future work.

5.9.2 Worker Competence Transferability

In our experimental results in the abridged *tagging* task, we observed the implications of competent and least-competent workers on the quality, reliability and the diversity of the tags produced. In this case, we assessed the competence of workers based on the *logical reasoning* task and applied the resulting characterization to the tagging task. By doing so we found that competent workers exhibit a better performance. However, such transferability of worker competence from one type of task to another needs to be studied further. While we cannot assume the universal transferability of a worker's competence that is assessed in one domain alone, an understanding of transferable domains will reduce further costs (in terms of time and money) that are incurred through pre-selection processes. Our proposed approach is to rapidly prototype a given task and use worker self-assessments to assess worker competence in a pre-selection phase. Due to this reason, we carried out further evaluations of worker self-assessment based competence estimation in a sentiment analysis task, and an image validation task.

5.9.3 Training Crowd Workers to Increase Competence

In their studies, Kruger and Dunning also studied the effect of training less-competent individuals [KD99]. The authors found that through systemic feedback and training, less-competent individuals can progress towards higher competence, leading them to become more self-aware. However, the impact of learning or training on individuals' self-assessments has attracted several debates on both sides ([SDJK13], [MG11]). While Schosser et al. found no evidence of learning that leads to consequent improvement in performance of incompetent individuals [SDJK13], Miller and Geraci cite contrasting evidence through their experiments [MG11].

Recent works have studied the impact of providing feedback and training workers in crowdsourcing microtasks [GFK15b, LEHB10]. Through a series of empirical experiments on different types of microtasks, Gadiraju et al. have shown that the performance of workers can be improved by providing training. In the context of our work in this chapter, the findings of Gadiraju et al. [GFK15b] can be extrapolated to reason that training least-competent workers can help them improve their competence, and thereby improve the calibration of their self-assessments. However, further scrutiny is required in order to understand the impact of training on crowd workers' self-assessments and competence.

5.9.4 Other Considerations

It is important to explore whether there are cross-cultural differences in how the Dunning-Kruger effect manifests, that can further dictate the use of self-assessments for pre-selection in tasks. For example, does the perception of their own performance vary across worker groups having different ethnicity? We conducted a one-way between workers ANOVA to compare the effect of *ethnicity* of workers on the perception of their own performance across all grades in 7 ethinicity-group conditions (African American, American Indian, Asian, Hispanic, Pacific Islander, White, Other) as indicated by workers in Study I. We found that there was no significant effect of ethnicity of workers on the perception of their performance at the $p < .5$ level for the 7 ethnicity-group conditions [*F(6, 1725)=0.4229, p=0.86*]. Post-hoc comparisons using the Tukey HSD test confirmed that there was no significant difference in perception of the workers' own performance between any two of the different ethnicity groups.

To give workers a fair chance to participate in a task while using self-assessments as pre-screening method, an important caveat is to ensure that the workers are aware that the selection is based on both, their performance and the accuracy of their self-assessment. Otherwise, workers may inflate their self-assessment with the belief that a higher assessment would lead to their participation in the task. Isolating workers who miscalibrate their self-assessments due to such inflation is beyond the scope of our work. Nevertheless, it is noteworthy that our proposed approach for worker pre-selection is effective in yielding improved results.

5.10 Related Literature

5.10.1 Self-Assessment

Apart from the priorly discussed work of Kruger and Dunning [KD99], there have been several other noteworthy works in the realm of individual self-assessment. Research works have shown that people provide inflated self-evaluations on performance in a number of different real world settings. Dunning et al. showed and discussed the implications of such flawed self-assessments on health, educational settings and the general workplace [DHS04].

Kulkarni et al. showed that in an online course addressing a large number of students (MOOC), the students graded their work 7% higher than those assigned by the staff on average [KWL+15]. Other exisiting data from experiments reinforce the mistaken self-evaluation of performance [ED03, EJB+08]. These works show that incompetent individuals are worse at assessing the quality of performance and often tend to think that they outperform the majority, while in fact they belong to the lower rungs of the performance quartile. Complementing these existing works on self-assessment, in our work we aim to understand whether the flawed self-assessment theories hold among crowd workers in the crowdsourcing paradigm. In contrast to these studies that are largely based on self-selected groups of individuals leading to potential selection bias, we use the crowd as a source for a diverse landscape of individuals with respect to their demographics, skills and competence.

Despite a considerable number of works that assert the findings from the Dunning-Kruger effect, the underlying reasons that dictate the dual-curse resulting in the miscalibrated self-assessment have been widely contested [BLK06, KO08, KM02]. Several researchers have provided alternative accounts for the Dunning-Kruger effect, alluding it to regression to the mean and the above-average effect. These accounts have in turn resulted in rigorous theoretical responses and empirical refutations [EJB$^+$08], and are out of the scope of our work in this chapter.

In closely related work that proposes the use of self-assessments to improve crowd work, Dow et al. showed that self-assessments allowed workers to improve over time in a task involving writing consumer reviews of products they owned [DKKH12]. The authors of this work proposed the use of self-assessments to yield better work quality by promoting self-reflection and learning. In contrast, we propose to consider the accuracy of worker self-assessments alongside their task accuracy in a pre-selection phase as an indicator of their true competence and potential performance. Thus, we develop a distinct and novel approach by directly leveraging self-assessments as a worker filtering mechanism, rather than aiming to improve work through self-review.

5.10.2 Competence of Crowd Workers

The crux of prior research works in the realm of characterizing crowd workers has mainly focused on ensuring reliability of workers, and presenting a means to the requester to pre-select prospective workers [KNB$^+$13]. In this regard, researchers have suggested the use of pre-screening methods and qualification tests [Kaz11], trust models to predict the probability of reliable responses [YSMA12], hidden gold standard questions [OSL$^+$11], and the use of metrics that quantify acceptability of responses from the crowd [GKDD15a]. In this chapter, we propose a novel method for the preselection of workers, that outperforms traditional performance based pre-screening methods.

Kazai et al. [KKMF11] used behavioral observations to typecast workers as one of *Spammer*; *Sloppy*; *Incompetent*; *Competent*; or *Diligent*. Here the authors take a keen interest in designing this typology with an aim to attract workers with desirable features, rather than to understand the competencies of the worker population.

As discussed by Dukat and Caton [DC13], these existing approaches are seldom applied to ascertain actual worker competencies. They merely serve as an indicator for whether a worker is likely to possess the required ability to complete a microtask successfully, and whether a worker is trustworthy. In this chapter, we present an understanding of the diversity in competence of individual crowd workers.

In closely related works by Kosinski and Bachrach et al., the authors measured the performance of crowd workers on a standard IQ questionnaire [BGK+12, KBK+12]. The authors however, discuss factors that effect the overall performance such as compostition of the crowd, reputation of workers and monetary rewards. Finally, the authors discuss an approach to aggregate responses from crowd workers to boost performance. While in these works the authors show that aggregating responses from crowd workers is a profitable approach, in this chapter, we are more interested in the individual competence of workers, and therefore adopt a more granular view of responses.

Previous works have highlighted the importance of building tools that support crowd work from the perspective of workers, in order to address the power asymmetry in exisiting crowdsourcing platforms such as AMT [IS13, MHOG14, MOGH16b]. In addition to this, Kittur et al. identified *facilitation of learning* as an important next step towards building a bright future for crowd work [KNB+13]. Complementary to these initiatives, we propose the use of self-assessments in pre-selection of workers to aid requesters in recruiting the desired crowd. In the future, we can explore the potential use of self-assessments to help workers increase their self-awareness, identify and potentially facilitate learning where their skills are lacking. Thus, we believe that there can be promising new directions based on leveraging workers' self-assessments to support and improve crowd work in various domains.

5.11 Chapter Summary

Our work presented in this chapter has important implications on paid microtask crowdsourcing systems, since we show that there is a disparity in the crowd regarding the metacognitive ability of workers. This hinders the performance of workers and deprives learning. Through our experiments and results presented in this work, we see evidence of the Dunning-Kruger effect in the paid crowdsourcing paradigm. After studying the impact of inherent task difficulty in the logical reasoning task, and exploring three hypotheses, we draw conclusions and highlight the novel contributions.

The important contribution that our work adds to exisiting literature on self-assessment is the impact of task difficulty on the Dunning-Kruger effect among crowd workers. In tasks with relatively lower difficulty (lower grades), we clearly observe the Dunning-Kruger effect. However, we note that with an increase in grade levels, competent workers also tend to gradually shift towards over-estimation of their ability and performance. This is explained by the fact that the higher grades go beyond the capabilities of even the competent crowd workers (research questions **RQ#1, RQ#2**).

The capability of a worker to accurately self-evaluate is an integral aspect of the worker's competence. Through our rigorous evaluation in tagging, sentiment analysis and image validation tasks, we have observed that crowdsourcing microtask requesters can benefit by operationalizing workers' self-assessments as a means of assessing their competence rather than relying solely on their performance in worker pre-selection phases (**RQ#3**). We find that workers pre-selected using our proposed approach exhibit a significantly higher accuracy, than those that are obtained using a traditional pre-screening method.

6

The Role of Task Clarity in Microtask Crowdsourcing

> *"For me the greatest beauty always lies in the greatest clarity."*
>
> *— Gotthold Ephraim Lessing*

Workers of microtask crowdsourcing marketplaces strive to find a balance between the need for monetary income and the need for high reputation. Such balance is often threatened by poorly formulated tasks, as workers attempt their execution despite a sub-optimal understanding of the work to be done.

In this chapter we highlight the role of *clarity* as a characterising property of tasks in crowdsourcing. We surveyed 100 workers of the CrowdFlower platform to verify the presence of issues with task clarity in crowdsourcing marketplaces, reveal how crowd workers deal with such issues, and motivate the need for mechanisms that can predict and measure task clarity. Next, we propose a novel model for task clarity based on the *goal* and *role* clarity constructs. We sampled 7.1K tasks from the Amazon mTurk marketplace, and acquired labels for task clarity from crowd workers. We show that task clarity is coherently perceived by crowd workers, and is affected by the type of the task. We then propose a set of features to capture task clarity, and use the acquired labels to train and validate a supervised machine learning model for task clarity prediction. Finally, we perform a long-term analysis of the evolution of task clarity on Amazon mTurk, and show that clarity is not a property suitable for temporal characterisation.

6.1 Introduction

Microtask crowdsourcing has become an appealing approach for data collection and augmentation purposes, as demonstrated by the consistent growth of crowdsourcing marketplaces such as Amazon Mechanical Turk[1] and CrowdFlower[2].

Task consumption in microtask crowdsourcing platforms is mostly driven by a self-selection process, where workers meeting the required eligibility criteria select the tasks that they prefer to work on. Workers strive to maintain high reputation and performance to access more tasks, while maximizing monetary income. When discussing such a trade-off, the dominant narrative suggests that workers are more interested in obtaining their rewards, than in executing good work. We challenge this widespread opinion by focusing on an often neglected component of microtask crowdsourcing: the *clarity* of task description and instructions in terms of comprehensibility for workers.

Poor formulation of tasks has clear consequences: to compensate for a lack of alternatives in the marketplace, workers often attempt the execution of tasks despite a sub-optimal understanding of the work to be done. On the other hand, requesters are often not aware of issues with their task design, thus considering unsatisfactory work as evidence of malicious behaviour and deny rewards. As a result, crowd workers get demotivated, the overall quality of work produced decreases, and all actors lose confidence in the marketplace. Despite the intuitive importance of task *clarity* for microtask crowdsourcing, there is no clear understanding of the extent by which the lack of clarity in task description and instructions impacts worker performance, ultimately affecting the quality of work.

Research Questions and Original Contributions. This chapter aims at filling this knowledge gap by contributing novel insights on the nature and importance of task clarity in microtask crowdsourcing. By combining qualitative and quantitative analysis, we seek to answer the following research questions.

RQ1: What makes the specification of a task unclear to crowd workers? How do workers deal with unclear tasks?

First, we investigate if clarity is indeed a concern for workers. We designed and deployed a survey on the CrowdFlower platform, where we asked workers to describe what makes a task unclear, and to illustrate their strategies for dealing with unclear tasks. The survey involved 100 workers, and clearly highlights that workers confront unclear tasks on a regular basis.

Some workers attempt to overcome the difficulties they face with inadequate instructions, and unclear language by using external help, dictionaries or translators.

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

[2] http://www.crowdflower.com/

Several workers tend to complete unclear tasks despite not understanding the objectives entirely.

These results demonstrate the need for methods for task clarity measurement and prediction, and shaped the formulation of the following questions.

RQ2: How is the clarity of crowdsourcing tasks perceived by workers, and distributed over tasks?

Inspired by work performed in the field of organisational psychology, we consider clarity both in the context of *what* needs to be produced by the worker (*goal clarity*) and *how* such work should be performed (*role clarity*). We sampled 7.1K tasks from a 5 years worth dataset of the Amazon mTurk marketplace. Tasks were published on CrowdFlower to collect clarity assessments from workers. Results show that task clarity is coherently perceived by crowd workers, and is affected by the type of the task. We unveil a significant lack of correlation between the *clarity* and the *complexity* of tasks, thus showing that these two properties orthogonally characterise microwork tasks.

RQ3: Which features can characterise the goal and role clarity of a task? Using such features, to what extent can task clarity be predicted?

We propose a set of features based on the metadata of tasks, task type, task content, and task readability to capture task clarity. We use the acquired labels to train and validate a supervised machine learning model for task clarity prediction. Our proposed model to predict task clarity on a 5-point scale achieves a mean absolute error (*MAE*) of *0.4 (SD=.003)*, indicating that task clarity can be accurately predicted.

RQ4: To what extent is task clarity a macro-property of the Amazon mTurk ecosystem?

We analyzed 7.1K tasks to understand how task clarity evolves over time. We found that the overall task clarity in the marketplace fluctuates over time, albeit without a discernible pattern. We found a weak positive correlation between the average task clarity and the number of tasks deployed by requesters over time, but no significant effect of the number of tasks deployed by requesters on the magnitude of change in task clarity.

6.2 Related Literature

6.2.1 Text readability

Readability has been defined as the sum of all elements in text that affect a reader's understanding, reading speed and level of interest in the material [DC49]. There has been a lot of work in the past on analyzing the readability of text, as summarized in [CT14]. Early works range from simple approaches that focus on the semantic and syntactic complexity of text [KFJRC75], or vocabulary based approaches where semantic difficulty is operationalized by means of gathering information on the average vocabulary of a certain age or social status group [CD95]. More recently, authors proposed statistical language models to compute readability [CTC04]. Other works studied the lexical richness of text by capturing the range and diversity of vocabulary in given text [MR12]. Several machine learning models have also been proposed to predict the readability of text [PN08, KLP$^+$10]. De Clerq et al. recently investigated the use of crowdsourcing for assessing readability [DCHD$^+$14]. The vast body of literature corresponding to text readability has also resulted in several software packages and tools to compute readability [GMLC04, CKM15].

In this chapter, we draw inspiration from related literature on text readability in order to construct features that aid in the prediction of task clarity on crowdsourcing platforms.

6.2.2 Task Clarity in Microtask Crowdsourcing

Research works in the field of microtask crowdsourcing have referred to the importance of task clarity tangentially; several authors have stressed about the positive impact of task design, clear instructions and descriptions on the quality of crowdsourced work [MS13b, SHC11, KCS08, Ber16]. Grady and Lease pointed out the importance of wording and terminology in designing crowdsourcing tasks effectively [GL10]. Alonso and Baeza-Yates recommended providing 'clear and colloquial' instructions as an important part of task design [ABY11]. Kittur et al. identified 'improving task design through better communication' as one of the pivotal next steps in designing efficient crowdsourcing solutions in the future [KNB$^+$13]. The authors elaborated that task instructions are often ambiguous and incomplete, do not address boundary cases, and do not provide adequate examples. Khanna et al. studied usability barriers that were prevalent on AMT, which prevented workers with little digital literacy skills from participating and completing work on AMT [KRDT10]. The authors showed that the task instructions, user interface, and the workers' cultural context corresponded to key usability barriers. To overcome such usability obstacles on AMT and better enable access and participation of low-income workers in India, the authors proposed the use of simplified user interfaces, simplified task instructions, and language localization. More recently, Yang et al. investigated the role of *task complexity* in worker performance, with an aim to better the understanding of task-related elements that aid or deter crowd work [YRDB16].

While the importance of task clarity has been acknowledged in the microtask crowdsourcing community, there is neither a model that describes task clarity nor a measure to quantify it. In this chapter, we not only propose a model for task clarity, but we also present a means to measure it. To the best of our knowledge, this is the first work that thoroughly investigates the features that determine task clarity in microtask crowdsourcing, and provides an analysis of the evolution of task clarity.

6.2.3 Task Clarity in Other Domains

In the field of organizational psychology, researchers have studied how the sexual composition of groups affects the authority behavior of group leaders in cases where the task clarity is either *high* or *low* [RN77]. In this case, the authors defined task clarity as the degree to which the goal (i.e., the desired outcome of an activity) and the role (i.e., the activities performed by an actor during the course of a task) are clear to a group leader. In self-regulated learning, researchers have widely studied task interpretation as summarized in [RR15]. Hadwin proposed a model that suggests the role of the following three aspects in task interpretation and understanding; (i) implicit aspects, (ii) explicit aspects, and socio-contextual aspects [Had06, HOMW09]. Recent literature regarding task interpretation in the learning field has revolved around text decoding, instructional practices or perceptions of tasks on the one hand [BvHWRvdB02, JN04, LLT01], and socio-contextual aspects of task interpretation such as beliefs about expertise, ability, and knowledge on the other hand [CCE04, DBT05].

Inspired by the modeling of task clarity in the context of authority behavior in Psychology, we model task clarity as a combination of *goal clarity* and *role clarity* (as explained in Section 6.4).

6.3 Are Crowdsoucred Microtasks Always Clear?

We aim to investigate whether or not workers believe task clarity to impact their work performance (**RQ1**). We thereby deployed a survey consisting of various questions ranging from general demographics of the crowd to questions regarding their experiences while completing microtasks on crowdsourcing platforms.

6.3.1 Methodology

We deployed the survey on CrowdFlower[3] and gathered responses from 100 distinct crowd workers. To detect untrustworthy workers and ensure reliability of the responses received, we follow recommended guidelines for ensuring high quality results in surveys [GKDD15c]. To this end, we intersperse two attention check questions within the survey. In addition, we use the filter provided by CrowdFlower to ensure the participation of only high quality workers (i.e., *level 3* crowd workers as prescribed on the CrowdFlower platform). We flagged workers who failed to pass at least one of the two attention check questions and do not consider them in our analysis.

6.3.2 Analysis and Findings

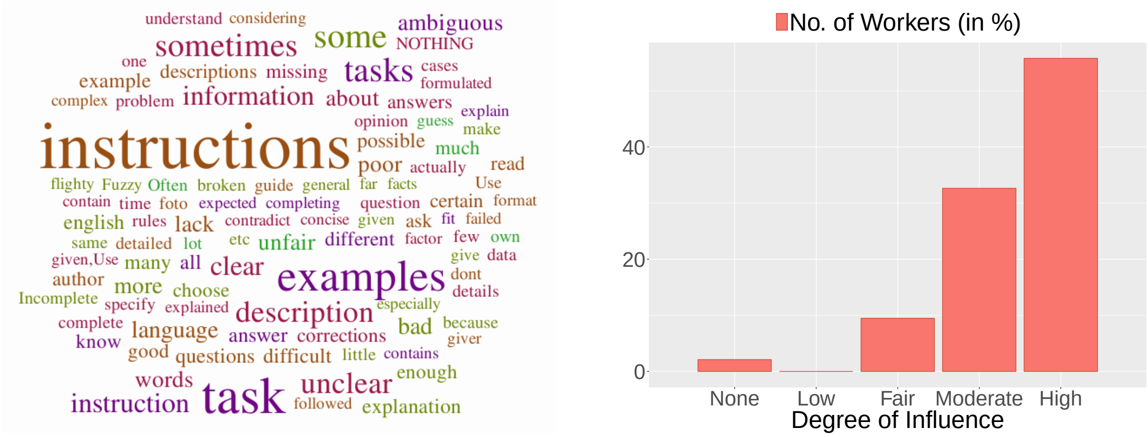
Worker's Experience. We found that around 36% of the workers who completed the survey earn their primary source of income through crowd work. 32.6% of the workers claim to have been contributing piecework through crowdsourcing platforms over the last 3 to 6 months. 63.2% of the workers have been doing so for the last 1 to 3 years. A small fraction of workers (3.2%) claim to have been working on microtasks for the last 3 to 5 years, while 1% of the worker population has been contributing to crowdsourced microtasks for over 5 years. During the course of this time, almost 74% of workers claim to have completed over 500 different tasks.

What factors make tasks unclear? We asked the workers to provide details regarding the factors that they believe make tasks unclear, in an open text field. The word-cloud in Figure 6.1(a) represents the responses collected from the crowd workers. Workers complained about the task instructions and descriptions being '*vague*', '*blank*', '*unclear*', '*inconsistent*', '*imprecise*', '*ambiguous*', or '*poor*'. Workers also complained about the language used; '*too many words*', '*high standard of English*', '*broken English*', '*spelling*', and so forth. Workers also pointed out that adequate examples are seldom provided by requesters. Excerpts of these responses are presented on the companion webpage[4].

Task Clarity and Influence on Performance. Around 49% of workers claimed that up to a maximum of 30% of the tasks that they worked on were unclear. 37% of workers claimed that between 31-60% of the tasks they completed lacked clarity, while 14% of the workers claimed that more than 60% of their completed tasks were unclear. We also asked the workers about the perceived influence of task clarity on their performance. Our findings are presented in the Figure 6.1(b). A large majority of workers believe that task clarity has a quantifiable influence on their performance. We also asked workers about the frequency of encounter for tasks containing difficult words, which might have hindered their performance. Figure 6.2(a) depicts our findings, indicating that workers observed tasks which contained

[3] http://crowdflower.com

[4] https://sites.google.com/site/ht2017clarity/



(a)

(b) Information Finding Tasks

Figure 6.1 (a) Word-cloud representing factors cited by workers that make tasks unclear. Size of words indicate frequency. (b) Degree of influence of task clarity on performance.

difficult words reasonably frequently.

How do workers deal with unclear tasks? We investigated the frequency with which workers complete tasks despite the lack of clarity. As shown in Figure 6.2(b), we found that nearly 27% of workers complete less than 10% of the unclear tasks that they encounter.

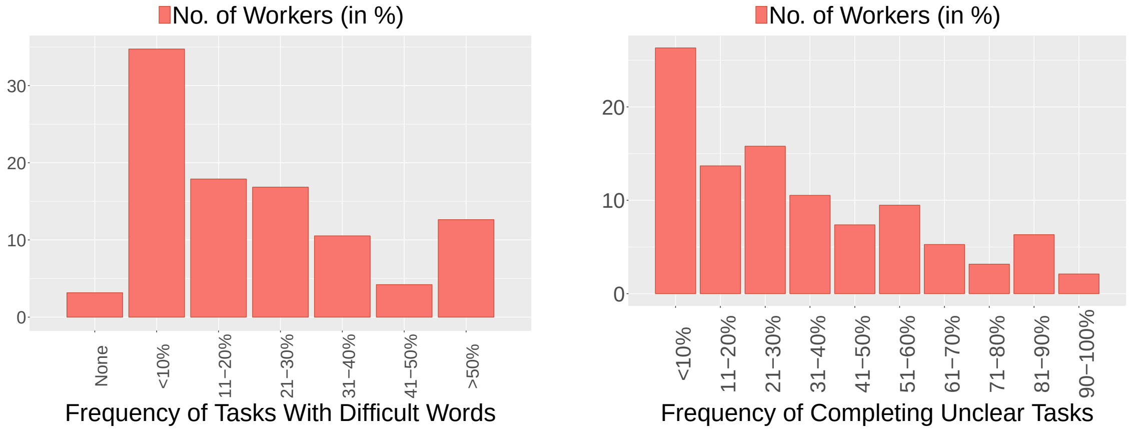
On the other hand, another 27% of workers completed more than 50% of all the unclear tasks they come across. In addition, around 18% of workers used dictionaries or other helpful means/tools to better understand over 50% of tasks they completed. 20% of workers used translators in more than 50% of the tasks that they completed.

6.4 Modeling Task Clarity

We address **RQ2** by modelling task clarity of crowdsourced microtasks as a combination of *goal clarity* and *role clarity*. Inspired by previous work in organizational psychology [RN77], we define task clarity as a combination of the extent to which the desired outcome of a task is clear (goal clarity), and the extent to which the workflow or activities to be carried out are clear (role clarity).

6.4.1 Assessing Task Clarity

Task clarity of microtasks in a marketplace is a notion that can be quantified by human assessors by examining task metadata such as the *title*, *keywords* associated with the task, *instructions* and *description*. Since these are the main attributes that



(a) (b) Information Finding Tasks

Figure 6.2 (a) Frequency of tasks with difficult words, and (b) frequency of workers completing unclear tasks.

requesters use to communicate the desired outcomes of the tasks, and prescribe how crowd workers should proceed in order to realize the objectives, we argue that they play an important role in shaping crowd work.

6.4.2 Acquiring Task Clarity Labels

With an aim to understand the distribution of task clarity across the diverse landscape of tasks on AMT [DCD+15b], we sampled 7,100 tasks that were deployed on AMT over a period of 1 year between October 2013 to September 2014. For every month spanning the year, we randomly sampled 100 tasks of each of the 6 task types proposed in previous work [GKD14d]; *content creation* (CC), *information finding* (IF), *interpretation and analysis* (IA), *verification and validation* (VV), *content access* (CA)[5] and *surveys* (SU). Next, we deployed a job[6] on CrowdFlower to acquire task clarity labels from crowd workers. We first provided detailed instructions describing task clarity, goal clarity and role clarity. An excerpt from the task overview is presented below:

"Task clarity defines the quality of a task in terms of its comprehensibility. It is a combination of two aspects; (i) goal clarity, i.e, the extent to which the objective of the task is clear, and (ii) role clarity, i.e., the extent to which the steps or activities to be carried out in the task are clear."

In each task workers were required to answer 10 questions on a 5-point Likert scale. The questions involved assessing the goal and role clarity of the corresponding

[5] Note that there were fewer than 100 tasks of the CA type in a few months during the time period considered. In those cases, we considered all available tasks.

[6] Preview available in the companion page: https://sites.google.com/site/ht2017clarity/.

task, the overall task clarity, the influence of goal and role clarity in assessing overall task clarity, clarity of title, instructions and description, the extent to which the title conveyed the task description, the extent to which the keywords conveyed the task description and goal of the task, and the quality of language in the task description. Apart from these 10 questions, workers were provided with an optional text field where they could enter comments or remarks about the AMT task they evaluated. We gathered responses to these questions for each of the 7,100 tasks from 5 distinct crowd workers. We controlled for quality by using the *highest quality* restriction on CrowdFlower, that allows only workers with a near perfect accuracy over hundreds of different tasks and varying task types. In addition, we interspersed attention check questions where workers were asked to enter alphanumeric codes that were displayed to them. Workers were compensated according to the hourly rate of 7.50 USD.

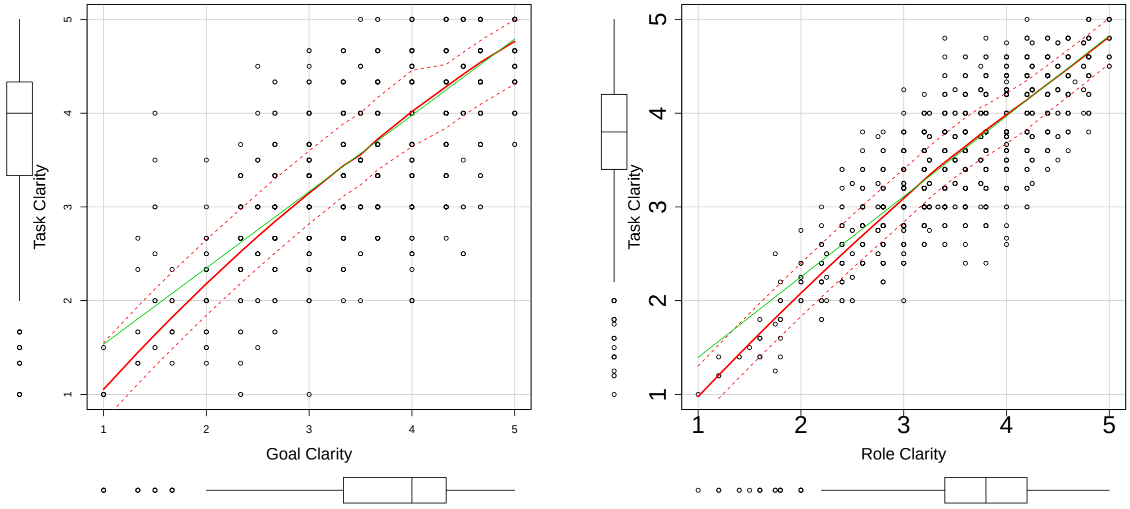
6.4.3 Perception of Task Clarity

We found that the mean task, goal and role clarity across the different tasks were nearly the same. On average, workers perceived tasks to be moderately clear (*M=3.77, SD=.53*). The same is the case with goal clarity (*M=3.76, SD=.53*) and role clarity (*M=3.76, SD=0.54*). On investigating the influence of goal and role clarity on the crowd workers in adjudicating the overall task clarity, we found that role clarity and goal clarity were both important in determining task clarity. On average, workers responded that goal clarity influenced their judgment of overall task clarity to an extent of 3.98/5 (*SD=.51*), and that in case of role clarity was 3.93/5 (*SD=.52*). We found that goal clarity was slightly more influential than role clarity in determining the task clarity, and this difference was statistically significant; $t(14199) = 25.28, p < .001$. We also analyzed the relationship of task clarity with goal and role clarity respectively. We found strong positive linear relationships in both cases, as shown in Figure 6.3.

We computed Pearson's **r** between task clarity with each of goal and role clarity; $r(14998) = .85, R^2 = .72, p < .001$ and $r(14998) = .86, R^2 = .74, p < .001$. These findings indicate that it is equally important for task requesters to ensure that the objective of the task, as well as the means to achieve the desired outcome are adequately communicated to crowd workers via the task title, instructions and description, and keywords associated with the task.

Inter-worker Agreement

To find out whether or not task clarity is coherently perceived by workers, we verify the presence of agreement of task clarity evaluations among workers. Given the subjective nature of task clarity evaluations, we apply the SOS Hypothesis [HSE11], which examines the extent to which individual evaluations of clarity spread around the mean clarity value per task. The SOS Hypothesis has proven to be more reliable than



(a) (b)

Figure 6.3 Relationship of Task Clarity with (a) Goal Clarity, and (b) Role Clarity. The trendline is represented in green, and the regression line is represented by the thick red line.

other inter-evaluator agreement measures such as Krippendorff's alpha, in subjective assessment tasks that involve a set of participants evaluating the same item – in our case, the same task [AMN14]. In SOS Hypothesis, we test the magnitude of the squared relationship between the standard deviation (i.e. SOS) of the evaluations and the mean opinion score (MOS; in our case, mean clarity score), denoted by α. The value of α can then be compared with those of other subjective assessment tasks that are deemed to be more (high α) or less prone to disagreement (low α) among evaluators. Specifically, for 5-point scale evaluations, SOS Hypothesis tests a square relationship between SOS and MOS by fitting the following equation:

$$SOS(i) = -\alpha MOS(i)^2 + 6\alpha MOS(i) - 5\alpha$$

considering each task i in the evaluation pool.

Table 6.1 SOS Hypothesis α values for Task Clarity, Goal Clarity and Role Clarity.

Clarity	Task Clarity	Goal Clarity	Role Clarity
α	0.3166	0.3229	0.3184

Table 6.1 shows the α values computed for task clarity, goal clarity and role clarity. All these evaluations have a value of 0.32, which is similar to what could be obtained

in other subjective assessment tasks such as smoothness of web surfing, VoIP quality, and cloud gaming quality [HSE11]. We therefore consider it acceptable. Figure 6.4 shows the fitted quadratic curve against worker evaluations for individual tasks. A significant correlation could be obtained between the fitted SOS value and the actual SOS value (Pearson's $r = 0.506$, $p < .001$). In conclusion, we find that task clarity is coherently perceived by workers. The substantial evidence of workers' agreement in perceiving task clarity helps establish the mean clarity score as ground truth for modeling task clarity using objective task features, as we report in Section 6.5.

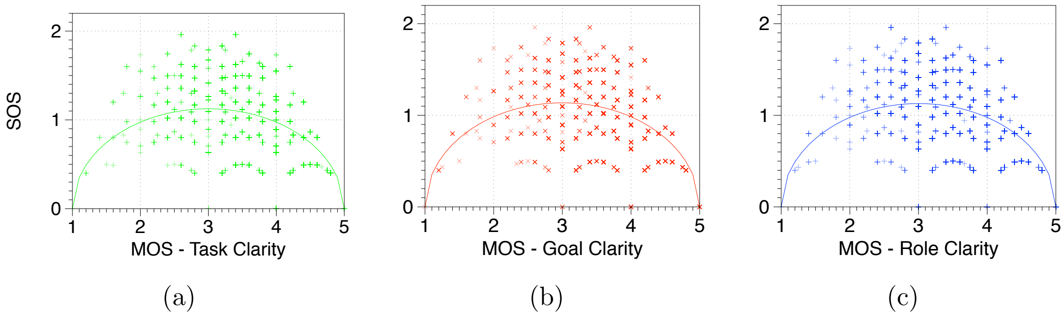


Figure 6.4 SOS Hypothesis plots for Task Clarity (green), Goal Clarity (red), and Role Clarity (blue). The quadratic curve depicts the fitting to worker evaluations for individual tasks.

Task Types and Perception of Task Clarity

We investigated the impact of task types on the perception of task clarity and the constructs of goal and role clarity. We note that Levene's test for homogeneity of variances was not violated across the different task types with respect to each of task, goal and role clarity. We conducted a one-way between workers ANOVAs to compare the effect of task types on the perception of task, goal and role clarity respectively. We found a significant effect of task type on the perception of task clarity at the $p < .01$ level, across the 6 task type conditions; $F(5, 7002) = 6.176, p < .001$. Post-hoc comparisons using the Tukey HSD test indicated that the perception of task clarity in some task types was significantly poorer than others; as presented in Table 6.2.

We also found a significant effect of task type on (i) the perception of goal clarity at the $p < .01$ level, across the 6 task type conditions; $F(5, 7002) = 5.918, p < .001$, and (ii) the perception of role clarity at the $p < .01$ level, across the 6 task type conditions; $F(5, 7002) = 8.074, p < .001$. Post-hoc comparisons using the Tukey HSD test (Tables 6.3 and 6.4) indicated that the perception of goal and role clarity in some task types was significantly poorer than others.

Table 6.2 Post-hoc comparisons using the Tukey HSD test to investigate the effect of task types on *task clarity*. Comparisons resulting in significant outcomes are presented here. (* indicates $p < .05$ and ** indicates $p < .01$)

Task Type	M	SD	Comparison	Tukey HSD p-value
CA	3.75	.51	*CA vs SU*	0.011^{*}
CC	3.76	.51	*CA vs VV*	0.004^{**}
IA	3.74	.52	*CC vs SU*	0.046^{*}
IF	3.77	.52	*CC vs VV*	0.020^{*}
SU	3.82	.50	*IA vs SU*	0.001^{**}
VV	3.82	.48	*IA vs VV*	0.001^{**}

Table 6.3 Post-hoc comparisons using the Tukey HSD test to investigate the effect of task types on *goal clarity*. Comparisons resulting in significant outcomes are presented here.
(* indicates $p < .05$ and ** indicates $p < .01$)

Task Type	M	SD	Comparison	Tukey HSD p-value
CA	3.76	0.52	*CA vs VV*	0.006^{**}
CC	3.76	0.50	*CC vs VV*	0.004^{**}
IA	3.74	0.51	*IA vs SU*	0.005^{**}
IF	3.78	0.52	*IA vs VV*	0.001^{**}
SU	3.82	0.51		
VV	3.83	0.50		

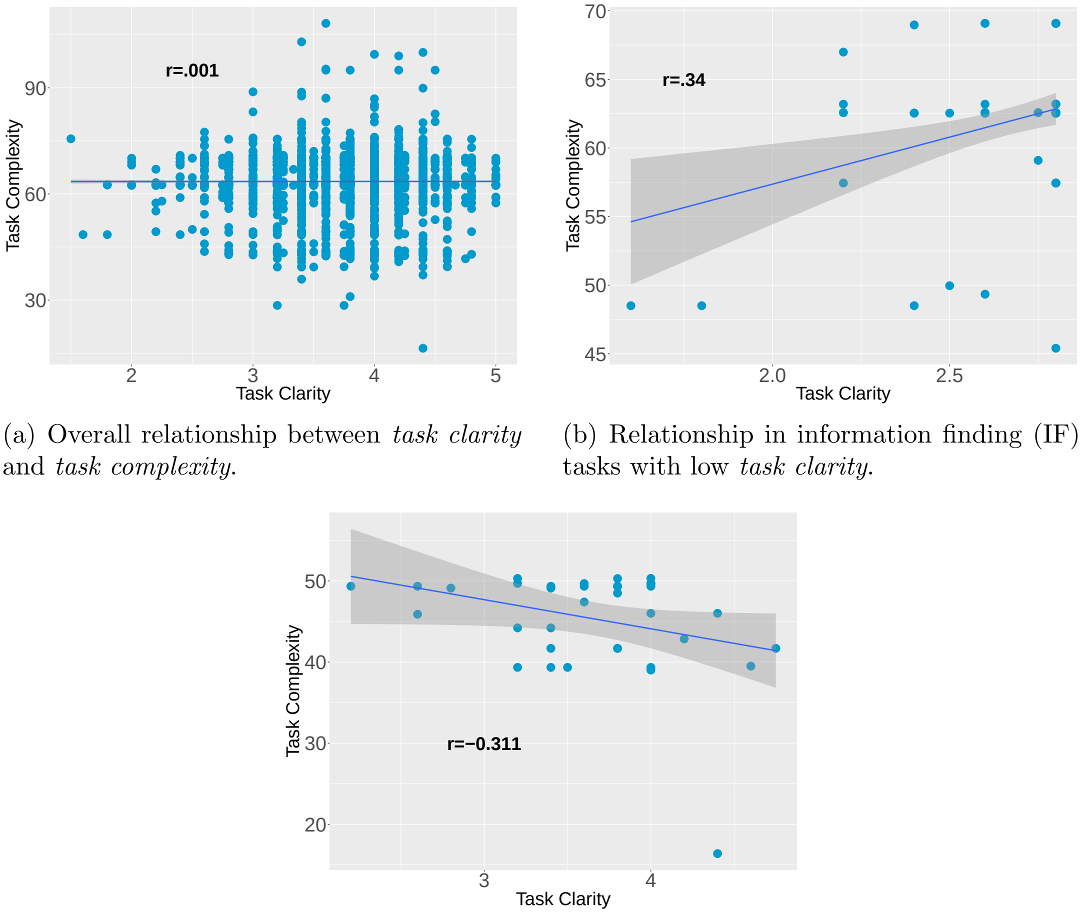
6.4.4 Task Clarity and Task Complexity

Recent work by Yang et al. modeled *task complexity* in crowdsourcing microtasks [YRDB16]. By using the task complexity predictor proposed by the authors, we explored the relationship between task clarity and task complexity. We found no significant correlation between the two variables across the different types of 7,100 tasks in our dataset (see Figure 6.5(a)). This absence of a linear relationship between task complexity and task clarity suggests that tasks with high clarity can still be highly complex or tasks with low clarity can have low task complexity at the same time.

We analyzed the relationship between *task clarity* and *complexity* across different types of tasks, and found that there is no observable correlation between the two variables across the different types of tasks. As can be observed from Figure 6.3, a majority of tasks are perceived to lie within the range of moderate to high clarity. We therefore further investigated tasks with low clarity or complexity.

Relationship in Tasks with Low Clarity

As shown earlier, task clarity was coherently perceived by workers. We reason that tasks corresponding to a clarity rating < 3 have relatively *low* clarity. We investigated the effect of task types on the relationship between task clarity and complexity



(a) Overall relationship between *task clarity* and *task complexity*.

(b) Relationship in information finding (IF) tasks with low *task clarity*.

(c) Relationship between *task clarity* and *complexity*.

Figure 6.5 SOS Hypothesis plots for Task Clarity (green), Goal Clarity (red), and Role Clarity (blue). The quadratic curve depicts the fitting to worker evaluations for individual tasks.

Table 6.4 Post-hoc comparisons using the Tukey HSD test to investigate the effect of task types on *role clarity*. Comparisons resulting in significant outcomes are presented here.
(* indicates $p < .05$ and ** indicates $p < .01$)

Task Type	M	SD	Comparison	Tukey HSD p-value
CA	3.75	0.51	*CA vs SU*	0.030^{*}
CC	3.76	0.50	*CA vs VV*	0.001^{**}
IA	3.73	0.52	*CC vs SU*	0.048^{*}
IF	3.78	0.51	*CC vs VV*	0.001^{**}
SU	3.82	0.50	*IA vs SU*	0.001^{**}
VV	3.84	0.48	*IA vs VV*	0.001^{**}
			IF vs VV	0.043^{*}

in tasks with low clarity. Using Pearson's **r**, we found a weak positive linear relationship between the two variables in information finding (IF) tasks with low clarity (see Figure 6.5(b)); *N=80*, **r**=*.34*. This can be explained as a consequence of complex workflows required to complete some IF tasks, where high task complexity is concomitant with relatively high task clarity. Accordingly, in IF tasks with low clarity, task complexity accounted for 11.56% of the variance in task clarity (the coefficient of determination, R^2*=.1156, p<.01*). We did not find a significant relationship between the two variables in the low clarity subsets of other task types.

Relationship in Tasks with Low Complexity

Similarly, we consider tasks having a complexity score < 50 have relatively *low* complexity. We investigated the effect of task types on the relationship between task clarity and complexity in tasks with low complexity. Using Pearson's **r**, we found a weak negative linear relationship between the two variables in content access (CA) tasks with low complexity (see Figure 6.5(c)); *N=41*, **r**=*.311*. Thus, in CA tasks with low complexity, task clarity accounted for 9.67% of the variance in task complexity (the coefficient of determination, R^2*=.0967, p<.05*).

Discussion

The lack of linear correlation between clarity and complexity yields interesting observations. While surprising (intuitively, one might assume that a better task formulation – high clarity – would yield a lower complexity), this result is aligned with the classical theory on cognitive load, by Sweller and Chandler [SC94]. The theory postulates the presence of two sources of cognitive load: *intrinsic* and *extraneous*. Intrinsic cognitive load refers to the inherent difficulty in the content of presented material, which approximates task complexity in our context; extraneous cognitive load, on the other hand, refers to the organization and presentation of material, i.e. task clarity in our context. Sweller and Chandler suggest in their theory that, while the intrinsic cognitive load is unalterable, the extraneous cognitive load can either be increased because of inappropriate instructional design, or be reduced by well-structured presentation. We show that the theory can find application in microtask crowdsourcing, as tasks of similar complexity can either be of high clarity or low clarity.

When considering tasks of specific types, however, we find correlation could be established. Specifically, we find a negative correlation with content access (CA) tasks, thus suggesting that (poorly formulated) tasks asking workers to interact with on line content (e.g. watch a video, click a link) can be perceived more complex to execute. With information finding (IF) tasks, high task complexity maps with high clarity, thus suggesting that requests for complex finding and retrieval operations can be associated with clearer instructions. These results provide further insights into the relationship between task clarity and complexity, and call for further investigation.

6.5 Prediction of Task Clarity

In this section we tackle **RQ3** and propose to model task clarity based on objective features that are extractable from tasks. We envision a system that could automatically predict task clarity and thus provides feedback to requesters on task design and to workers on task selection and execution. To test the feasibility of this idea, our study starts by designing task features that are potentially predictive for task clarity; we then build a predictive model to automatically learn task clarity based on these features.

6.5.1 Features Sets

We explore four classes of task features, namely: *metadata features*, *task type features*, *content features*, and *readability features*. In the following we provide a brief introduction to each feature class, and refer the readers to the companion page for a full description of the feature set.

Metadata Features are the task attributes associated with the definition of tasks when they are created. Typical metadata features include the number of `initial HITs`, attributes of the descriptions about desired activities to be performed by workers (e.g., `title length` and `description length`), the required qualification of workers (e.g., worker `location` and minimum `approval rate`), the estimated execution time (i.e. `allotted time`) and `reward`. These features characterize a task from different aspects that might be correlated with task clarity. For example, we assume that a longer description could entail more efforts from the requester in explaining the task.

Task Type Features categorize a task into one of the six task types defined by [GKD14d]. They are therefore high level features that comprehensively describe what knowledge is in demand. Through previous analysis, we have observed that task type has a significant effect on the perception of task clarity. We therefore assume that task type could be indicative of task clarity in prediction.

Content Features capture the semantics of a task. These features use the high-dimensional bag of words (BOW) representation. To maximize the informativeness of the content features while minimizing the amount of noise, one-hot (i.e. binary) coding was applied to the BOW feature of task title and keywords, while TF-IDF weighting was applied to the BOW feature of task description. It has been shown by research in related domains (e.g., community Q&A systems [YHBH14]) that the use of words is indicative of the quality of task formulation, therefore we are interested in understanding the effect of language use on workers' perception of task clarity.

Readability Features are by nature correlated with task clarity: tasks with higher readability are better formulated, and are thereby expected to have a higher clarity. We experiment with several widely used readability metrics in our clarity prediction task to understand their predictive power of task clarity. These include the use of long words (`long_words`), long sentence (`words_per_sentence`), the use of `preposition`, `nominalization`, and more comprehensive readability metrics such as `ARI`, `LIX`, and in particular, `Coleman_Liau`, which approximates the U.S. grade level necessary to comprehend a piece of text.

6.5.2 Prediction Results

Due to the high dimension of the content features (size of vocabulary = 10,879), we apply the Lasso method, which does feature selection and regression simultaneously. We adopt 5-fold cross-validation and mean absolute error (MAE) for evaluation. Table 6.5 shows the prediction results. The prediction on task clarity achieves a MAE of 0.4032 ($SD = 0.0031$). The relatively small error compared to the scale of ground truth (i.e. 1-5) indicates that task clarity can be predicted accurately. In addition, the small standard deviation shows that the prediction is robust across different tasks. Similar results also hold for the prediction of goal clarity and role clarity, which confirms our previous observation that both are highly correlated with the overall task clarity.

Table 6.5 Prediction results for Task Clarity, Goal Clarity and Role Clarity, shown by $\mu \pm \sigma$.

Clarity	Task Clarity	Goal Clarity	Role Clarity
MAE	0.4032±0.0031	0.4076±0.0067	0.4008± 0.0070

Predictive Features. In the following we analyze the predictive features selected by Lasso. Table 6.6 shows the features with positive and negative coefficients in the Lasso model after training for task clarity prediction, i.e. features that are positively

Table 6.6 Predictive features for task clarity prediction.

Feat. Class	Feat. w. Positive Coef.		Feat. w. Negative Coef.	
	Feature	Coef.*	Feature	Coef.*
Metadata	`number_keywords`	0.719	`external_links`	-0.598
	`description_length`	0.295		
	`number_images`	0.071		
	`total_approved`	0.011		
Task Type	`VV`	0.434	`IA`	-0.922
	`SU`	0.413		
Content	keyword: `audio`	2.673	keyword: `id`	-2.658
	keyword: `transcription`	1.548		
	keyword: `survey`	1.178		
Readability	`preposition`	1.748	`ARI`	-1.982
	`GunningFogIndex`	1.467	`long_words`	-0.671
	`Coleman_Liau`	0.855	`syllables`	-0.478
	`words_per_sentence`	0.620	`nominalization`	-0.136
	`characters`	0.237	`pronoun`	-0.104
	`LIX`	0.150	`FleschReadingEase`	-0.075
			`RIX`	-0.038
	(all about title)		(all about title)	

* For the sake of comparison, each value is shown with original coefficient $\times 10^2$.

and negatively correlated with task clarity. Similar observations can be obtained for predicting goal and role clarity.

With regard to metadata features, it can be observed that longer descriptions and more keywords are positively correlated with task clarity. This suggests that more description and keywords could potentially improve the clarity of task formulation. We also observe that the increased use of images, or less use of external links could enhance task clarity. These are reasonable, since intuitively, images can help illustrate task requirements, while external links would bring in extra ambiguity to task specification in the absence of detailed explanations.

With regard to task type features, we find that tasks of type SU and VV are in general of higher clarity, while tasks of type IA are of lower clarity. This result confirms our previous findings.

With regard to content features, we observe that keyword features are more predictive than other types of content features (e.g. words in title or description). Predictive keywords include `audio`, `transcription`, `survey`, etc., which can actually characterize the majority of tasks in AMT. We therefore reason that workers' familiarity with similar tasks could enhance their perception of task clarity.

Finally, several interesting findings with regard to task readability are observed as follows. First, many types of readability scores are indicative of task clarity, indicating a strong correlation between task readability and task clarity. Second, compared with description or keyword readability, title readability is most predictive of task clarity. As an implication for requesters, putting efforts in designing better titles can improve task clarity. Third, we observe a positive correlation between task clarity and Coleman_Liau, which approximates the U.S. grade level necessary to comprehend the text. The increase of Coleman_Liau (i.e. more requirements on workers capability to comprehend the title) therefore does not lead to lower task clarity perceived by workers. The result is not surprising, given the demographic statistics of crowdworkers [DCD+15b]. However, it raises questions on the suitability of this class of microtask crowdsourcing tasks for other types of working population.

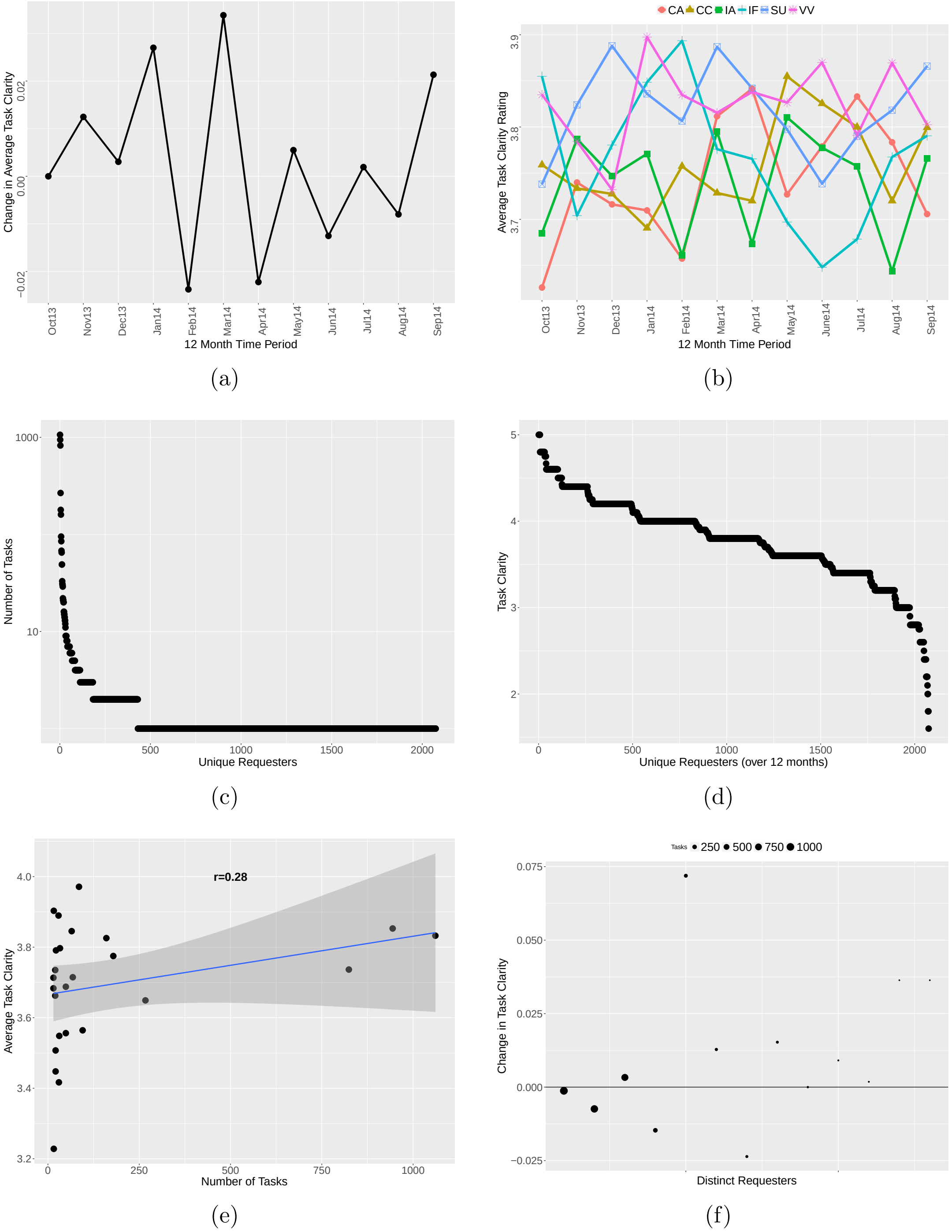
On decomposing Coleman_Liau and exploring the effect of length of words (in terms of #letters) and length of sentences (in terms of #words), it can be observed that longer words (i.e., long_words) would decrease task clarity, while longer sentence (i.e., words_per_sentence) can enhance task clarity. This suggests that workers can generally comprehend long sentences, while the use of long words would decrease task clarity. This is consistent with our findings from **RQ1**, where workers identified difficult words as a factor that decreased task clarity and also suggested that tasks with difficult words are commonplace in the microtask crowdsourcing market. We also found a positive correlation between preposition and task clarity, in contrast to the negative correlation between syllables (or nominalization) and task clarity. These results suggest that partitioning sentences with prepositions could increase task clarity, while complicating individual words decreases task clarity.

6.6 Evolution of Task Clarity

6.6.1 Role of Task Types

To address **RQ4**, we investigated the evolution of task clarity over time (see Figure 6.6). We found that there was no monotonous trend in the overall average task clarity over time, as shown in Figure 6.6(a). We also investigated the effect of task type on the evolution of task clarity. We found no discernible trend in the evolution of task clarity of different types of tasks over the 12 month period considered in the dataset (Figure 6.6(b)). We conducted a one-way ANOVA to compare the effect of task type on the evolution of task clarity over time. We did not find a significant effect of task type on the evolution of task clarity at the $p < .05$ level, across the 6 task type conditions; $F(5, 66) = 0.081, p = .994$.

These findings suggest that the overall task clarity in the marketplace varies over time but does not follow a clear pattern. This can be attributed to the organic influx of new task requesters every month [DCD+15b]. To identify whether the experience of task requesters plays a role in the evolution of task clarity, i.e., whether individual requesters deploy tasks with increasing task clarity over time we investigated the role of requesters in the evolution of task clarity.



(a) (b)

(c) (d)

(e) (f)

Figure 6.6 (a) Evolution of overall task clarity and (b) with respect to different types of tasks from Oct'13-Sep'14, (c) distribution of tasks corresponding to requesters who deployed them, (d) distribution of the average task clarity of tasks corresponding to distinct requesters across the 12 months, (e) relationship between the average task clarity and the number of tasks deployed by experienced requesters, (f) $\Delta TaskClarity$ of requesters who deployed tasks during more than 6/12 months in our dataset.

6.6.2 Role of Requesters

Recent analysis of the AMT marketplace, revealed that there is an organic growth in the number of active requesters and a constant growth in the number of new requesters (at the rate of 1,000 new requesters per month) on the platform [DCD+15b]. Poor task design leading to a lack of task clarity can be attributed to inexperienced requesters. To assess the role of requesters in the evolution of task clarity, we analyzed the evolution of task clarity of different types of tasks with respect to individual requesters.

We analyzed the distribution of unique requesters corresponding to the 7.1K tasks in our dataset. We found that a few requesters deployed a large portion of tasks, as depicted by the power law relationship in Figure 6.6(c). We also found that over 40% of the requesters exhibited an overall average task clarity of $\geq 4/5$, and in case of nearly 75% of the requesters it was found to be over 3.5/5 (as presented in Figure 6.6(d)). We considered requesters who deployed ≥ 15 tasks within the 12-month period as being experienced requesters, and analyzed the relationship between the number of tasks they deployed with the corresponding overall task clarity. Using Pearson's **r**, we found a weak positive correlation between the average task clarity and the number of tasks deployed by experienced requesters (see Figure 6.6(e)); **r**= .28. Thus, the experience of requesters (i.e., the number of tasks deployed) explains over 8% of the variance in the average task clarity of tasks deployed by the corresponding requesters; the coefficient of determination, $R^2 = 0.081$.

Considering the requesters who deployed tasks during more than 6 months in the 12-month period, we investigated the overall change in terms of average task clarity of the tasks deployed from one month to the next. We measure the overall change in task clarity for each requester using the following equation.

$$\Delta TaskClarity_r = \frac{1}{n} \sum_{i=1}^{n} (TC_{i+1} - TC_i)$$

where, TC_i represents the average task clarity of tasks deployed by a requester in the month i, n is the total number of months during which requester r deployed tasks.

Figure 6.6(f) presents our findings with respect to the overall change in task clarity corresponding to such requesters. The size of the points representing each requester depict the number of tasks deployed by that requester. We did not find a significant effect of the number of tasks deployed by requesters on the magnitude of change in task clarity.

Based on our findings, we understand that the overall task clarity in the marketplace fluctuates over time. We found a weak positive linear relationship between the number of tasks deployed by individual task requesters and the associated task clarity over time. However, we did not find evidence that the magnitude of change in task clarity is always positive in case of experienced requesters.

6.6.3 Top Requesters

We note that the top-3 task requesters accounted for around 67% of the tasks that were deployed between Oct'13 to Sep'14. The requesters were found to be *SpeechInk*–1,061 tasks, *AdTagger*–944 tasks, and *CastingWords*–824 tasks. The evolution of task clarity of the tasks corresponding to these requesters over time is presented in the Figure 6.7 below.

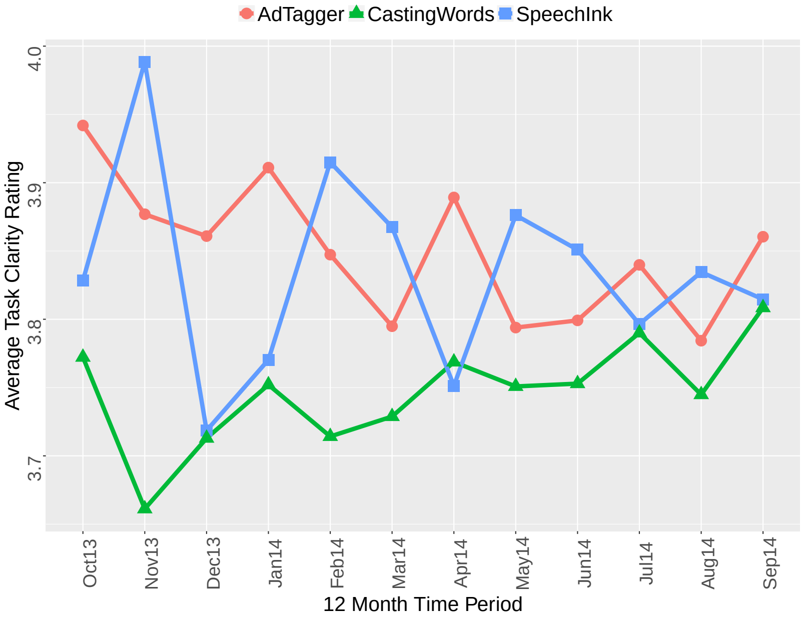


Figure 6.7 Top-3 task requesters w.r.t. the number of tasks deployed, and the evolution of their task clarity over time.

To understand the effect of the task requesters on the evolution of task clarity over time, we conducted a one-way between requesters ANOVA. We found a significant effect of task requesters on the evolution of task clarity across the three different requester conditions (SpeechInk, AdTagger, CastingWords) over the 12-month period; $F(2, 33) = 11.837$, $p < .001$. Post-hoc comparisons using the Tukey HSD test revealed that the evolution of task clarity corresponding to tasks from *SpeechInk* and *AdTagger* were significantly different in comparison to *CastingWords*.

We observe a gradual increase in the task clarity of *CastingWords* tasks over time in contrast to the other two top requesters. In the context of these requesters and the time period of Oct'13-Sep'14, we explored the Turkopticon ratings [IS13] corresponding to the requesters. Turkopticon collects ratings from workers on the following qualities : *fairness* of a requester in approving/rejecting work, *communicativity*– the responsiveness of a requester when contacted, *generosity*– quality of pay with respect to the amount of time required for task completion, *promptness* of the requester in approving work and paying the workers. Figure 6.8 presents a comparison of the

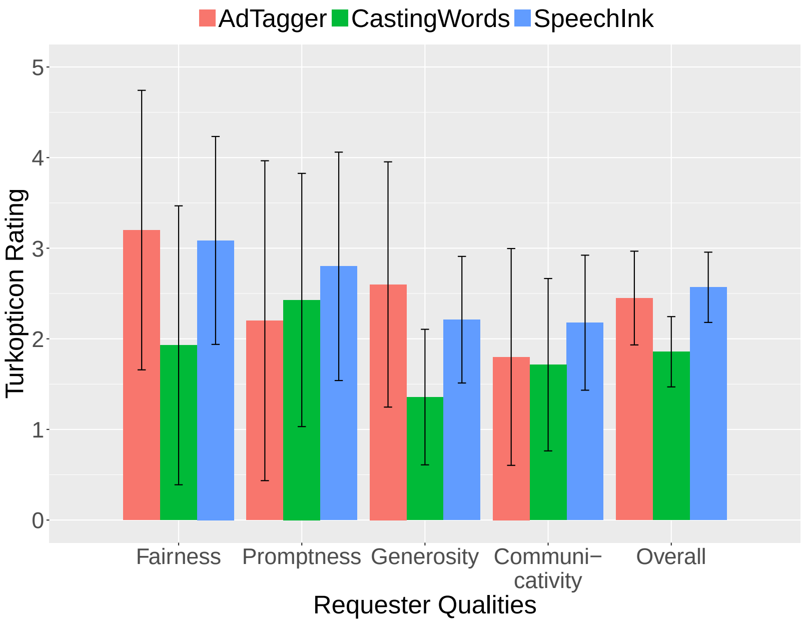


Figure 6.8 Average Turkopticon ratings of the top requesters from Oct'13–Sep'14.

Turkopticon ratings of the 3 requesters for each of the four qualities. We note that *SpeechInk* received consistently better ratings across all qualities within the given period. This coincides with the relatively higher task clarity of *SpeechInk* (*M=3.83, SD=0.47*) tasks when compared to *CastingWords* (*M=3.73, SD=0.48*) tasks over the 12 months (see Figure 6.7). A two-tailed T-test revealed a significant difference in the task clarity between *SpeechInk* and *CastingWords*; *t(1883)=18.43, p* < .001. We did not find ratings of tasks deployed by *AdTagger* on Turkopticon during the time period considered. However, we present a comparison based on the ratings received by *AdTagger* prior to Oct'13. Once again, in comparison to *CastingWords* we note that the higher overall quality ratings of *AdTagger* on Turkopticon coincide with the higher task clarity over the 12 months (*M=3.85, SD=0.48*); *t(1766)=25.23, p* < .001.

Through our findings it is clear that task clarity is not a global, but a local property of the AMT marketplace. It is influenced by the actors in the marketplace (i.e., tasks, requesters and workers) and fluctuates with the changing market dynamics.

6.7 Chapter Summary

In this chapter we examined *task clarity*, an important, yet largely neglected aspect of crowdsourced microtasks. By surveying 100 workers, we found that workers confront unclear tasks on a regular basis. They deal with such tasks by either exerting extra effort to overcome the suboptimal clarity, or by executing them without a clear understanding. Poor task formulation thereby greatly hinders the progress of workers' in obtaining rewards, and in building up a good reputation.

To better understand how clarity is perceived by workers, we collected workers' assessments for 7.1K tasks sampled from a 5 years worth dataset of the AMT marketplace. With an extensive study we revealed that clarity is coherently perceived by workers, and that it varies by the task type. In addition, we found compelling evidence about the lack of direct correlation between clarity and complexity, showing the presence of a complex relationship that requires further investigation. We proposed a supervised machine learning model to predict task clarity and showed that clarity can be accurately predicted. We found that workers' perception of task clarity is most influenced by the number of keywords and title readability. Finally, through temporal analysis, we show that clarity is not a macro-property of the AMT ecosystem, but rather a local property influenced by tasks and requesters.

In conclusion, we demonstrated the importance of clarity as an explicit property of microwork crowdsourcing tasks, we proposed an automatic way to measure it, and we unveiled interesting relationships (or lack thereof) with syntactical and cognitive properties of tasks. Our findings enrich the current understanding of crowd work and bear important implications on structuring workflow. Predicting task clarity can assist workers in task selection and guide requesters in task design.

7
The Role of Work Environments in Microtask Crowdsourcing

> *"Sometimes the Internet fee is greater than the rewards I earn (due) to images, audios or videos in tasks."*
>
> *– CrowdFlower Worker from India*

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

7.1 Introduction

We are currently in an age of pervasive computing where various kinds of sensors facilitate *smart* environments at home or work, improving our lives in numerous ways [ZBC+14]; ranging from optimizing energy consumption [MA13] to facilitating structural health monitoring [LL06]. Recent work has showcased how visual sensors (in the form of CCTV cameras) and social sensors (such as Twitter feeds) can be combined to improve event detection and aid in understanding the evolution of situations [WK15]. The rapid growth and ubiquity of mobile devices has resulted in making participatory sensing feasible on a large scale [BEH+06]. Such opportunities of using people as sources of sensory information allows us to build useful applications that have implications on urban mobility [XW15], environment [Kan11], personal health monitoring [RPH+07], and so forth. An effective way to use people as sources of sensory information is to collect data from them directly [KRF+17]. Online marketplaces like Amazon's Mechanical Turk[1] (AMT) or CrowdFlower[2] provide a large and diverse workforce of people accessible around the clock and on demand, for participation in return for monetary rewards.

Microtask crowdsourcing is being used widely these days to solve a multitude of problems that go beyond the capability of machine intelligence. Over time, crowdsourcing platforms like AMT have been used for diverse purposes including content creation and running surveys [DCD+15b]. The potential to reach thousands of people around the globe at will and on demand [Ipe10b] has led to innovative and unprecedented solutions to problems in the space of ubiquitous and pervasive computing [HAB16, SSK+16, MOR+16]. On the other hand, the ubiquity of the Internet and rise in prevalence of electronic devices have led to the rise of applications centered on mobile crowdsourcing [Eag09, YMH+09, FZZ+14, GTCB12, NGR+11]. Recent work by Laput et al. that introduced *Zensors*, leverages real-time human intelligence from online crowd workers to create robust, adaptive and intelligent sensors for any visually observable property [LLW+15]. This is a great example of how human input and intelligence can play a pivotal role in pervasive computing.

An important aspect at the core of crowd-powered solutions (especially those based on the microtask crowdsourcing paradigm) is controlling the quality of work that is produced by workers. Several research works have focused on improving the quality of crowdsourcing results by using a variety of techniques ranging from worker pre-screening methods and effective crowdsourcing task design [KNB+13], to using gamification and incentive mechanisms [FSVKS15b, RZS16], and answer aggregation methods [VGK+14b]. Due to the low entry barrier, crowdsourcing has become truly ubiquitous [VKG10]. Prior ethnographic works have also shown that workers who participate in microtask crowdsourcing are embedded in diverse environmental contexts that impact their work routines [GMHO14, MHOG14]. A considerable

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

[2] http://www.crowdflower.com

amount of research effort has focused on the *motivations* behind worker participation in crowdsourced microtasks [KSV11b, CK13]. Yet, little is understood about the *tools* that support and drive such worker participation. While some previous works have focused on the question of *why* crowd workers spend time and complete tasks on crowdsourcing platforms, in this work we focus on the question of *how* crowd workers complete tasks, and investigate the affect of different work environments on worker performance.

In this chapter we draw attention to the less addressed realm of the *modus operandi* of workers in crowdsourcing microtasks. How exactly do workers participate and contribute to crowd work? We are particularly interested in the work environments that influence the observable behavior of crowd workers. We aim to understand how the different *work environments*, defined and characterized by the hardware and software that enable worker participation in crowd work, affect the quality of work that is produced. With the rampant increase in mobile device use around the world[3], do crowd workers use mobile devices more commonly than laptops and PCs? Are some devices more suitable to specific task types than others? How up-to-date are crowd workers with the available software and hardware? How do UI elements and work environments interact while facilitating crowd work? By '*suitability*', we refer to the quality of being appropriate in supporting workers complete a given task at hand. For example, a work environment that is characterized by low Internet connection bandwidth may impede workers in completing tasks that contain high resolution images or videos, as opposed to a work environment that is characterized by a high Internet connection bandwidth. Similarly, some devices may be more appropriate in supporting workers complete tasks. For example, desktops or laptops are intuitively more appropriate for tasks that involve searching the web and retrieving information than mobile phones. Even in a novel system such as *WearWrite* [NTG+16], that enables users to write documents from their smartwatches, crowd workers who tested the system found issues with the interface such as its complexity and language. To elucidate, one user said "I don't know what a bullet point is", with one of the recommendations being better support in providing a cross-device functionality so workers can work on smartwatches, mobile phones or desktop computers. Understanding how crowd workers operate during their participation in microtasks can help us in the following ways:

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

To this end, we carried out three studies. In the first study, we gathered responses from 100 distinct workers on CrowdFlower, in a survey regarding common issues with

[3] Forecasted to reach 11.6 billion mobile devices by 2020, exceeding the world's projected population at that time (7.8 billion) [cis].

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

7.2 Related Literature

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

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

7.2.1 Crowdsourcing for Ubiquitous & Pervasive Computing

Over the last decade, many parallels have been drawn between crowdworkers and artificial intelligence systems; crowdsourcing even being called artificial artificial intelligence [Ira15], making the worker and their work invisible, an issue that the field of 'Social Computing' has been trying to tackle over the years.

Prior work in participatory sensing [SD10, DWW^{+}14] have identified issues with usability in the design of crowd applications that could lead to significant data quality issues. For instance, Ding et al. have studied the issues with using crowd-based personal spectrum sensors (such as through smartphones and in-vehicle sensors) where the sensing data may have been unreliable or untrustworthy due to unexpected equipment failures or malicious behaviors, amounting to some abnormal data, and, making crowd-sensing schemes ineffective [DWW^{+}14]. They discussed reactive measures to robustly cleanse out abnormal data components from the original corrupted sensing data. Hara et al. identified that heterogeneous devices come to be involved in crowdsourcing environments [HSBS13]. While finding generic infrastructures can be very difficult, they think having a generic reusable platform to support development of crowdsourcing applications would help in supporting heterogeneous mobile devices

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

as well as manage large numbers of users. Chatzimilioudis et al. discuss multiple crowdsourcing applications by the use of smartphones [CKLZY12]. They deliberate on the issues of running crowdsourcing applications with smartphones such as different Internet connection modalities (2G, 3G, 4G) each with different energy and data transfer rates.

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

7.2.2 Task Types in Microtask Crowdsourcing

In our work we focus on the impact of crowd work environments and investigate their effects on different types of tasks. A taxonomy of task types in microtask crowdsourcing platforms has been developed in [GKD14d] where a two-level structure with 6 categories at the top level has been proposed. In our work we leverage such top level categorization to compare the effects of work environments on different types of tasks.

In our previous work we ran a large scale supervised classification job to analyze 130 million HITs published on AMT over 5 years with the goal of understanding patterns in task type changes over time [DCD$^+$15b]. We observed, for example, that content creation tasks (i.e., where workers are asked to generate some content like an audio transcription or a document summarization task) are the most popular on AMT, and in our experiments, are the ones in which workers performed poorly (See Table 7.6). Another popular task type on AMT are surveys [DCD$^+$15b], as a crowdsourcing platform allows easy and immediate access to large populations of participants.

7.2.3 Worker Differences & Participation Bias

Crowd workers are not all the same. For example, different workers have different skills and interests. In [DDCM13] we previously showed that it is possible to profile workers based on their social network activities and assigned tasks based on such profiles which model their interests, to increase the accuracy of crowd work.

Other types of differences in the crowd that participates and completes specific tasks are caused by incentive schemes. For example, different reward levels may attract different types of workers [EdV13] thus creating a participation bias in a study run on platforms like AMT. Jiang et al. analyzed the perceived benefits of participation in crowd work, and found that American and Indian workers differed in their perceptions of non-monetary benefits of participation. Indian workers valued self-improvement benefits, whereas American workers valued emotional benefits [JWN15]. Hsieh and Kocielnik showed how different reward strategies (e.g., lottery-based reward models) result in different types of crowd workers deciding to participate and complete the available tasks. They highlighted the consequent difference in the crowdsourced task results [HK16]. Along the same line, in [Har15] authors showed that rewarding workers when they quit their participation in a batch of HITs allows to filter out low-quality workers early, thus retaining only highly accurate workers. Recently, Findlater et al. showed that results of online HCI experiments are similar to those achieved in the lab for desktop interactions, but this was less so in the case of mobile devices [FZFM17].

Prior work has studied the reasons that drive senior adults to participate in crowd work and show both a low participation of such population as well as an interest for incentive types that differ from monetary ones [BMP16]. In contrast to this, in our work we analyze potential barriers to crowd work from a technological perspective showing important geographical differences in the type of devices and tools used by crowd workers which can also create participation bias in crowdsourcing studies. We also focus on how the technical infrastructure used by workers has an impact on participation and work quality in paid microtask crowdsourcing platforms.

7.2.4 Crowd Work Context and Barriers

Recently, ethnography-based research has been carried out to understand the contexts in which crowd workers are embedded. Authors of [MOGH16a] focused on the Indian and US worker communities and highlighted the effects of current crowdsourcing platforms and marketplaces on crowd worker experience. In [MHOG14], authors studied how crowd workers on AMT use web forums to create communities where they share experiences and voice crowd work-related problems. McInnis et al. report on the qualitative analysis of 437 comments where AMT workers were asked to comment on parts of the AMT participation agreement through an online discussion website [MCNL16]. They indicate 'unfair rejection' as a major issue for workers, and identify risk factors that lead to this issue. Workers discuss 'flaws in task or interface design' as a key factor, while the authors suggest sounding an alarm through various tools to intimate requesters about a broken task.

Narula et al. noted that microtask marketplaces were often inaccessible to workers in developing countries, and introduced a mobile-based crowdsourcing platform called Mobileworks for OCR tasks, thereby lowering a barrier for participation [NGR$^+$11]. Khanna et al. studied usability barriers that were prevalent on AMT, which prevented workers with little digital literacy skills from participating and completing work on AMT [KRDT10]. Authors showed that the task instructions, user interface, and the workers' cultural context corresponded to key usability barriers. To overcome such usability obstacles on AMT and better enable access and participation of low-income workers in India, the authors proposed the use of simplified user interfaces, simplified task instructions, and language localization. Vasantha et al. report an initial study of the demographic of 22 rural homeworkers in Scotland including computer skills, views on rural infrastructure and their skills in solving spatial visualization tests [VVC$^+$14]. The authors present results equivalent to survey-based studies conducted in the past, and suggest that the homeworkers can solve knowledge-intensive industrial spatial reasoning problems with minimum training. They asked participants to report on their computer and Internet skills, to which most participants reported 'good', while some reported 'fair'. In their work, the authors also call for more research on rural infrastructure (such as Internet connection and road connectivity) that support crowdsourcing work, as most participants expressed satisfaction with their infrastructure yet a few did not find them adequate for crowdwork. Jones et al. explored what it means to be a mobile phone user situated at the lower end of the socio-economic ladder in developing economies like India, South Africa and Kenya to own and operate digital and mobile devices with almost no access to computers [JRP$^+$16]. The authors suggest to engage with such users to help sketch out a technology road-map that will lead to devices and services which will be of value in the near future.

Several prior works have stressed the positive impact of good task design, clear instructions and descriptions on the quality of work produced [KCS08, MS13b, SHC11]. However, as pointed out by Kittur et al. task interfaces are often poorly designed or even have bugs that make it impossible to complete tasks [KNB$^+$13]. Poor quality work often arises from poorly designed crowdsourcing tasks. Morris et al. discuss the value of subcontracting microtask work and present value propositions for doing so [MBB$^+$17]. In that they hypothesize a contracting model specifically based upon the need for task improvement such that workers can fix issues with user interface components and task structure amongst other things, which currently takes place by way of informal back-channels [GSAK16, IS13].

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

7.3 Study I : UI Elements

The aim of this first study was to identify typical problems that crowd workers face during interactions with different UI elements embedded in tasks. During the course of task completion, crowd workers are exposed to the various UI elements that may or may not be carefully designed by requesters publishing tasks on a crowdsourcing platform. Recent work that analyzed 5 years of crowd work on AMT [DCD$^+$15b], found that there is an organic growth in the number of new requesters (over 1,000 new requesters each month in 2013, 2014). Such new requesters are typically unfamiliar with the process of task design and may put less effort to ensure adequate UI design before deployment. Even experienced requesters do not necessarily consider the work environments in which crowd workers contribute to piecework. Prior studies have highlighted the importance of appropriate task presentation, reflecting on the impact it has on a worker's perception of task complexity, cognitive load, and eventually on worker performance within the task [ASRI14]. Recently, Yang et al. investigated the role of *task complexity* in worker performance, with an aim to better the understanding of task-related elements that aid or deter crowd work [YRDB16]. In this chapter, we study the interplay between task design (in terms of the UI elements) and work environments (i.e., the context and differing conditions that crowd workers are embedded in). To this end, Study-I plays an important role to understand the typical issues that workers confront on a regular basis in crowd work.

7.3.1 Methodology and Survey Design

We designed a survey[5] asking workers about the issues that they typically faced with during their contributions in previous crowdsourcing tasks. The survey consisted of a few background questions, followed by questions corresponding to worker experiences while dealing with various UI input elements (*input boxes*, *text areas* spanning multiple lines, *checkboxes*, *dropdown menus*, *radio buttons* and *submit buttons*). Questions also covered other UI aspects such as *external navigation*, use of *colors*, experiences with *audio / video* content. To avoid misinterpretation, we presented workers with pictorial examples of each UI element. Finally, we provided workers with an opportunity to raise UI issues that were not addressed by the preceding questions in an open text field. We deployed the survey on CrowdFlower[6] and gathered responses from 100 distinct crowd workers. On average, each worker took just over 5 minutes to complete our survey and was compensated according to a fixed hourly rate of 7.5 USD on task completion. To detect untrustworthy workers and ensure reliability of the responses received, we followed the recommended guidelines for ensuring high quality results in surveys [GKDD15c]. To this end, we interspersed two attention check questions within the survey. We also used the filter provided by CrowdFlower to ensure the participation of high quality workers only (i.e., *level 3* crowd workers as prescribed on the CrowdFlower platform). We flagged 7 (out of 100) workers who failed to pass at least one of the two attention check questions and do not consider them in our analysis.

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

[6] http://crowdflower.com

7.3.2 Survey Results

We found that 43% of the workers who participated in the survey identified themselves as females (and 57% were males). Crowdsourcing microtasks served as a primary source of income for 42% of the workers. Table 7.1 presents the distribution of workers according to their age groups. We note a fairly even distribution of workers with respect to their age. As shown in Table 7.2, the workers who participated in the survey were also highly experienced in crowdsourcing work.

Table 7.1 Distribution of workers accoding to their age

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

Table 7.2 Experience of workers

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

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

1. **Input Boxes & Text Areas** – We found that 36% of workers raised issues that they faced with input boxes and text areas. 64% of the workers suggested that they did not experience problems in this regard. A recurring issue cited by workers with respect to input boxes and text areas was that the size of the input box and character limit were disproportionate, often leading to only a part of the entered text being visible upon entry (mentioned by 5% of the workers). The following is an excerpt from a worker who raised this issue:

> *'Usually, text fields just work well as I can put things that I want in the list. The only big issue is when people want a long text answer*

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

Over a decade ago, researchers suggested that matching the size of input boxes to the expected length is an important guideline to follow when designing forms on the Web [CDS07, CTL01, Wro08]. Another issue cited by workers was that of *input format validation* and *auto-correction*. Workers described occasions where text that was input was not accepted due to flawed format validation (especially, in case of URLs). These issues were cited by 15% of the workers. In other cases, input text was unfavorably auto-corrected, thereby hindering workers (mentioned by 6% of workers). Yet again, we found that a guideline that was suggested years ago with respect to accurate format validation [BAOP+11] is sometimes violated by requesters during task design. Workers also reported that sometimes default text in the input field gets appended to the text that is entered, instead of disappearing on input. Finally, workers brought to light that input fields are not enabled sometimes, leading to a situation where it is not possible for workers to enter any response (mentioned by 6% of workers).

2. **Checkboxes, Radio Buttons & Dropdown Menus** – We found that nearly 70% of the workers on average claimed to never have faced issues with checkboxes, radio buttons and dropdown menus. A recurring issue with checkboxes, radio buttons and dropdown menus was cases with too many options (mentioned by 10% of workers on average). This is a well-studied issue in UX design on the Web [BABT+11, BATO+10]. Another common problem was found to be the small size of checkboxes, radio buttons or the dropdown menu icon (mentioned by 6% of workers on average). Several workers referred to issues with selecting checkboxes and radio buttons due to the active region not including the corresponding text, or multiple clicks being required for selection (mentioned by 10% of workers on average). The following is an excerpt from a worker who raised this issue corresponding to checkboxes:

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

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

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

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

   > *'Slow loading, waiting very long for the audio/video to load completely.'*

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

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

   > *'many dead links, many old test questions linking to sites that have been changes so the test question is no longer valid'*

> *'Some links don't work, pages never open, by the way, i lost a level 3 badge with a problem like this.'*

5. **Colors Used & Submit Button** – We found that 80% of the workers did not typically face issues with the colors used in tasks. Some workers however, pointed out that poor color contrasts used in interfaces sometimes makes it hard to read the content of given tasks (nearly 15% of workers mentioned related problems). Around 50% of the workers claimed to have faced problems with the submit button. A common issue raised regarding the submit button was the poor positioning of the button (mentioned by 5% of the workers). Workers also complained that in some cases, the submit button was not enabled or had to be clicked multiple times (mentioned by 20% of the workers); another design violation [LF04]. Other issues pointed out include missing or unresponsive submit buttons and errors on clicking. Some excerpts from the responses of workers that reflect these issues are presented below.

> *'For me, some colors applied to data make it difficult to read the data - e.g., light grey.'*

> *'when you hit the keyboard 'enter' the task is automatically submit your work, even though you're not yet done.'*

> *'Sometimes the submit button didn't work at all. Unable to be pressed at all for a few minutes.'*

At first glance from a crowd worker's perspective, some of the issues raised might appear to be trivial to resolve or overcome through trial and error. However, prior ethnographic work has rightly stressed on the importance of considering the environmental context of crowd workers [MHOG14, GMHO14]. It is also well understood that a fair portion of workers are not entirely familiar with using computers and other devices which play a crucial role in their participation on crowdsourcing platforms [KRDT10]. We thereby believe that these issues play an important role in shaping the quality of crowd work and the fluidity with which it is completed.

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

7.4 Study II : Work Environments

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

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

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

7.4.1 Methodology and Task Design

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

To analyze the influence of different design considerations with respect to UI elements on worker performance, and their interplay with varying worker environments, we manually created a batch of 129 microtasks accounting for each of the 43 variations (each variation × 3 tasks), shown in the table on the companion webpage.These tasks consist of different types; *information finding*, *verification and validation*, *interpretation and analysis*, *content creation*, *surveys* and *content access* [GKD14d]. The table in the companion page also presents sample tasks that we created corresponding to each of the UI element variations; these tasks are noticeably designed to reflect real-world microtasks that have previously been deployed on crowdsourcing platforms.

Since understanding work environments would be a crucial part of this study, we deployed two identical batches of the 129 tasks on CrowdFlower, one addressing workers based in USA and the other addressing workers based in India. We considered USA and India since they represent two of the largest populations of crowd workers [Ipe10b], and due to the potentially different work environments they entail. In both cases, we used the inbuilt CrowdFlower feature to restrict participation to only the

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

highest quality level of the crowd. We gathered 50 distinct judgments from workers in response to each of the 129 tasks in the batch, resulting in 6,450 × 2 responses (from USA and India workers). Workers were randomly assigned to tasks in the batch and the order of tasks was also randomized. Workers were compensated in proportion to the amount of work completed at a fixed hourly piece-rate of 7.5 USD.

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

7.4.2 Results and Analysis

Note that we report our results including the test statistic, degrees of freedom and effect sizes (Cohen's *d*, Hedge's *g*) for statistically significant findings. We acquired responses from 90 Indian and 95 American workers. The overall comparison between the two groups of workers is presented in Table 7.3. We did not find significant differences between Indian and American workers in terms of their overall accuracy or retention rate (i.e., number of HITs completed by each worker). However, we found that the American workers were significantly faster in completing the tasks, with an average task completion time (TCT) of 0.89 mins compared to the Indian workers (1.39 mins); *t(183)=3.06, p<.05, Cohen's d=.45.*

Table 7.3 Overall comparison between workers from India and USA. The asterix ('*') indicates a statistically significant difference between the two groups.

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

We also investigated the overall differences in the average accuracy and task completion times of Indian and American workers on *good* and *bad* variations. Table 7.4 presents our findings. By using multiple t-tests with Bonferroni correction, we

Table 7.4 Overall comparison between workers from India and USA with respect to *good* and *bad* variations of UI elements. The asterix ('*') indicates statistically significant differences between the *good* and *bad* variations for a group.

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

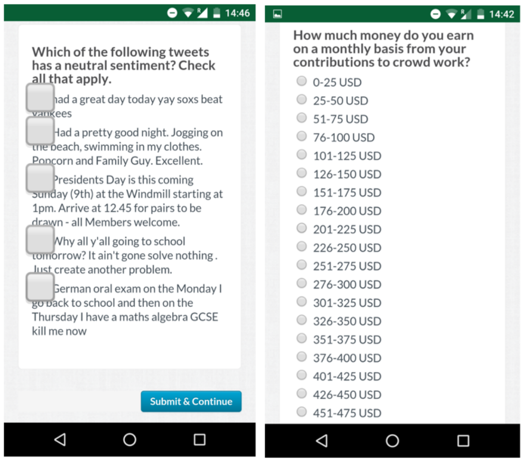
note that there are statistically significant differences between the accuracy of Indian workers on tasks with *good* versus *bad* variations; *t(178)=2.15, p<.05, Cohen's d=.36*. On the other hand, we found that American workers exhibited a significant difference in the task completion time across *good* and *bad* variations; *t(188)=3.66, p<.001, Cohen's d=.29*, requiring significantly less time to complete tasks with *good* design variations as intuitively expected. We also observe that workers from USA take less time to complete tasks (see Table 7.3) than workers from India (independently from the *good* or *bad* design). In summary, this suggests that American workers deal better with tasks that are poorly designed.

Performance Across UI Element Variations

To understand how workers coped with tasks involving the different UI element variations, we analyzed the performance of workers corresponding to each UI element grouped into *good* and *bad* variations. The results of our analyses are presented in Table 7.5.

Statistical Significance – We computed the statistical significance of our observations in Table 7.5, using multiple t-tests. To control for Type-I error inflation in our multiple comparisons, we use the Holm-Bonferroni correction for family-wise error rate (FWER) [Hol79], at the significance level of $\alpha < .05$. Statistically significant differences between tasks with *good* and *bad* UI element variations are marked with an asterix (*).

- **Input Boxes** – Both Indian and American workers performed considerably better in tasks having *good* variations of input boxes, in comparison to *bad* variations. Further investigation revealed that disproportionately large input boxes (`ib_largeSize`) corresponded to the least accuracy: 30.67% in the case of Indian workers and 34.67% in the case of American workers. We did not find significant differences in the average task completion time (TCT) for both Indian and American workers between the two variations. This indicates that workers spent roughly the same amount of time on tasks with *bad* design variations



Which of the following tweets has a neutral sentiment? Check all that apply.

□ had a great day today yay soxs beat yankees

□ Had a pretty good night. Jogging on the beach, swimming in my clothes. Popcorn and Family Guy. Excellent.

□ Presidents Day is this coming Sunday (9th) at the Windmill starting at 1pm. Arrive at 12.45 for pairs to be drawn - all Members welcome.

□ Why all y'all going to school tomorrow? It ain't gone solve nothing . Just create another problem.

□ German oral exam on the Monday I go back to school and then on the Thursday I have a maths algebra GCSE kill me now

Submit & Continue

How much money do you earn on a monthly basis from your contributions to crowd work?

○ 0-25 USD
○ 25-50 USD
○ 51-75 USD
○ 76-100 USD
○ 101-125 USD
○ 126-150 USD
○ 151-175 USD
○ 176-200 USD
○ 201-225 USD
○ 226-250 USD
○ 251-275 USD
○ 276-300 USD
○ 301-325 USD
○ 326-350 USD
○ 351-375 USD
○ 376-400 USD
○ 401-425 USD
○ 426-450 USD
○ 451-475 USD

Figure 7.1 Left– An *interpretation and analysis* task with large checkboxes (`cb_largeSize`); Right– A *survey* task with many radio buttons (`rb_manyOptions`), as rendered on an Android mobile phone and viewed on a Chrome browser.

Table 7.5 Performance of Indian and American workers corresponding to each of the UI elements, grouped into *good* and *bad* variations. Statistically significant differences between tasks with *good* and *bad* UI element variations are marked with an asterix (*).

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

despite the potential hindrance, reflecting their genuine attempt to provide high quality responses.

- **Text Areas** – In the case of tasks corresponding to text areas, Indian workers depict a better performance in the *good* variations when compared to the *bad* variations. No significant differences were observed for American workers. We found that both groups of workers performed with the least accuracy in the (`ta_smallSize`) variation where the visible size of the text area is small, making it inconvenient for workers to see all the text that is entered. Similar to our findings with respect to tasks with variations of input boxes, we did not find significant differences in the average TCT of workers between the *good* and *bad* variations.

- **Checkboxes** – We found that both Indian and American workers performed much better in tasks with *good* variations of checkboxes than in *bad* variations. An example of a bad variation is shown in Figure 7.1. In contrast to Indian workers where we found no significant difference in their average TCT, the American workers were faster in tasks with *good* variations. Further investigation revealed that the (`cb_manyOptions`) variation corresponded to the least accuracy among both Indian and American workers in tasks with checkbox variations.

- **Radio Buttons** – Tasks with variations of radio buttons correspond to better performance from both Indian and American workers in the *good* as opposed to in the *bad* variations. However, the performance of workers in the *bad* radio button variations when compared to that in case of the checkboxes is significantly higher (Indian workers: $t(125)=6.452,\ p<.001$, American workers: $t(163)=4.805,\ p<.001$). This is explained by the simpler nature of radio buttons, where questions are modeled to accept a single response (even in case there are many options; as in `rb_manyOptions`) (an example is shown in Figure 7.1).

- **Audios** – In tasks with audio variations, we note that compared to other tasks, the overall level of accuracy drops in the case of Indian and American workers, both in the *good* and *bad* variations. We attribute this drop in accuracy to the inherently more complex nature of audio transcription tasks [PE10], where work environment specifics (such as device volume, headsets or other equipment) may play a role. This is exacerbated in the case of audios with poor quality (`audio_poorQuality`). We found that Indian workers perform considerably better in audio transcription tasks with good quality variations when compared to the poor quality variations. They also take more time on tasks with the `audio_goodQuality` variations. On further scrutiny, we found that in tasks with poor audio quality, several Indian workers gave up after trying to transcribe the audio, condemning the poor quality of the audio in their responses. In contrast, we found that American workers performed similarly in both the *good* and *bad* variations, without a significant difference in the average TCTs.

- **Images** – We found that Indian and American workers performed similarly in the tasks with either *good* or *bad* image variations, without significant differences in TCTs between the two variations. Across both groups of workers we found that the `img_smallSize` variation corresponded to the lowest accuracy of workers in image variation tasks.

- **Videos** – In case of tasks with videos, we note that both Indian and American workers do not exhibit a significant difference in their performance between the *good* and *bad* design variations. We found that American workers took less time to complete the tasks with *good* variations as opposed to *bad* variations (i.e., tasks with the `video_poorQuality` variation).

- **Dropdown Menus** – In the case of tasks with dropdown menu variations (related to active region and icon size), we found that both Indian and American workers perform with similar accuracy and take similar amounts of time in both the *good* and *bad* variations.
  External Links – We found no effect on the accuracy of Indian and American workers or their task completion times based on the type of external links (`elink_HTTP` or `elink_HTTPS`), or whether the links opened in the same or new tab (`elink_sameTab` or `elink_newTab`).

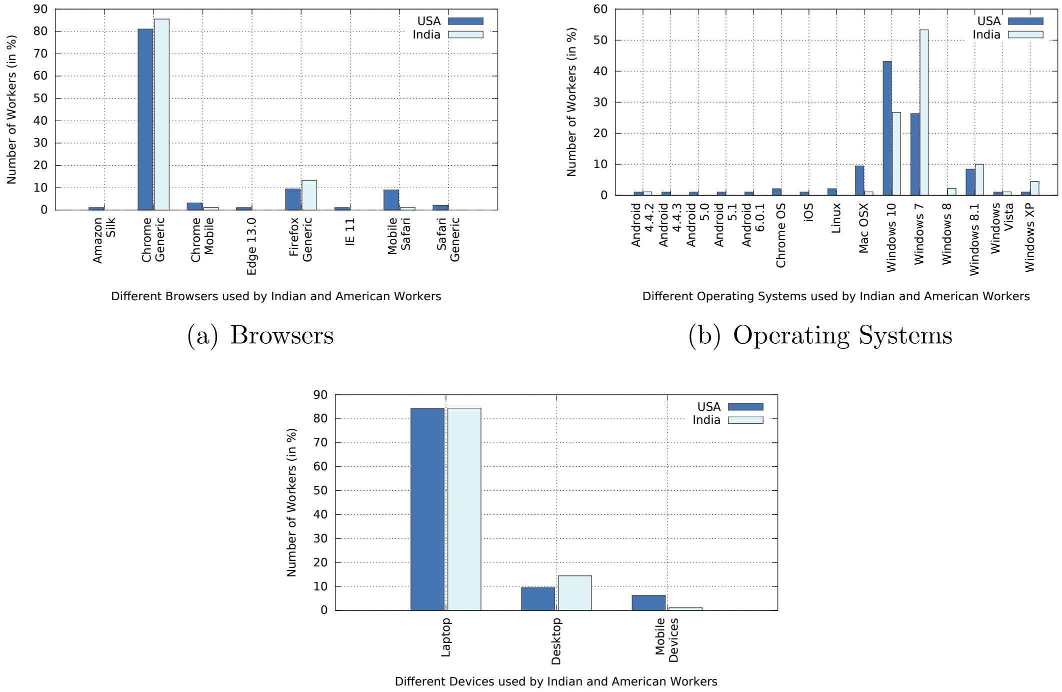
- **Rating Scales** – Due to the subjective nature of rating scales, we only consider the TCT of workers as a measure of performance. We did not find significant differences across different rating scales.

7.4.3 Role of Work Environments

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

Browsers, Operating Systems and Devices

By resolving the `user-agent` strings of workers who participated in our tasks, we identified their browsers, operating systems and devices. The distribution of workers according to these specifics are presented in Figure 7.2. We found that there is a far greater variety in browsers, operating systems and devices used by crowd workers from the USA than those from India. However, the most popularly used browsers (*Chrome*



(a) Browsers

(b) Operating Systems

(c) Devices

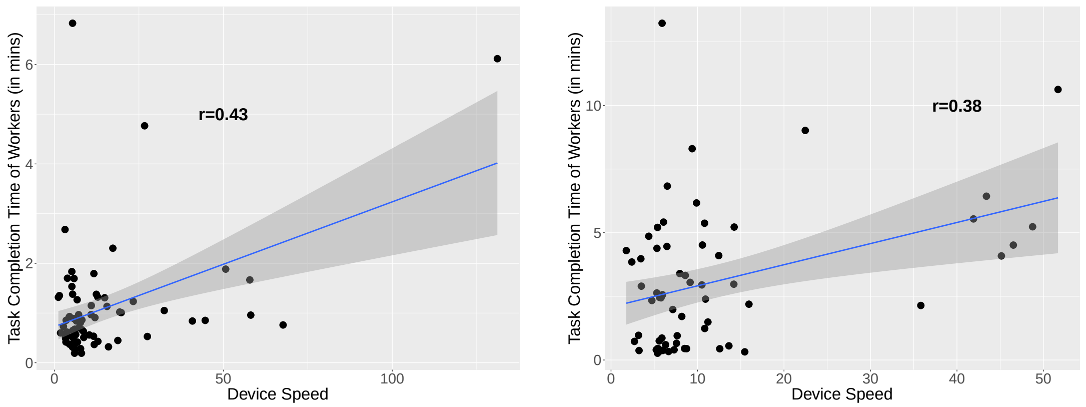
Figure 7.2 Distribution of Indian and American workers in the 129 tasks according to their browser, operating system and devices used.

Generic, *Firefox Generic*), operating systems (*Windows 7*, *Windows 10*) and devices (*Laptops*, *Desktops*) are similar across both groups of workers. It is noteworthy that the American workers appear to use more of the latest available technology such as the Windows 10 operating system and Macbooks, in comparison to Indian workers in our tasks.

Impact of Device Speed

We investigated the relationship between the speed of devices that Indian and American workers used, and their accuracy and TCTs across different tasks. Figure 7.3 presents the statistically significant relationships that we found.

We found a moderate positive correlation between the speed of the device used and the task completion time of American workers in tasks with text areas (see Figure 7.3(a)); *r(72)=.43*, $p<.001$. This suggests that American workers use more time to complete tasks with text areas when the device used is relatively faster. Thus in tasks with text areas, we found that the speed of the device used by American workers accounts for nearly 18.5% of the variance in their task completion times (the coefficient of determination, R^2=0.184). Further scrutiny revealed that faster



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

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

Figure 7.3 Impact of device speed on different types of tasks (only statistically significant linear relationships are presented).

devices led to American workers providing more tags as well as more unique tags in the tagging task (see Table 1 in the companion webpage) corresponding to text areas. We investigate and discuss these findings further in a follow-up qualitative study described in Section 7.5.

Similarly, in tasks involving audio media, we found a moderate positive correlation between the task completion time of American workers and the speed of the devices used (see Figure 7.3(b)); $r(62)=.38$, $p<.001$. Accordingly, the speed of the devices used by American workers in tasks with audio variations accounted for nearly 14.5% of the variances in their task completion times (the coefficient of determination, R^2=0.144). We did not find significant correlations with respect to the devices used by Indian or American workers across other task variations.

Impact of Devices on Worker Performance

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

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

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

Our findings regarding the impact of devices on worker performance are further explored and discussed in Section 7.5, through follow-up individual interviews with Indian and American workers.

Impact of Screen Resolution

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

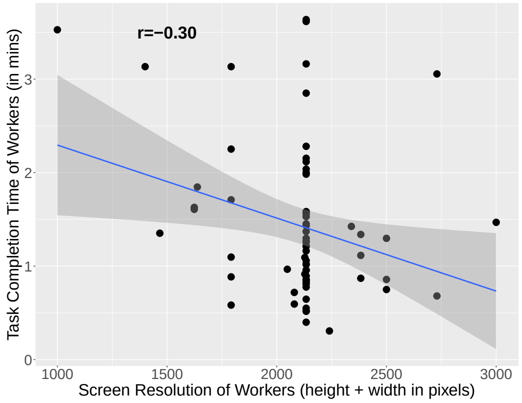


Figure 7.4 Relationship between *screen resolution* and *TCT* of Indian workers who completed tasks with *image* variations.

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

We found a weak negative correlation between the screen resolution and the task completion time of Indian workers in tasks with UI element variations corresponding to images (see Figure 7.4); *r(88)=-.30*, p<.001. This indicates that lower screen resolutions can hinder workers in tasks that involve images, resulting in longer task completion times. Accordingly, the screen resolution of Indian workers accounted for 9% of the variance in their task completion times (the coefficient of determination, R^2=.09). We did not find significant correlations with screen resolution of devices across other UI element variations.

Impact on Different Task Types

Based on the taxonomy of microtasks proposed in [GKD14d], we analyzed the impact of work environments on the different task types; *information finding* (IF), *interpretation & analysis* (IA), *verification & validation* (VV), *content creation* (CC), *surveys* (SU) and *content access* (CA). Table 7.6 presents the distribution of the 129 tasks according to the taxonomy, and the overall average work accuracy of Indian (IND-Acc) and American (USA-Acc) workers corresponding to each task type. While we found differences in accuracy within each group across the different task types, we did not find statistically significant differences in worker performance between Indian and American workers across the task types.

Table 7.6 Distribution of the tasks deployed according to their type.

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

Although we did not find significant differences in worker accuracy in each of the task types across the devices (desktops, laptops or mobile devices), we found that in information finding (IF) and content creation (CC) tasks, both Indian and American workers using mobile devices required significantly more time for task completion. This indicates that laptops and desktops are more suitable than mobile devices for certain task types. We reason that content creation tasks typically involve typing content, which is inherently easier to accomplish using a keyboard on a laptop or desktop computer, as opposed to a mobile device (considering that it is easier to deal with active regions corresponding to input boxes and text areas on laptops and desktops). In the case of information finding tasks, workers are typically required to search the Web, return to the task, and provide their responses. We reason that such toggling between tabs or windows is easier to accomplish on desktops and laptops in comparison to mobile devices.

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

7.5 Study III : Follow-up Personal Interviews

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

7.5.1 Methodology

To better understand the influence of work environments on crowdsourced microtasks, we conducted 7 semi-structured interviews with CrowdFlower workers [TON+14] who completed all tasks in the batches described earlier in Study II.

We randomly selected 20 American and 20 Indian workers who completed all the tasks, and contacted them via e-mail requesting their participation in a follow-up interview over Skype or Google Hangout. In the email[8], workers were notified about the purpose and nature of the interview, the participation reward and mode of compensation (i.e., a bonus payment on CrowdFlower according to the hourly-rate of 7.5 USD), and an assurance of anonymity. We sent out the recruitment emails to workers nearly 3 months after their participation in Study II, to avoid any bias in their responses and perception of tasks stemming from the recency of participating in our batch of tasks with *good* and *bad* variations. 5 American and 10 Indian workers showed interest to participate in the interviews. Of these, 2 American and 5 Indian workers scheduled and completed the interviews. Table 7.7 presents some characteristics of the interview participants in Study III.

Two of the authors conducted semi-structured interviews with the interested workers. Participants were first briefed about the identity and work of the authors, and

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

Table 7.7 Characteristics of American (AW) and Indian (IW) Interview Participants

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

the structure of the interview. They were informed that, with their consent, the interviews would be audio recorded and transcribed. After receiving a verbal consent, workers were asked a series of questions ranging from their general background, experience and motivation for participation in crowd work, to details about their work environment. The wordings of questions were made intentionally neutral to elicit responses from participants without being influenced by the interviewer. Some of the questions asked during the interviews are presented below.

A sample of questions asked during the interviews with participants.

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

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

7.5.2 Results and Findings

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

Work Environments – Device Usage

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

In contrast to these workers, **IW4**, **AW1** and **AW2** reported the usage of a single device to complete crowdsourced microtasks. **IW4** uses an assembled desktop computer, **AW1** uses a laptop, and **AW2** uses a desktop. **AW1** indicated her preference to use her laptop by saying, "*I only use my laptop. I think mobile devices are too small, (my) laptop is big. I like to be able to see what I'm doing. So there we go!*".

AW2 indicated her preference to use her desktop since most of the tasks she completes involve typing, and she is better at typing on a desktop with a keyboard. She also said, "*I can see the screen better on my desktop*".

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

Work Environments — Internet Connection, Device Speed and Screen Resolution

We asked workers about the quality of their Internet connections as well as their devices. **IW1** reported having a 2 MBPS unlimited broadband connection. His desktop is 10 years old, and his laptop is 2 years old. **IW1** said he experienced issues due to bandwidth limitations in tasks that contained images. **IW2** mentioned that his laptop is 2 months old, and that he replaced the old one due to the slow processor, and small screen size and keypad. He reported that he bought a new laptop to mainly do CrowdFlower tasks and described his Internet connection as follows; "*In the past I had a very slow Internet connection. But now I have high-(speed) and unlimited Internet connection and everything is fine. New laptop is making a huge difference in speed and especially helping my accuracy*". However, he cited having issues with loading tasks that contain media such as images, audios/videos. **IW3** also reported having an unlimited broadband connection. He explained that despite his high speed connection, he faced problems with tasks containing videos, "*Sometimes I feel a bit odd, some videos might not be like buffering fast and I have to wait and wait for sometime. And yesterday I did some tasks and the videos were buffering, so too slow. I got a bit frustrated. I think it was a problem with that task actually. It's not happening with all the tasks, but some tasks, videos are not opening, its taking time to buffer, even though my Wi-Fi has speed but its taking time, buffering buffering buffering. I just pressed the play button and went to the kitchen and did my cooking and then yeah, came back and started doing the task. Luckily had enough time to do that. It's a bit more time-consuming. I have to be looking into it, let it buffer, it takes time. So I am losing time, time is money. If I can do it fast I can go for the next task. It's money*". **IW4** reported having a dial-up connection that he uses via a modem. He said, "*Sometimes the Internet is slow, so I convert my Internet to 3G to download images. Time is sufficient, it downloads ... to finish it before the task ends*". **IW5** uses a data card. Referring to the slow speeds he deals with, he said

"*The service is pathetic, very slow Internet. The network drops often, sometimes in the middle of task and this means a lot of time and effort wasted*". In contrast, **AW1** affirmed that she has no issues with her Internet connection. She said, "*I have high bandwidth, the highest you can get*". **AW2** also claimed to have a high-speed Internet connection and said she did not face any problems due to her Internet connection.

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

Language Proficiency

Previous works [Gup17, KRDT10] have studied how work is completed within different crowdsourcing tasks on AMT and establish that there are language proficiency differences amongst workers. These differences where shown to range from bare-minimal functional knowledge of English to highly sophisticated, contextual use of the language. Language proficiency can have a large impact in tasks with poorly designed UI elements, creating amplified difficulties for workers. For example in tasks with poor audio or large interface elements that hide text, workers who are highly proficient in the language have a better chance of guessing unclear words and providing accurate transcriptions and responses. The range in proficiency became apparent in Study III as we asked participants questions about their use of digital devices for personal use in the day to day, and gauged their effective use of English during the interviews. For example, worker **IW4** who has a degree in Homeopathic Science and ran a Homeopathic clinic in a small town operating in a regional language in India, acquired a basic education of English but hardly used the language on a daily basis except for when he worked on CrowdFlower.

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

Dealing with Issues Emerging from Poor Task Design

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

Summary of Insights from Study III

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

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

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

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

7.6 Main Findings and Implications

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

General Insights –

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

American Workers versus Indian Workers –

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

American and Indian Workers : Finer Details –

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

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

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

7.6.1 ModOp – A Tool for Environment-Aware Task Design

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

- *Input Box / Text Area size* – Warning triggered if the size is disproportionately small or large.
- *Image size and resolution* – Warning triggered if the image size or resolution is disproportionately small.
- *Checkboxes* – Warning triggered if number of checkboxes is not optimal; re-

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

questers are advised to split checkbox questions where there are more than 10 options.

- *Radio Buttons* – Warning triggered if the number of radio buttons corresponding to a question is > 4, based on [BATO+10].

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

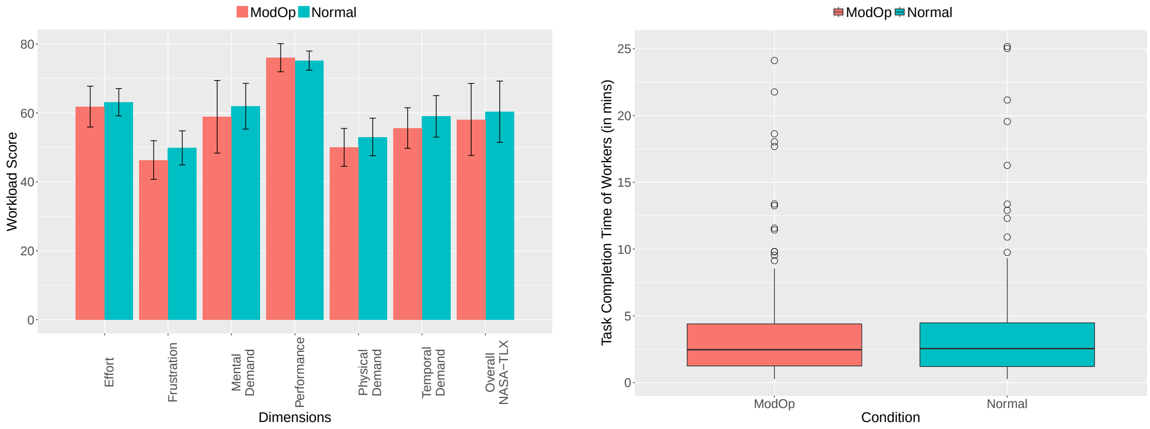
- *Device Type* – *ModOp* automatically detects the device type of workers by leveraging the user agent string.
- *Device Speed* – *ModOp* automatically provides an estimate of the worker's device speed based on a target relative speed.
- *Screen Resolution* – *ModOp* automatically detects the screen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

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

We believe that this tool can help crowdsourcing requesters in designing better tasks by directing requesters' attention to otherwise neglected attributes of UI element design and work environments. We envision that this would improve the overall crowd worker experience and help in reducing their cognitive load [Whi13].

7.6.2 Evaluation of *ModOp*

We performed an evaluation of the ModOp tool using real-world microtasks by considering the impact of *ModOp* on the cognitive load of workers. Cognitive load refers to the total amount of mental effort being used in the working memory of a person with respect to the task at hand. Early work showed that instructional design can be used to reduce cognitive load in learners [Swe88]. Cognitive load theory has been used widely to inform and evaluate several web-based tools and systems; Feinberg et al. proposed to leverage cognitive load to inform web-based design [FM00], Oviatt used the theme of cognitive load to design human-centered interfaces [Ovi06], Wang et al. studied website complexity from the cognitive load perspective [WYL+14], Schnabel et al. proposed the use of shortlists to improve the performance of recommender systems and reduce the cognitive load of users [SBDJ16]. Thus, our intuition behind using cognitive load as a metric to evaluate *ModOp* was to observe whether the input propelled by *ModOp* related to the design of UI elements affects the task perception of crowd workers.



(a) Cognitive Load

(b) Task Completion Time

Figure 7.5 Cognitive load of workers who completed tasks in the *Normal* condition compared to the *ModOp* condition, measured using the NASA Task Load Index (NASA-TLX), and the corresponding task completion time.

We considered a dataset of 61 tasks that were previously deployed on Amazon's Mechanical Turk, and consisting of different task types [YRDB16]. By running the *ModOp* tool on the original task designs, we identified over 20 tasks for which *ModOp* suggested UI design improvements. This accounts for nearly one-third of tasks in the dataset. However, not all modifications suggested by *ModOp* were feasible to implement given the dataset constraints: for example, increasing image resolution is infeasible without possession of high resolution sources. Thus, we consider 9 different tasks where improvements suggested by *ModOp* were feasible to implement. The goals of these tasks included guided learning, interpretation and analysis, image classification and content moderation. In the *Normal* condition, we deployed these tasks on CrowdFlower in their original form and collected 20 judgments from distinct crowd workers. In an identical setup barring the modifications in task design suggested by ModOp, i.e., the *ModOp* condition, we deployed the improved task designs on CrowdFlower. In both conditions, workers were compensated with the same reward level computed based on a fixed hourly rate of 7.5 USD. On completion of each task, workers were asked to complete an online version of the NASA-TLX questionnaire [Har06] to measure the associated cognitive load of completing the crowdsourcing task.

Figure 7.5(a) presents results comparing the cognitive load of tasks measured using the 6 sub-scales of NASA-TLX as well as the overall workload between the two conditions. Through a two-tailed T-test we found that workers in the *ModOp* condition (*M=58.11, SD=11.45*) perceived a statistically significant lower workload than in the *Normal* condition (*M=60.35, SD=9.52*); $t(41)=2.524, p < .05$. Figure 7.5(b) draws a comparison in the task completion time of workers in the two conditions.

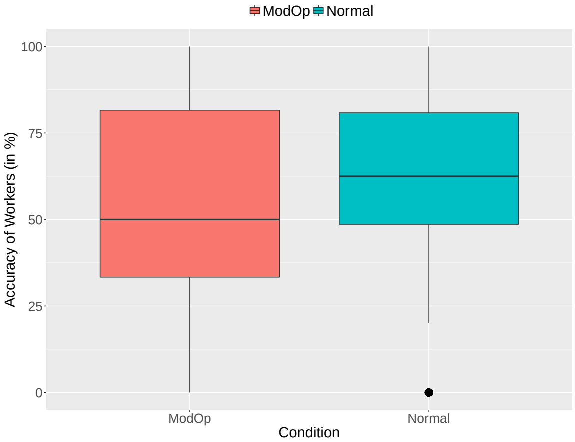


Figure 7.6 Performance of Workers in the two Conditions

This suggests that the modifications in task design suggested by *ModOp* can reduce the cognitive load experienced by workers.

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

Finally, we analyzed the performance of workers in both the *Normal* (*M=57.20, SD=33.23*) and *ModOp* (*M=52.29, SD=36.59*) conditions. Figure 7.6 presents our findings. We did not find a statistically significant difference in the accuracy of workers across the two conditions using a two-tailed T-Test; *t(105)=0.525, p=.47*.

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

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

7.7 Discussion, Caveats and Limitations

From the results of our first study, we found that UI elements in crowdsourced microtasks pose several issues to crowd workers. We note that some of the widely cited issues from workers, emerge from violations of classic UI design guidelines highlighted in previous work related to web form design [BATO$^+$10, Wro08]. This indicates that requesters do not take special care to ensure optimal design of UI elements. By accounting for *good* and *bad* design of UI elements in crowdsourcing tasks, we explored the role of crowd work environments in determining the quality of work that is produced.

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

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

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

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

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

7.8 Chapter Summary

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

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

8 Conclusions and Future Work

> *"The future belongs to those who believe in the beauty of their dreams."*
> — *Eleanor Roosevelt*

In this thesis, we have identified and addressed some important problems in the space of microtask crowdsourcing. We investigated key obstacles that hinder the effectiveness of the paradigm, and we proposed and evaluated several novel methods to overcome these challenges. Our findings enrich the current understanding of crowd work and bear important implications on structuring workflow. In this chapter, we draw main conclusions from our findings presented in this thesis and set precedents for future work.

8.1 Main Conclusions

This dissertation has addressed three main challenges that hinder the effectiveness of microtask crowdsourcing; (i) a limited understanding of crowdsourced tasks and worker behavior, (ii) inadequate means for worker pre-selection, and (iii) an incomplete consideration of factors that influence the quality of work produced.

Through our work presented in chapters 2 and 3, we have advanced the understanding of tasks as well as worker behavior and characteristics. We proposed a two-level categorization scheme for crowdsourced microtasks. This fine-grained categorization of tasks has important implications for user modeling of crowd workers, task design, deployment and recommendation. Showcasing the task type of *surveys*, we analyzed the wide range of trustworthy and untrustworthy behavior exhibited by crowd workers, which go beyond exisiting works and are better justified through data. For the benefit of requesters and to ensure adequate utilization of the platform, we also proposed behavioral metrics that can be used to counter undesirable activity in crowdsourced microtasks.

In Chapter 4, we showed that worker behavioral traces can be leveraged to classify them according to a fine-grained worker typology, which in turn can be used for more effective worker pre-selection. Since our approach is based on gathering behavioral signals from a worker during the pre-screening phase, no prior information about a woker is required. This has important implications since requesters typically have little or no prior knowledge of a worker's behavioral type.

In Chapter 5, by investigating the Dunning-Kruger effect in the crowd, we showed that there is a disparity in the crowd regarding the metacognitive ability of workers. This hinders the performance of workers and deprives learning. We argued that the capability of a worker to accurately self-evaluate is an integral aspect of the worker's competence. Through our rigorous evaluation in tagging, sentiment analysis and image validation tasks, we showed that requesters can benefit by operationalizing workers' self-assessments as a means of assessing their competence rather than relying solely on their performance in worker pre-selection phases.

We then investigated the role of hidden factors such as *task clarity* and *work environments* in influencing crowd work (chapters 6 and 7). With an extensive study in Chapter 6 we revealed that clarity is coherently perceived by workers, and that it varies with respect to the task type. We proposed a supervised machine learning model to predict task clarity and showed that clarity can be accurately predicted. Finally, through temporal analysis, we show that clarity is not a macro-property of the crowdsourcing marketplace ecosystem, but rather a local property influenced by tasks and requesters. In Chapter 7, we studied the effect of work environments (characterized by the hardware and software affordances of crowd workers) on the efficiency and effectiveness of microtask crowdsourcing across two different worker populations. Our findings from multiple studies indicate a significant impact of good and bad HIT designs for certain task types and across American and Indian crowd workers. We found substantial evidence that confirms the invisible role of work environments in shaping crowd work through experimental findings and supported by individual interviews.

By carrying out various studies, proposing different methods to deal with the key challenges that were identified, and through extensive evaluations, we have made the following noteworthy contributions in improving the effectiveness of microtask crowdsourcing – (i) we have advanced the understanding of task types, worker behavior and quality control, (ii) we have proposed novel mechanisms for worker pre-selection that outperform existing methods, and (iii) we reveal the influence of hidden factors such as task clarity and work environments in shaping the quality of crowd work.

8.2 Future Directions

Building on our observations and findings presented in this thesis, we plan to investigate the following aspects of microtask crowdsourcing in the imminent future.

- **Measuring learning in microtask crowdsourcing marketplaces**.

  Microtask crowdsourcing presents a unique learning context where workers have to learn to complete tasks on-the-fly by applying their learning immediately through the course of tasks. With thousands of workers around the world turning to microtask crowdsourcing platforms to earn their livelihood, it is worthy to investigate the learning that occurs through the course of task completion on such platforms. Measuring learning can pave way to intelligent task grouping and ordering to optimize the learning outcomes for participating workers.

- **Inducing competence-based self-selection of tasks in crowdsourcing platforms**.

  Task consumption in crowdsourcing marketplaces is largely defined by the self-selection of workers. In order to ensure that workers refrain from participating in microtasks that are beyond their competence, they first need to be aware of their limitations. By providing workers with an assessment of their competence in particular microtasks, we hypothesize that workers can better select microtasks which are suitable to their competence. Crowdsourcing marketplaces can greatly benefit from this by training their workforce to progress towards higher competence and improved reputation. This in turn would help workers to qualify for a larger spectrum of tasks, resulting in a greater turnover for workers. In the future, we plan to experimentally investigate the affect of competence-based feedback on consequent task selection behavior of workers.

- **Worker activity dashboards to promote self-reflection and improve worker peformace**.

  Several tools and techniques have been developed to support task requesters in identifying desirable workers, filtering out undesirable workers, and visualizing worker activity. However, supporting worker reflection, learning and development has received a relatively limited amount of focus so far. In the near future, we aim to fill this gap by proposing the use of activity dashboards as a means to promote worker reflection and metacognition with the ultimate aim of improving worker engagement, learning and development.

  We believe that presenting workers with real-time feedback with respect to their low-level behavior within tasks can promote reflection on how they are completing tasks, thereby fostering a better understanding of how they can be more effective (i.e., complete tasks faster, adopt better workflows, improve their accuracy and so forth).

Bibliography

[ABY11] Omar Alonso and Ricardo Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In *ECIR*, pages 153–164. Springer, 2011.

[ALV14] Ioannis Arapakis, Mounia Lalmas, and George Valkanas. Understanding within-content engagement through pattern analysis of mouse gestures. In *Proceedings of CIKM '14*. ACM, 2014.

[AMN14] Omar Alonso, Catherine Marshall, and Marc Najork. Crowdsourcing a subjective labeling task: a human-centered framework to ensure reliable results. Technical report, MSR-TR-2014-91, 2014.

[ASRI14] Harini Alagarai Sampath, Rajeev Rajeshuni, and Bipin Indurkhya. Cognitively inspired task design to improve user performance on crowdsourcing platforms. In *CHI*, pages 3665–3674. ACM, 2014.

[BABT$^+$11] Javier A Bargas-Avila, Olivia Brenzikofer, Alexandre N Tuch, Sandra P Roth, and Klaus Opwis. Working towards usable forms on the worldwide web: optimizing multiple selection interface elements. *Advances in Human-Computer Interaction*, 2011:4, 2011.

[BAOP$^+$11] Javier A Bargas-Avila, Sébastien Orsini, Hannah Piosczyk, Dominic Urwyler, and Klaus Opwis. Enhancing online forms: Use format specifications for fields with format restrictions to help respondents. *Interacting with Computers*, 23(1):33–39, 2011.

[BATO$^+$10] Javier A Bargas-Avila, AN Tuch, K Opwis, O Brenzikofer, S Orsini, and SP Roth. *Simple but crucial user interfaces in the world wide web: introducing 20 guidelines for usable web form design*. INTECH Open Access Publisher, 2010.

[BBC+13] Alessandro Bozzon, Marco Brambilla, Stefano Ceri, Matteo Silvestri, and Giuliano Vesci. Choosing the right crowd: Expert finding in social networks. In *Proceedings of EDBT'13*, pages 637–648. ACM, 2013.

[BBD+16a] Jakob Beetz, Ina Blümel, Stefan Dietze, Besnik Fetahu, Ujwal Gadiraju, Martin Hecher, Thomas Krijnen, Michelle Lindlar, Martin Tamke, Raoul Wessel, and Ran Yu. Enrichment and preservation of architectural knowledge. In *3D Research Challenges in Cultural Heritage II - How to Manage Data and Knowledge Related to Interpretative Digital 3D Reconstructions of Cultural Heritage*, pages 231–255. 2016.

[BBD+16b] Jakob Beetz, Ina Blümel, Stefan Dietze, Besnik Fetahu, Ujwal Gadiraju, Martin Hecher, Thomas Krijnen, Michelle Lindlar, Martin Tamke, Raoul Wessel, and Ran Yu. Enrichment and preservation of architectural knowledge. In *3D Research Challenges in Cultural Heritage II - How to Manage Data and Knowledge Related to Interpretative Digital 3D Reconstructions of Cultural Heritage*, pages 231–255. 2016.

[BEH+06] Jeffrey A Burke, Deborah Estrin, Mark Hansen, Andrew Parker, Nithya Ramanathan, Sasank Reddy, and Mani B Srivastava. Participatory sensing. *Center for Embedded Network Sensing*, 2006.

[Ber16] Janine Berg. Income security in the on-demand economy: findings and policy lessons from a survey of crowdworkers. *Comparative Labor Law & Policy Journal*, 37(3), 2016.

[BGK+12] Yoram Bachrach, Thore Graepel, Gjergji Kasneci, Michal Kosinski, and Jurgen Van Gael. Crowd iq: aggregating opinions to boost performance. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 535–542. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

[BKK+13] Yukino Baba, Hisashi Kashima, Kei Kinoshita, Goushi Yamaguchi, and Yosuke Akiyoshi. Leveraging crowdsourcing to detect improper tasks in crowdsourcing marketplaces. In *Twenty-Fifth IAAI Conference*, 2013.

[BL11] Sabine Buchholz and Javier Latorre. Crowdsourcing preference tests, and how to detect cheating. In *INTERSPEECH*, pages 3053–3056, 2011.

[BLK06] Katherine A Burson, Richard P Larrick, and Joshua Klayman. Skilled or unskilled, but still unaware of it: how perceptions of difficulty drive miscalibration in relative comparisons. *Journal of personality and social psychology*, 90(1):60, 2006.

[BLL04] Bruce Lawrence Berg, Howard Lune, and Howard Lune. *Qualitative research methods for the social sciences*, volume 5. 2004.

[BMP16] Robin Brewer, Meredith Ringel Morris, and Anne Marie Piper. "why would anybody do this?": Understanding older adults' motivations and challenges in crowd work. In *CHI*, CHI '16, pages 2246–2257, New York, NY, USA, 2016.

[Bou13] David Boud. *Enhancing learning through self-assessment*. Routledge, 2013.

[Bra08] Daren C. Brabham. Crowdsourcing as a Model for Problem Solving. *Convergence: The International Journal of Research into New Media Technologies*, 14(1):75–90, February 2008.

[BSMW11] Tara S Behrend, David J Sharek, Adam W Meade, and Eric N Wiebe. The viability of crowdsourcing for survey research. *Behavior research methods*, 43(3):800–813, 2011.

[BSP$^+$14] Rick Bonney, Jennifer L Shirk, Tina B Phillips, Andrea Wiggins, Heidi L Ballard, Abraham J Miller-Rushing, and Julia K Parrish. Next steps for citizen science. *Science*, 343(6178):1436–1437, 2014.

[BvHWRvdB02] Hein Broekkamp, Bernadette HAM van Hout-Wolters, Gert Rijlaarsdam, and Huub van den Bergh. Importance in instructional text: teachers' and students' perceptions of task demands. *Journal of Educational Psychology*, 94(2):260, 2002.

[CCE04] Francisco Cano and María Cardelle-Elawar. An integrated analysis of secondary school students conceptions and beliefs about learning. *European Journal of Psychology of Education*, 19(2):167–187, 2004.

[CD95] Jeanne Sternlicht Chall and Edgar Dale. *Readability revisited: The new Dale-Chall readability formula.* Brookline Books, 1995.

[CDS07] Leah Melani Christian, Don A Dillman, and Jolene D Smyth. Helping respondents get it right the first time: the influence of words, symbols, and graphics in web surveys. *Public Opinion Quarterly*, 71(1):113–125, 2007.

[CGF15] Andrea Ceroni, Ujwal Kumar Gadiraju, and Marco Fisichella. Improving event detection by automatically assessing validity of event occurrence in text. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 1815–1818, 2015.

[CGF17] Andrea Ceroni, Ujwal Gadiraju, and Marco Fisichella. Justevents: A crowdsourced corpus for event validation with strict temporal constraints. In *Advances in Information Retrieval - 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings*, pages 484–492, 2017.

[CGG$^+$14] Andrea Ceroni, Mihai Georgescu, Ujwal Gadiraju, Kaweh Djafari Naini, and Marco Fisichella. Information evolution in wikipedia. In *Proceedings of The International Symposium on Open Collaboration, OpenSym 2014, Berlin, Germany, August 27 - 29, 2014*, pages 24:1–24:10, 2014.

[CGM$^+$16] Andrea Ceroni, Ujwal Gadiraju, Jan Matschke, Simon Wingert, and Marco Fisichella. Where the event lies: Predicting event occurrence in textual documents. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 1157–1160, 2016.

[CHMA10] Lydia B Chilton, John J Horton, Robert C Miller, and Shiri Azenkot. Task search in a human computation market. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 1–9. ACM, 2010.

[cis] Cisco visual networking index: Global mobile data traffic forecast update, 2015-2020. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html. Accessed: 2016-10-01.

[CIT16] Carrie J. Cai, Shamsi T. Iqbal, and Jaime Teevan. Chain reactions: The impact of order on microtask chains. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*, pages 3143–3154, 2016.

[CK13] Dana Chandler and Adam Kapelner. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, 90:123–133, 2013.

[CKLZY12] Georgios Chatzimilioudis, Andreas Konstantinidis, Christos Laoudias, and Demetrios Zeinalipour-Yazti. Crowdsourcing with smartphones. *IEEE Internet Computing*, 16(5):36–44, 2012.

[CKM15] Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. The tool for the automatic analysis of text cohesion (taaco): automatic assessment of local, global, and text cohesion. *Behavior research methods*, pages 1–11, 2015.

[CKT$^+$10] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, et al. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760, 2010.

[CS08] Juliet Corbin and Anselm Strauss. *Basics of qualitative research: Techniques and procedures for developing grounded theory.* Sage, 2008.

[CT14] Kevyn Collins-Thompson. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135, 2014.

[CTC04] Kevyn Collins-Thompson and James P Callan. A language modeling approach to predicting reading difficulty. In *HLT-NAACL*, pages 193–200, 2004.

[CTIB15] Justin Cheng, Jaime Teevan, Shamsi T. Iqbal, and Michael S. Bernstein. Break it down: A comparison of macro- and microtasks. In *SIGHI'15*, 2015.

[CTL01] Mick P Couper, Michael W Traugott, and Mark J Lamias. Web survey design and administration. *Public opinion quarterly*, 65(2):230–253, 2001.

[DBG$^+$13] Stefan Dietze, Jakob Beetz, Ujwal Gadiraju, Georgios Katsimpras, Raoul Wessel, and René Berndt. Towards preservation of semantically enriched architectural knowledge. In *Proceedings of the 3rd International Workshop on Semantic Digital Archives co-located with 17th International Conference on Theory and Practice of Digital Libraries (TPDL 2013), Valetta, Malta, September 26, 2013.*, pages 4–15, 2013.

[DBT05] Tove I Dahl, Margrethe Bals, and Anne Lene Turi. Are students' beliefs about knowledge and learning associated with their reported use of learning strategies? *British journal of educational psychology*, 75(2):257–273, 2005.

[DC49] Edgar Dale and Jeanne S Chall. The concept of readability. *Elementary English*, 26(1):19–26, 1949.

[DC13] Christoph Dukat and Simon Caton. Towards the competence of crowdsourcees: Literature-based considerations on the problem of assessing crowdsourcees' qualities. In *Cloud and Green Computing (CGC), 2013 Third International Conference on*, pages 536–540. IEEE, 2013.

[DCD$^+$15a] D. E. Difallah, M. Catasta, G. Demartini, P. G. Ipeirotis, and P. Cudré-Mauroux. The Dynamics of Micro-Task Crowdsourcing – The Case of Amazon MTurk. In *24th International Conf. on World Wide Web (WWW)*, 2015.

[DCD$^+$15b] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G Ipeirotis, and Philippe Cudré-Mauroux. The dynamics of micro-task crowdsourcing: The case of amazon mturk. In *WWW*, pages 238–247. International World Wide Web Conferences Steering Committee, 2015.

[DCHD$^+$14] Orphée De Clercq, Véronique Hoste, Bart Desmet, Philip Van Oosten, Martine De Cock, and Lieve Macken. Using the crowd for readability prediction. *Natural Language Engineering*, 20(03), 2014.

[DDCM12a] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-scale Entity Linking. In *WWW'12*, pages 469–478, New York, NY, USA, 2012.

[DDCM12b] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. In *CrowdSearch*, pages 26–30, 2012.

[DDCM13] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 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, pages 367–374, New York, NY, USA, 2013. ACM.

[DHL16] Brandon Dang, Miles Hutson, and Matt Lease. Mmmturkey: A crowdsourcing framework for deploying tasks and recording worker behavior on amazon mechanical turk. *arXiv preprint arXiv:1609.00945*, 2016.

[DHS04] David Dunning, Chip Heath, and Jerry M Suls. Flawed self-assessment implications for health, education, and the workplace. *Psychological science in the public interest*, 5(3):69–106, 2004.

[DKB$^+$11] Steven Dow, Anand Kulkarni, Brie Bunge, Truc Nguyen, Scott Klemmer, and Björn Hartmann. Shepherding the crowd: managing and providing feedback to crowd workers. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 1669–1674. ACM, 2011.

[DKKH12] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1013–1022. ACM, 2012.

[DRH11] Anhai Doan, Raghu Ramakrishnan, and Alon Y Halevy. Crowdsourcing systems on the world-wide web. *Communications of the ACM*, 54(4):86–96, 2011.

[Dun11] David Dunning. The dunning-kruger effect: On being ignorant of one's own ignorance. *Advances in experimental social psychology*, 44:247, 2011.

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

[Eag09] Nathan Eagle. txteagle: Mobile crowdsourcing. In *International Conference on Internationalization, Design and Global Development*, pages 447–456. Springer, 2009.

[EAGLdG12] Enrique Estellés-Arolas and Fernando González-Ladrón-de Guevara. Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2):189–200, 2012.

[Eck10] Peter Eckersley. How unique is your web browser? In *Privacy Enhancing Technologies*, pages 1–18. Springer, 2010.

[ED03] Joyce Ehrlinger and David Dunning. How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of personality and social psychology*, 84(1):5, 2003.

[EdV11] Carsten Eickhoff and Arjen de Vries. How crowdsourcable is your task. In *Proceedings of the workshop on crowdsourcing for search and data mining (CSDM) at the fourth ACM international conference on web search and data mining (WSDM)*, pages 11–14, 2011.

[EdV13] Carsten Eickhoff and Arjen P de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval*, 16(2):121–137, 2013.

[EHdVS12] Carsten Eickhoff, Christopher G Harris, Arjen P de Vries, and Padmini Srinivasan. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *Proceedings of SIGIR'12*, pages 871–880. ACM, 2012.

[EJB$^+$08] Joyce Ehrlinger, Kerri Johnson, Matthew Banner, David Dunning, and Justin Kruger. Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational behavior and human decision processes*, 105(1):98–121, 2008.

[FGD14] Besnik Fetahu, Ujwal Gadiraju, and Stefan Dietze. Crawl me maybe: Iterative linked dataset preservation. In *Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, ISWC 2014, Riva del Garda, Italy, October 21, 2014.*, pages 433–436, 2014.

[FGD15] Besnik Fetahu, Ujwal Gadiraju, and Stefan Dietze. Improving entity retrieval on structured data. In *The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part I*, pages 474–491, 2015.

[Fis14] Carmen Fishwick. Tomnod–the online search party looking for malaysian airlines flight mh370. *The Guardian*, 14:37, 2014.

[FLS$^+$15] Oluwaseyi Feyisetan, Markus Luczak-Rösch, Elena Simperl, Ramine Tinati, and Nigel Shadbolt. Towards hybrid NER: A study of content and crowdsourcing-related performance factors. In *Proceedings of ESWC'15*, pages 525–540, 2015.

[FM00] Susan Feinberg and Margaret Murphy. Applying cognitive load theory to the design of web-based instruction. In *Proceedings of IEEE professional communication society international professional communication conference and Proceedings of the 18th annual ACM international conference on Computer documentation: technology & teamwork*, pages 353–360. IEEE Educational Activities Department, 2000.

[FSVKS15a] Oluwaseyi Feyisetan, Elena Simperl, Max Van Kleek, and Nigel Shadbolt. Improving paid microtasks through gamification and adaptive furtherance incentives. In *Proceedings of WWW'15*, pages 333–343, Republic and Canton of Geneva, Switzerland, 2015.

[FSVKS15b] Oluwaseyi Feyisetan, Elena Simperl, Max Van Kleek, and Nigel Shadbolt. Improving paid microtasks through gamification and adaptive furtherance incentives. In *WWW*, pages 333–343. ACM, 2015.

[FZFM17] Leah Findlater, Joan Zhang, Jon E Froehlich, and Karyn Moffatt. Differences in crowdsourced vs. lab-based mobile and desktop input performance data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6813–6824. ACM, 2017.

[FZZ$^+$14] Zhenni Feng, Yanmin Zhu, Qian Zhang, Lionel M Ni, and Athanasios V Vasilakos. Trac: Truthful auction for location-aware collaborative sensing in mobile crowdsourcing. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pages 1231–1239. IEEE, 2014.

[GA08] Qi Guo and Eugene Agichtein. Exploring mouse movements for inferring query intent. In *SIGIR'08*, pages 707–708. ACM, 2008.

[Gal07] Francis Galton. Vox populi (the wisdom of crowds). *Nature*, 75(7):450–451, 1907.

[GCGD17] Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. 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):29, 2017.

[GD17a] Ujwal Gadiraju and Stefan Dietze. Improving learning through achievement priming in crowdsourced information finding microtasks. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference, Vancouver, BC, Canada, March 13-17, 2017*, pages 105–114, 2017.

[GD17b] Ujwal Gadiraju and Stefan Dietze. Improving learning through achievement priming in crowdsourced information finding microtasks. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pages 105–114. ACM, 2017.

[GDD15] Ujwal Gadiraju, Stefan Dietze, and Ernesto Diaz-Aviles. Ranking buildings and mining the web for popular architectural patterns. In *Proceedings of the ACM Web Science Conference, WebSci 2015, Oxford, United Kingdom, June 28 - July 1, 2015*, pages 8:1–8:10, 2015.

[GDDC16] Ujwal Gadiraju, Gianluca Demartini, Djellel Eddine Difallah, and Michele Catasta. It's getting crowded!: how to use crowdsourcing effectively for web science research. In *Proceedings of the 8th ACM Conference on Web Science, WebSci 2016, Hannover, Germany, May 22-25, 2016*, page 11, 2016.

[GFK15a] Ujwal Gadiraju, Besnik Fetahu, and Ricardo Kawase. Training workers for improving performance in crowdsourcing microtasks. In *Design for Teaching and Learning in a Networked World - 10th European Conference on Technology Enhanced Learning, EC-TEL 2015, Toledo, Spain, September 15-18, 2015, Proceedings*, pages 100–114, 2015.

[GFK15b] Ujwal Gadiraju, Besnik Fetahu, and Ricardo Kawase. Training workers for improving performance in crowdsourcing microtasks. In *Proceedings of the 10th European Conference on Technology Enhanced Learning. EC-TEL 2015*, 2015.

[GFK$^+$17] Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(4), 2017.

[GG16] Ujwal Gadiraju and Neha Gupta. Dealing with sub-optimal crowd work: Implications of current quality control practices. In *International Reports on Socio-Informatics (IRSI), Proceedings of the CHI 2016 - Workshop: Crowd Dynamics: Exploring Conflicts and Contradictions in Crowdsourcing*, pages 15–20, 2016.

[GGP10] Rosario Gennaro, Craig Gentry, and Bryan Parno. Non-interactive verifiable computing: Outsourcing computation to untrusted workers. In *Advances in Cryptology–CRYPTO 2010*, pages 465–482. Springer, 2010.

[GK17] Ujwal Gadiraju and Ricardo Kawase. Improving reliability of crowdsourced results by detecting crowd workers with multiple identities. In *Web Engineering - 17th International Conference, ICWE 2017, Rome, Italy, June 5-8, 2017, Proceedings*, pages 190–205, 2017.

[GKD14a] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. Extracting architectural patterns from web data. In *Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, ISWC 2014, Riva del Garda, Italy, October 21, 2014.*, pages 461–464, 2014.

[GKD14b] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. A taxonomy of microtasks on the web. In *25th ACM Conference on Hypertext and Social Media, HT '14, Santiago, Chile, September 1-4, 2014*, pages 218–223, 2014.

[GKD14c] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. A taxonomy of microtasks on the web. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 218–223. ACM, 2014.

[GKD14d] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. A taxonomy of microtasks on the web. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 218–223. ACM, 2014.

[GKDD15a] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*, pages 1631–1640, 2015.

[GKDD15b] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of SIGCHI'15*, pages 1631–1640, 2015.

[GKDD15c] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. Understanding malicious behavior in crowdsourcing platforms: the case of online surveys. In *CHI*, pages 1631–1640. ACM, 2015.

[GL10] Catherine Grady and Matthew Lease. Crowdsourcing document relevance assessment with mechanical turk. In *HLT-NAACL workshop on creating speech and language data with Amazon's mechanical turk*, pages 172–179. Association for ComputationalLinguistics, 2010.

[GMG$^+$16] Snehalkumar Neil S Gaikwad, Durim Morina, Adam Ginzberg, Catherine Mullings, Shirish Goyal, Dilrukshi Gamage, Christopher Diemert, Mathias Burton, Sharon Zhou, Mark Whiting, et al. Boomerang: Rebounding the consequences of reputation feedback on crowdsourcing platforms. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 625–637. ACM, 2016.

[GMHO14] Neha Gupta, David Martin, Benjamin V Hanrahan, and Jacki O'Neill. Turk-life in india. In *Proceedings of the 18th International Conference on Supporting Group Work*, pages 1–11, 2014.

[GMLC04] Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. Coh-metrix: analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202, 2004.

[GMN^{+}15] Ujwal Gadiraju, Sebastian Möller, Martin Nöllenburg, Dietmar Saupe, Sebastian Egger-Lampl, Daniel Archambault, and Brian Fisher. Crowdsourcing versus the laboratory: Towards human-centered experiments using the crowd. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments - Dagstuhl Seminar 15481, Dagstuhl Castle, Germany, November 22-27, 2015, Revised Contributions*, pages 6–26, 2015.

[GNC^{+}14] Ujwal Gadiraju, Kaweh Djafari Naini, Andrea Ceroni, Mihai Georgescu, Dang Duc Pham, and Marco Fisichella. Wikipevent: Temporal event data for the semantic web. In *Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, ISWC 2014, Riva del Garda, Italy, October 21, 2014.*, pages 125–128, 2014.

[GPF^{+}14] Mihai Georgescu, Dang Duc Pham, Claudiu S. Firan, Ujwal Gadiraju, and Wolfgang Nejdl. When in doubt ask the crowd: Employing crowdsourcing for active learning. In *4th International Conference on Web Intelligence, Mining and Semantics (WIMS 14), WIMS '14, Thessaloniki, Greece, June 2-4, 2014*, pages 12:1–12:12, 2014.

[GSAK16] Mary L Gray, Siddharth Suri, Syed Shoaib Ali, and Deepti Kulkarni. The crowd is a collaborative network. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 134–147. ACM, 2016.

[GSD16a] Ujwal Gadiraju, Patrick Siehndel, and Stefan Dietze. Estimating domain specificity for effective crowdsourcing of link prediction and schema mapping. In *Proceedings of the 8th ACM Conference on Web Science, WebSci 2016, Hannover, Germany, May 22-25, 2016*, pages 323–324, 2016.

[GSD16b] Ujwal Gadiraju, Patrick Siehndel, and Stefan Dietze. Estimating domain specificity for effective crowdsourcing of link prediction and schema mapping. In *Proceedings of the 8th ACM Conference on Web Science, WebSci 2016, Hannover, Germany, May 22-25, 2016*, pages 323–324, 2016.

[GSFK15] Ujwal Gadiraju, Patrick Siehndel, Besnik Fetahu, and Ricardo Kawase. Breaking bad: Understanding behavior of crowd workers in categorization microtasks. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media, HT 2015, Guzelyurt, TRNC, Cyprus, September 1-4, 2015*, pages 33–38, 2015.

[GSS+11] David Geiger, Stefan Seedorf, Thimo Schulze, Robert C Nickerson, and Martin Schader. Managing the crowd: Towards a taxonomy of crowdsourcing processes. In *AMCIS*, 2011.

[GTCB12] Aakar Gupta, William Thies, Edward Cutrell, and Ravin Balakrishnan. mclerk: enabling mobile crowdsourcing in developing regions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1843–1852. ACM, 2012.

[Gup17] Neha Gupta. An ethnographic study of crowdwork via amazon mechanical turk in india. 2017.

[GV95] Robert L Glass and Iris Vessey. Contemporary application-domain taxonomies. *IEEE Software*, 12(4):63–76, 1995.

[GYB17] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 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, Prague, Czech Republic, July 4-7, 2017*. ACM, 2017.

[HAB16] Ting-Hao Kenneth Huang, Amos Azaria, and Jeffrey P Bigham. Instructablecrowd: Creating if-then rules via conversations with the crowd. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1555–1562. ACM, 2016.

[Had06] Allison Hadwin. Student task understanding. In *Learning and Teaching Conference. University of Victoria, Victoria, British Columbia, Canada.*, 2006.

[Har06] Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage Publications Sage CA: Los Angeles, CA, 2006.

[Har15] Christopher G. Harris. The effects of pay-to-quit incentives on crowdworker task quality. In *CSCW*, CSCW '15, pages 1801–1812, New York, NY, USA, 2015. ACM.

[HK16] Gary Hsieh and RafałKocielnik. You get who you pay for: The impact of incentives on participation bias. In *CSCW*, CSCW '16, pages 823–835, New York, NY, USA, 2016. ACM.

[Hol79] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[HOMW09] AF Hadwin, M Oshige, M Miller, and P Wild. Examining student and instructor task perceptions in a complex engineering design task. In *international conference on innovation and practices in engineering design and engineering education. McMaster University, Hamilton, ON, Canada*, 2009.

[How06] Jeff Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

[HRZ11] John J Horton, David G Rand, and Richard J Zeckhauser. The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, 14(3):399–425, 2011.

[HSBS13] Tenshi Hara, Thomas Springer, Gerd Bombach, and Alexander Schill. Decentralised approach for a reusable crowdsourcing platform utilising standard web servers. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*, pages 1063–1074. ACM, 2013.

[HSE11] T Hoßfeld, Raimund Schatz, and Sebastian Egger. Sos: The mos is not enough! In *QoMEX*, pages 131–136. IEEE, 2011.

[HT11] Joseph M Hellerstein and David L Tennenhouse. Searching for jim gray: a technical overview. *Communications of the ACM*, 54(7):77–87, 2011.

[HWD11] Jeff Huang, Ryen W. White, and Susan Dumais. No clicks, no problem: Using cursor movements to understand and improve search. In *SIGCHI'11*, pages 1225–1234, New York, NY, USA, 2011. ACM.

[Ipe10a] Panagiotis G Ipeirotis. Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, 17(2):16–21, 2010.

[Ipe10b] Panagiotis G Ipeirotis. Demographics of mechanical turk. 2010.

[IPW10] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67. ACM, 2010.

[Ira15] Lilly Irani. The cultural work of microwork. *New Media & Society*, 17(5):720–739, 2015.

[IS13] Lilly C Irani and M Silberman. Turkopticon: interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 611–620. ACM, 2013.

[JN04] Diane Lee Jamieson-Noel. *Exploring task definition as a facet of self-regulated learning*. PhD thesis, Faculty of Education-Simon Fraser University, 2004.

[JRP$^+$16] Matt Jones, Simon Robinson, Jennifer Pearson, Manjiri Joshi, Dani Raju, Charity Chao Mbogo, Sharon Wangari, Anirudha Joshi, Edward Cutrell, and Richard Harper. Beyond yesterdays tomorrow: future-focused mobile interaction design by and for emergent users. *Personal and Ubiquitous Computing*, pages 1–15, 2016.

[JWN15] Ling Jiang, Christian Wagner, and Bonnie Nardi. Not just in it for the money: A qualitative investigation of workers' perceived benefits of micro-task crowdsourcing. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, pages 773–782. IEEE, 2015.

[Kan11] Salil S Kanhere. Participatory sensing: Crowdsourcing data from mobile smartphones in urban spaces. In *Mobile Data Management (MDM), 2011 12th IEEE International Conference on*, volume 2, pages 3–6. IEEE, 2011.

[Kaz11] Gabriella Kazai. In search of quality in crowdsourcing for search engine evaluation. In *Advances in information retrieval*, pages 165–176. Springer, 2011.

[KBK$^+$12] Michal Kosinski, Yoram Bachrach, Gjergji Kasneci, Jurgen Van-Gael, and Thore Graepel. Crowd iq: Measuring the intelligence of crowdsourcing platforms. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 151–160. ACM, 2012.

[KCS08] Aniket Kittur, Ed H Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM, 2008.

[KD99] Justin Kruger and David Dunning. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121, 1999.

[KDC+11] Firas Khatib, Frank DiMaio, Seth Cooper, Maciej Kazmierczyk, Miroslaw Gilski, Szymon Krzywda, Helena Zabranska, Iva Pichova, James Thompson, Zoran Popović, et al. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology*, 18(10):1175–1177, 2011.

[KFJRC75] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document, 1975.

[KKMF11] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1941–1944. ACM, 2011.

[KKMF12] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 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*, pages 2583–2586. ACM, 2012.

[KKMF13a] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Inf. Retr.*, 16(2):138–178, 2013.

[KKMF13b] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information retrieval*, 16(2):138–178, 2013.

[KLP+10] Rohit J Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J Mooney, Salim Roukos, and Chris Welty. Learning to predict readability using diverse linguistic features. In *ACL*, pages 546–554. Association for Computational Linguistics, 2010.

[KM02] Joachim Krueger and Ross A Mueller. Unskilled, unaware, or both? the better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of personality and social psychology*, 82(2):180, 2002.

[KNB+13] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1301–1318. ACM, 2013.

[KO08] Marian Krajc and Andreas Ortmann. Are the unskilled really that unaware? an alternative explanation. *Journal of Economic Psychology*, 29(5):724–738, 2008.

[KRDT10] Shashank Khanna, Aishwarya Ratan, James Davis, and William Thies. Evaluating and improving the usability of mechanical turk for low-income workers in india. In *DEV*, page 12. ACM, 2010.

[KRF$^+$17] Vassilis Kostakos, Jakob Rogstadius, Denzil Ferreira, Simo Hosio, and Jorge Goncalves. Human sensors. In *Participatory Sensing, Opinions and Collective Awareness*, pages 69–92. Springer, 2017.

[KSG14] Ricardo Kawase, Patrick Siehndel, and Ujwal Gadiraju. Technology enhancing learning: Past, present and future. In *Open Learning and Teaching in Educational Communities - 9th European Conference on Technology Enhanced Learning, EC-TEL 2014, Graz, Austria, September 16-19, 2014, Proceedings*, pages 193–206, 2014.

[KSV11a] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. More than fun and money. worker motivation in crowdsourcing - a study on mechanical turk. In *AMCIS*, 2011.

[KSV11b] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. More than fun and money. worker motivation in crowdsourcing–a study on mechanical turk. 2011.

[KWL$^+$15] Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R Klemmer. Peer and self assessment in massive online classes. In *Design Thinking Research*, pages 131–168. Springer, 2015.

[KZ16] Gabriella Kazai and Imed Zitouni. Quality management in crowdsourcing using gold judges behavior. In *WSDM*, pages 267–276, 2016.

[LCGM09] Greg Little, Lydia B Chilton, Max Goldman, and Robert C Miller. Turkit: tools for iterative tasks on mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 29–30. ACM, 2009.

[LEHB10] John Le, Andy Edmonds, Vaughn Hester, and Lukas Biewald. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*, pages 21–26, 2010.

[LF04] Matthew Linderman and Jason Fried. *Defensive Design for the Web: How to improve error messages, help, forms, and other crisis points*. New Riders Publishing, 2004.

[LL06] Jerome P Lynch and Kenneth J Loh. A summary review of wireless sensors and sensor networks for structural health monitoring. *Shock and Vibration Digest*, 38(2):91–130, 2006.

[LLT01] Lieve Luyten, Joost Lowyck, and Francis Tuerlinckx. Task perception as a mediating variable: A contribution to the validation of instructional knowledge. *British Journal of Educational Psychology*, 71(2):203–223, 2001.

[LLW$^+$15] Gierad Laput, Walter S Lasecki, Jason Wiese, Robert Xiao, Jeffrey P Bigham, and Chris Harrison. Zensors: Adaptive, rapidly deployable, human-intelligent sensor feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1935–1944. ACM, 2015.

[LOYT14] Mounia Lalmas, Heather L O'Brien, and Elad Yom-Tov. Measuring user engagement. *Morgan and Claypool Publishers*, 2014.

[MA13] Marija Milenkovic and Oliver Amft. An opportunistic activity-sensing approach to save energy in office buildings. In *Proceedings of the fourth international conference on Future energy systems*, pages 247–258. ACM, 2013.

[MBB$^+$17] Meredith Ringel Morris, Jeffrey P Bigham, Robin Brewer, Jonathan Bragg, Anand Kulkarni, Jessie Li, and Saiph Savage. Subcontracting microwork. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, page To Appear. ACM, 2017.

[MCNL16] Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. 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*, pages 2271–2282. ACM, 2016.

[MG11] Tyler M Miller and Lisa Geraci. Training metacognition in the classroom: the influence of incentives and feedback on exam predictions. *Metacognition and Learning*, 6(3):303–314, 2011.

[MHOG14] David Martin, Benjamin V Hanrahan, Jacki O'Neill, and Neha Gupta. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 224–235. ACM, 2014.

[MOGH16a] David Martin, Jacki O'Neill, Neha Gupta, and Benjamin V. Hanrahan. Turking in a global labour market. *Computer Supported Cooperative Work (CSCW)*, 25(1):39–77, 2016.

[MOGH16b] David Martin, Jacki ONeill, Neha Gupta, and Benjamin V Hanrahan. Turking in a global labour market. *Computer Supported Cooperative Work (CSCW)*, 25(1):39–77, 2016.

[MOR$^+$16] Róisín McNaney, Mohammad Othman, Dan Richardson, Paul Dunphy, Telmo Amaral, Nick Miller, Helen Stringer, Patrick Olivier, and John Vines. Speeching: Mobile crowdsourced speech assessment to support self-monitoring and management for people with parkinsons. In *Proceedings of the 2016 ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 7–12, 2016.

[MR12] David Malvern and Brian Richards. Measures of lexical richness. *The Encyclopedia of Applied Linguistics*, 2012.

[MS12] Winter Mason and Siddharth Suri. Conducting behavioral research on amazons mechanical turk. *Behavior research methods*, 44(1):1–23, 2012.

[MS13a] Catherine C. Marshall and Frank M. Shipman. Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, pages 234–243, New York, NY, USA, 2013. ACM.

[MS13b] Catherine C Marshall and Frank M Shipman. Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 234–243. ACM, 2013.

[MW10] Winter Mason and Duncan J Watts. Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter*, 11(2):100–108, 2010.

[NGR$^+$11] Prayag Narula, Philipp Gutheim, David Rolnitzky, Anand Kulkarni, and Bjoern Hartmann. Mobileworks: A mobile crowdsourcing platform for workers at the bottom of the pyramid. *Human Computation*, 11:11, 2011.

[NR16] Edward Newell and Derek Ruths. How one microtask affects another. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*, pages 3155–3166, 2016.

[NTDM14] Klimis Ntalianis, Nicolas Tsapatsoulis, Anastasios Doulamis, and Nikolaos Matsatsinis. Automatic annotation of image databases based on implicit crowdsourcing, visual concept modeling and evolution. *Multimedia Tools and Applications*, 69(2):397–421, 2014.

[NTG+16] Michael Nebeling, Alexandra To, Anhong Guo, Adrian A de Freitas, Jaime Teevan, Steven P Dow, and Jeffrey P Bigham. Wearwrite: Crowd-assisted writing from smartwatches. In *Proceedings of CHI*, 2016.

[OSL+11] David Oleson, Alexander Sorokin, Greg P Laughlin, Vaughn Hester, John Le, and Lukas Biewald. Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. *Human computation*, 11:11, 2011.

[Ovi06] Sharon Oviatt. Human-centered design meets cognitive load theory: designing interfaces that help people think. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 871–880. ACM, 2006.

[PBSA17] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163, 2017.

[PCI10] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running experiments on amazon mechanical turk. 2010.

[PE10] Gabriel Parent and Maxine Eskenazi. Toward better crowdsourced transcription: Transcription of a year of the let's go bus information system data. In *Spoken Language Technology Workshop (SLT), 2010 IEEE*, pages 312–317. IEEE, 2010.

[PN08] Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *EMNLP*, pages 186–195. Association for Computational Linguistics, 2008.

[QB11] Alexander J Quinn and Benjamin B Bederson. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1403–1412. ACM, 2011.

[RIS+10] Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 2863–2872. ACM, 2010.

[RK11] Jeffrey M Rzeszotarski and Aniket Kittur. 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*, pages 13–22. ACM, 2011.

[RK12] Jeffrey Rzeszotarski and Aniket Kittur. Crowdscape: interactively visualizing user behavior and output. In *UIST'12*, pages 55–62. ACM, 2012.

[RN77] Libby O Ruch and Rae R Newton. Sex characteristics, task clarity, and authority. *Sex Roles*, 3(5):479–494, 1977.

[RPH$^+$07] Sasank Reddy, Andrew Parker, Josh Hyman, Jeff Burke, Deborah Estrin, and Mark Hansen. Image browsing, processing, and clustering for participatory sensing: lessons from a dietsense prototype. In *Proceedings of the 4th workshop on Embedded networked sensors*, pages 13–17. ACM, 2007.

[RR15] Presentacion Rivera-Reyes. Students' task interpretation and conceptual understanding in electronics laboratory work. 2015.

[RZS15] Markus Rokicki, Sergej Zerr, and Stefan Siersdorfer. Groupsourcing: Team competition designs for crowdsourcing. In *Proceedings of WWW'15*, pages 906–915, Republic and Canton of Geneva, Switzerland, 2015.

[RZS16] Markus Rokicki, Sergej Zerr, and Stefan Siersdorfer. Just in time: Controlling temporal performance in crowdsourcing competitions. In *WWW*, pages 817–827, 2016.

[SBDJ16] Tobias Schnabel, Paul N Bennett, Susan T Dumais, and Thorsten Joachims. Using shortlists to support decision making and improve recommender system performance. In *Proceedings of the 25th International Conference on World Wide Web*, pages 987–997. International World Wide Web Conferences Steering Committee, 2016.

[SC94] John Sweller and Paul Chandler. Why some material is difficult to learn. *Cognition and instruction*, 12(3):185–233, 1994.

[Sch04] Barry Schwartz. The paradox of choice: Why less is more. *New York: Ecco*, 2004.

[SD10] Matthias Stevens and Ellie DHondt. Crowdsourcing of pollution data using smartphones. In *Workshop on Ubiquitous Crowdsourcing*, 2010.

[SDJK13] Thomas Schlösser, David Dunning, Kerri L Johnson, and Justin Kruger. How unaware are the unskilled? empirical tests of the signal extraction counterexplanation for the dunning–kruger effect in self-evaluation of performance. *Journal of Economic Psychology*, 39:85–100, 2013.

[SG16] Patrick Siehndel and Ujwal Gadiraju. Unlock the stock: User topic modeling for stock market analysis. In *Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference, EDBT/ICDT Workshops 2016, Bordeaux, France, March 15, 2016.*, 2016.

[SHC11] Aaron D Shaw, John J Horton, and Daniel L Chen. Designing incentives for inexpert human raters. In *CSCW*, pages 275–284. ACM, 2011.

[SL13] Aashish Sheshadri and Matthew Lease. SQUARE: A Benchmark for Research on Computing Crowd Consensus. In *Proceedings of HCOMP'13*, pages 156–164, 2013.

[SR14] Parnia Samimi and Sri Devi Ravana. Creation of reliable relevance judgments in information retrieval systems evaluation experimentation through crowdsourcing: a review. *The Scientific World Journal*, 2014, 2014.

[SSK$^+$16] Oliver S Schneider, Hasti Seifi, Salma Kashani, Matthew Chun, and Karon E MacLean. Hapturk: Crowdsourcing affective ratings of vibrotactile icons. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3248–3260. ACM, 2016.

[Str87] Anselm L Strauss. *Qualitative analysis for social scientists.* Cambridge University Press, 1987.

[Sur05] James Surowiecki. *The wisdom of crowds.* Anchor, 2005.

[SW04] Barry Schwartz and Andrew Ward. Doing better but feeling worse: The paradox of choice. *Positive psychology in practice*, pages 86–104, 2004.

[Swe88] John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285, 1988.

[Tar02] Maddalena Taras. Using assessment for learning and learning from assessment. *Assessment & Evaluation in Higher Education*, 27(6):501–510, 2002.

[TNK+15] Tuan A. Tran, Claudia Niederée, Nattiya Kanhabua, Ujwal Gadiraju, and Avishek Anand. Balancing novelty and salience: Adaptive learning to rank entities for timeline summarization of high-impact events. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 1201–1210, 2015.

[TON+14] Rannie Teodoro, Pinar Ozturk, Mor Naaman, Winter Mason, and Janne Lindqvist. The motivations and experiences of the on-demand mobile workforce. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 236–247. ACM, 2014.

[VA08] Luis Von Ahn. Human computation. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 1–2. IEEE, 2008.

[VAD04] Luis Von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM, 2004.

[VAD08] Luis Von Ahn and Laura Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67, 2008.

[VDV12] Jeroen BP Vuurens and Arjen P De Vries. Obtaining high-quality relevance judgments using crowdsourcing. *Internet Computing, IEEE*, 16(5):20–27, 2012.

[VGK+14a] Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. Community-based bayesian aggregation models for crowdsourcing. In *WWW'14*, pages 155–164, New York, NY, USA, 2014. ACM.

[VGK+14b] Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. Community-based bayesian aggregation models for crowdsourcing. In *WWW*, pages 155–164, 2014.

[VKG10] Maja Vukovic, Soundar Kumara, and Ohad Greenshpan. Ubiquitous crowdsourcing. In *Proceedings of the 12th ACM international conference adjunct papers on Ubiquitous computing-Adjunct*, pages 523–526. ACM, 2010.

[VVC+14] Annamalai Vasantha, Gokula Vijayumar, Jonathan Corney, Nuran Acur Bakir, Andrew Lynn, Ananda Prasanna Jagadeesan, Marisa Smith, and Anupam Agarwal. Social implications of crowdsourcing in rural scotland. *International Journal of Social Science & Human Behavior Study*, 1(3):47–52, 2014.

[WHA12] Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. Strategies for crowdsourcing social data analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 227–236, New York, NY, USA, 2012. ACM.

[Whi13] Kathryn Whitenton. Minimize cognitive load to maximize usability. *Pozyskano*, 4:2014, 2013.

[WIP11] Jing Wang, Panagiotis G Ipeirotis, and Foster Provost. Managing crowdsourcing workers. In *WCBI'11*, 2011.

[WK15] Yuhui Wang and Mohan S Kankanhalli. Tweeting cameras for event detection. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1231–1241. ACM, 2015.

[Wro08] Luke Wroblewski. *Web form design: filling in the blanks.* Rosenfeld Media, 2008.

[WWZ$^+$12] Gang Wang, Christo Wilson, Xiaohan Zhao, Yibo Zhu, Manish Mohanlal, Haitao Zheng, and Ben Y Zhao. Serf and turf: crowdturfing for fun and profit. In *Proceedings of the 21st international conference on World Wide Web*, pages 679–688. ACM, 2012.

[WYL$^+$14] Qiuzhen Wang, Sa Yang, Manlu Liu, Zike Cao, and Qingguo Ma. An eye-tracking study of website complexity from cognitive load perspective. *Decision support systems*, 62:1–10, 2014.

[XW15] Xiao-Feng Xie and Zun-Jing Wang. An empirical study of combining participatory and physical sensing to better understand and improve urban mobility networks. In *Transportation Research Board 94th Annual Meeting*, number 15-3238, 2015.

[YFGD16] Ran Yu, Besnik Fetahu, Ujwal Gadiraju, and Stefan Dietze. A survey on challenges in web markup data for entity retrieval. In *Proceedings of the ISWC 2016 Posters & Demonstrations Track co-located with 15th International Semantic Web Conference (ISWC 2016), Kobe, Japan, October 19, 2016.*, 2016.

[YGFD15] Ran Yu, Ujwal Gadiraju, Besnik Fetahu, and Stefan Dietze. Adaptive focused crawling of linked data. In *Web Information Systems Engineering - WISE 2015 - 16th International Conference, Miami, FL, USA, November 1-3, 2015, Proceedings, Part I*, pages 554–569, 2015.

[YGZ+16] Ran Yu, Ujwal Gadiraju, Xiaofei Zhu, Besnik Fetahu, and Stefan Dietze. Towards entity summarisation on structured web markup. In *The Semantic Web - ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29 - June 2, 2016, Revised Selected Papers*, pages 69–73, 2016.

[YHBH14] Jie Yang, Claudia Hauff, Alessandro Bozzon, and Geert-Jan Houben. Asking the right question in collaborative q&a systems. In *Hypertext*, pages 179–189. ACM, 2014.

[YKL11] Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. A survey of crowdsourcing systems. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 766–773. IEEE, 2011.

[YMH+09] Tingxin Yan, Matt Marzilli, Ryan Holmes, Deepak Ganesan, and Mark Corner. mcrowd: a platform for mobile crowdsourcing. In *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, pages 347–348. ACM, 2009.

[YRDB16] Jie Yang, Judith Redi, Gianluca Demartini, and Alessandro Bozzon. Modeling task complexity in crowdsourcing. In *HCOMP*, pages 249–258. AAAI, 2016.

[YSMA12] Han Yu, Zhiqi Shen, Chunyan Miao, and Bo An. Challenges and opportunities for trust management in crowdsourcing. In *2012 IEEE/WIC/ACM International Conferences on Intelligent Agent Technology, IAT 2012, Macau, China, December 4-7, 2012*, pages 486–493, 2012.

[ZBC+14] Andrea Zanella, Nicola Bui, Angelo Castellani, Lorenzo Vangelista, and Michele Zorzi. Internet of things for smart cities. *IEEE Internet of Things journal*, 1(1):22–32, 2014.