



University of
Strathclyde
Glasgow

**Effective Task Design to Improve Worker Performance and
Engagement in Crowdsourcing Image-Labelling Tasks**

By
Huda Umar Banuqitah

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Computer Science

Department of Computer and Information Science

University of Strathclyde

2026

Declaration

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.50. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Signed: *huda banugitah*

Date: 2026-9-02

Acknowledgment

I wish to convey my profound gratitude and appreciation to everybody who has helped me during my doctoral studies and the completion of this thesis. This study's work would not have been possible without the support of numerous individuals. I am profoundly thankful to my supervisors, Dr. Mark Dunlop (University of Strathclyde), Dr. Sotirios Terzis (University of Strathclyde), and Dr. Maysoon F. Abulkhair (King Abdulaziz University), for their direction, wisdom, and steadfast support. Their experience, patience, and passion have been important in refining this research and improving its quality. I am genuinely privileged to have received their guidance during this process.

I would like to convey my profound appreciation to my colleagues who participated in the experiment and survey, contributing their time, insights, and experience. Their significant contributions were essential to the completion of this research, and I sincerely value their willingness to participate. I would also like to extend my sincere thanks to Entisar Alkayal for her valuable contribution as an external expert in validating the paper classification process, which strengthened the reliability of the research findings.

I extend my profound appreciation to my father, Umar, whose love and support have been a continual source of motivation and encouragement in my academic pursuits. I extend my gratitude to my sisters (Hana, Hanadi and Nuha) and my brothers (Mohammed, Ahmed, Saeed and Salem) for their unwavering support and presence. They consistently provided support and encouragement through their best wishes. I wish to express my profound gratitude to my devoted husband, Abdulrahman Bahabri, for his steadfast support, patience, and understanding over this entire journey. His faith in my capabilities, his support, and his sacrifices has been the foundation of my success. I would also like to convey my profound gratitude to my three exceptional sons (Abdullah, Aseel and Saba) and my sweet daughter (Jouri). Their unwavering love, comprehension, and adaptability have consistently served as a source of inspiration and drive. Their forbearance throughout my extended study hours and their comprehension of the sacrifices necessitated by my academic pursuit have profoundly moved me. I appreciate their assistance and the happiness they contribute to my existence.

Abstract

Nowadays, crowdsourcing has become a baseline approach for large-scale data annotation, yet maintaining workers' engagement, and outcome quality in repetitive tasks remains a persistent challenge. Current systems often rely on individual workers, offer minimal social interaction and apply gamification in a fragmented manner without a systematic mechanism to fulfil motivational needs. Consequently, there remains limited understanding of how collaboration combined with gamification could improve worker performance, and engagement in crowdsourcing environments. This thesis argues that effective and high-quality crowdsourcing requires a shift from an isolated, incentive-driven task execution paradigm to a collaborative socially embedded, and motivation-aligned system design. It investigates how collaborative gamification (e.g., teamwork, real-time interactions and motivational design components) can enhance performance and engagement in image-labelling tasks. The work is grounded in Self Determination Theory (SDT) and Flow Theory, which provide a theoretical foundation for understanding how the system design influences worker motivation and behaviour. To address this, the thesis adopts a multi-phase explanatory sequential framework research programme that incorporates three interrelated studies. First, an exploratory study empirically investigates how workers' engagement and performance differ according to the crowdsourcing task types. Second, a systematic scoping review examines how gamification features have been applied across crowdsourcing archetypes, identifying the most popular design strategies, the underpinning theory, results and gaps in the literature. Third, a real-time collaborative gamified image-labelling platform is designed, developed, and evaluated through four controlled between-subjects conditions to assess how collaboration and gamification affect worker performance and engagement, and their perceived workload. The thesis makes four primary contributions: 1) it establishes empirical baselines of worker behaviour, engagement, and performance across different crowdsourcing task types, 2) it provides a systematic scoping review that maps and summarises current studies gamification features and the motivational theories across crowdsourcing archetypes, 3) it presents a custom-built platform for real-time collaborative gamified crowdsourcing with a robust evaluation pipeline, and 4) it provides large-scale experimental evidence that collaborative gamification improves worker performance (operationalised as labelling accuracy and workers' agreement strength), and engagement while reducing perceived workload in image-labelling task compared to traditional individual approaches. Collectively, these contributions advance our understanding of motivational design in crowdsourcing and offer practical direction for the future of human-centred crowdsourcing, and human-AI applications.

Contents

Chapter 1 : Introduction	1
1.1 Background and Motivation.....	1
1.2 Research Objectives	2
1.3 Thesis Statement.....	3
1.4 Research Questions	3
1.5 Summary of Key Findings	4
1.6 The Significance and Contributions of the Study	5
1.7 Thesis Structure.....	7
Chapter 2 : Background and Conceptual Foundations	9
2.1 Introduction	9
2.2 Literature strategy.....	9
2.3 Foundation of Crowdsourcing and Human Computation in ML/AI	10
2.4 Crowdsourcing Components	13
2.5 Motivational Mechanisms	17
2.6 Theoretical lens	22
2.7 Conclusion.....	24
Chapter 3 : Research Methodology.....	27
3.1 Introduction	27
3.2 Research paradigms and methodologies	27
3.3 Methodological Framework: Multi-Phase Mixed Methods Research.	29
3.4 Methods for Data Collection and Analysis Instruments	30
3.5 Integration and Consistency	34
3.6 Ethics Consideration.....	35
3.7 Conclusion.....	35
Chapter 4 : Exploratory Study of Worker Engagement in Different Crowdsourcing Tasks	37
4.1 Introduction	37
4.2 Study Objective	37
4.3 Study Design	38
4.4 Experiments Setup.....	40
4.5 Participants and Procedures.....	43
4.6 Datasets	43

4.7	Evaluation metrics.....	45
4.8	Results	49
4.9	Discussion	59
4.10	Conclusion.....	61
Chapter 5 : Scoping Review of Gamified Tasks in Crowdsourcing.....		63
5.1	Introduction	63
5.2	Scoping Review Objective and Research Questions.....	63
5.3	Scoping Review Methodological Approach.....	64
5.4	Scoping Review Results.....	66
5.5	Discussions.....	95
5.6	Conclusion.....	99
Chapter 6 : Collaborative Gamified Image-labelling Task Design.....		101
6.1	Introduction	101
6.2	Study Objective	102
6.3	Design Rationale	105
6.4	Prototype Architectural Framework Components and Implementation.....	108
6.5	Prototype frameworks	110
6.6	MTurk Integration	111
6.7	Workflow Logic in Real Time	112
6.8	Reduced-Feature Experimental Conditions.....	116
6.9	Dataset.....	120
6.10	Pilot studies	120
6.11	Conclusion.....	121
Chapter 7 : Experimental Evaluation of the Collaborative Gamified Image-labelling Task.....		122
7.1	Introduction	122
7.2	Overview of Analysis and Result Organisation	122
7.3	Experimental Conditions.....	123
7.4	Sample Size Determination and Power Analysis	124
7.5	Participant Recruitment and Qualification Criteria.....	125
7.6	Evaluation Measures	126
7.7	Evaluation Metrics and Analysis Plan.....	130
7.8	Results	131
7.9	Discussion	145
7.10	Threats to Validity.....	146
7.11	Conclusion.....	148

Chapter 8 : Discussion.....	150
8.1 Introduction	150
8.2 Synthesis Across the Three Studies.....	150
8.3 Contributions to knowledge and Practice.....	153
8.4 Limitation and Methodological Reflections.....	155
8.5 Conclusion.....	157
Chapter 9 : Conclusion and Future Work.....	159
9.1 Overview of the Research	159
9.2 Thesis Research Questions.....	160
9.3 Summary of Key Findings	161
9.4 Thesis Contributions.....	161
9.5 Future Work	163
9.6 Thesis Conclusion	164
References	166
Appendix A 192	
A.1 Ethics Approval for Study 1	192
A.2 Users ‘Consent Form of Study 1	194
A.3 Survey Form of Study 1	195
A.4 Demographic Details of Participants in Study 1	203
Appendix B 206	
B.1 Ethics Approval for Study 3	206
B.2 Consent Form for Study 3.....	209
B.3 Scripts of Publishing the Task on MTurk.....	210
B.4 Task Instruction of Full-Feature Condition in Study 3.....	211
B.5 Task Instruction of Chat-Only Condition in Study 3.....	212
B.6 Task Instruction of No-Element Condition in Study 3	213
B.7 Task Instruction of Individual Condition in Study 3	214
B.8 Demographic Details of Participants in Study 3.....	215
B.9 The scoring and compensation logic.	219
B.10: Survey of Study 3	222

List of Figures

Figure 1.1: Overview of the research framework.....	6
Figure 1.2: Mapping of research questions (RQ#), objectives (Obj#), contributions (C#), and corresponding chapters	8
Figure 2.1: Flow diagram for literature review reading and writing	10
Figure 2.2: Adding Machine Learning with Human Computation as a means of solving computational problems, adapted from (Quinn and Bederson, 2011)	11
Figure 2.3: Crowdsourcing categories adapted from Geiger et al.(2012)	15
Figure 3.1: The interconnection of worldviews, design, and research method (Creswell and Creswell, 2017).....	28
Figure 3.2: Integration of the pragmatic paradigm with the multi-phase mixed methods research design adopted in this thesis.	31
Figure 4.1: Tweet classification task	41
Figure 4.2: Dog classification task	42
Figure 4.3: Survey task.....	42
Figure 4.4: Gender distribution in all tasks	50
Figure 4.5: Age distribution in all tasks	50
Figure 4.6: Education level distribution in all tasks	51
Figure 4.7: Employment status distribution across all tasks.....	51
Figure 4.8: Distribution of task load.....	52
Figure 4.9: Engagement level distribution.....	53
Figure 4.10: Distribution of personality traits	54
Figure 4.11: Rank of workers' task selection criteria	55
Figure 4.12: The percentage of workers with a certain accuracy rate in the tweet classification task	57
Figure 4.13: The percentage of workers with a certain accuracy rate in the dog classification task	58
Figure 5.1: Scoping review PRISMA flow-chart process	66
Figure 5.2: Confusion matrix between raters	69
Figure 5.3: Cooperative vs competitive across all archetypes.....	84
Figure 5.4: Achievement element usage across all archetypes	84
Figure 5.5: Immersion element usage across archetypes.....	85

Figure 5.6: Social Gamification elements across all archetypes	85
Figure 5.7:Heatmap of gamification elements by archetype	90
Figure 6.1:Conceptual framework for the design and implementation of the collaborative gamified image-labeling task	103
Figure 6.2: Prototype system architectural framework.....	111
Figure 6.4:Consent form.....	113
Figure 6.3: HITs Dashboard in Amazon Mechanical Turk	113
Figure 6.5: Registration page.....	114
Figure 6.6: Full-Feature interface	115
Figure 6.7: Level-End feedback message	117
Figure 6.8: End of task message.....	117
Figure 6.9: Chat-only condition interface.....	118
Figure 6.10:No-Elements condition interface.....	119
Figure 6.11: Individual control-condition interface.....	119
Figure 7.1:F-Distribution Power Analysis for One-Way ANOVA (Cohen, 1992).....	124
Figure 7.2:Boxplot of Majority Accuracy Across Conditions.....	133
Figure 7.3: Agreement strength	134
Figure 7.4:Average accuracy of workers across conditions	136
Figure 7.5:Histograms of average accuracy in Full-Features experimental conditions	136
Figure 7.6:Histograms of average accuracy in Chat-Only experimental conditions.....	137
Figure 7.7:Histograms of average accuracy in No-Elements experimental conditions.....	137
Figure 7.8:Histograms of average accuracy in Individual experimental conditions	138
Figure 7.9: NASA-TLX distribution across conditions.....	141
Figure 7.10:Average Likert ratings for gamification elements across team-based conditions.	143

List of Tables

Table 2.1: Crowdsourcing archetypes	16
Table 2.2: Motivational mechanisms in crowdsourcing.....	23
Table 2.3:Alignment of SDT and Flow constructs with motivational mechanisms.....	24
Table 4.1:Confusion matrix for positive and negative labels	45
Table 4.2:Mapping of Sub_Research questions to results sections	48
Table 4.3:Workers' workload in different tasks	56
Table 4.4:Workers' personality traits in different tasks	58
Table 5.1: Classification of reviewed studies	70
Table 5.2: Participant Types in the Reviewed Studies	72
Table 5.3:Gamification elements distribution across reviewed studies	76
Table 5.4:Behavioural and psychological outcomes on gamified crowdsourcing types.....	91
Table 5.5:Implemented theories based on different crowdsourcing types.	94
Table 6.1: Features Design mapping across theoretical constructs, design rationale.....	108
Table 7.1: Mapping of Sub-Research Questions to Measures and Result Sections	123
Table 7.2: Experimental conditions and Gamification Feature	123
Table 7.3:Confusion matrix for positive and negative labels	126
Table 7.4:Engagement Measures.....	129
Table 7.5:NASA-TLX Measures.....	131
Table 7.6:Tukey's Post-Hoc and Rank Biserial Correlation Analysis for Majority Accuracy.....	133
Table 7.7:Ratio of different majorities analysis:	135
Table 7.8: Statistical Analysis of Majority-Decision Differences Across Conditions	135
Table 7.9: Overall Engagements Across all Conditions	139
Table 7.10: Average NASA-TLX Constructs Across Conditions	141
Table 7.13:Chat messages analysis.....	144

Chapter 1 : Introduction

1.1 Background and Motivation

Artificial Intelligence (AI) systems have heavily relied on large quantities of high-quality datasets for training and validation. Even with rapid development in automation, human efforts remain critical in the supply, labelling, and validation of the data. As such, human-AI collaboration stands as the pillar of modern machine learning (ML), given the necessity of human cognition as a complement to computational efficiency. However, the need to maintain the accuracy, attentiveness, and motivation of human participants engaged in repetitive annotation tasks has been a significant challenge.

Human computation and crowdsourcing bring together large numbers of humans to perform data-centric work, spreading the load out across a diverse pool within the online community. Although these paradigms have led to impressive results, such as the large-scale annotation efforts behind landmark datasets, they also reveal human-related issues of lack of motivation and fatigue and uneven participation. Empirical studies have demonstrated that a small number of workers finish most tasks, while vague or monotonous work may cause the participants to leave and this results in inconsistent quality. These problems are particularly harmful to the reliability of machine-learning models, which heavily rely on human-labelled data.

Crowdsourcing platforms stem workers from diverse cultural, educational, and motivational sources. Such diversity ostensibly mitigates data generation bias; however, it also presents myriad challenges in creating incentives that motivate participation and reliability on their outcome. The effectiveness data quality and motivation among workers is a primary challenge in digital labour systems.

The workers' engagement is generally considered an important factor in supporting motivated participation in online collective labour. It includes behavioural effort, emotional engagement, and cognitive investment; however, the antecedents are not well explored in distributed settings at a large scale. Motivation is what makes people want to engage, and it can come from both extrinsic rewards (for example, payment or recognition) and intrinsic motivations (the experience of participating because one is curious, enjoys the activity, wants to give back as an act of altruism, or experience social engagement). The traditional crowdsourcing mechanism, in which the main incentive provided is monetary reward, frequently proves inadequate to keep users' performance quality continuously engaged and satisfied.

Recent studies have revealed an imbalance in the motivational design of crowdsourcing systems by which most existing implementations focus on competitive features, like contests, points, and leaderboards, whereas knowledge regarding collaborative mechanisms is relatively limited (Humayun et al., 2023). This excessive weighting to competition fails to enable the development of engagement models that could leverage cooperation, and intrinsic fulfilment.

Gamification, the use of game design elements in non-game contexts, was identified as a serious (potential) candidate to enhance data quality, engagement, and enjoyment (Hamari et al., 2014, Huotari and Hamari, 2017, Seaborn and Fels, 2015). However, the majority of the existing designs limit themselves to a narrow view on simple competitive elements (like points and badges) (Morschheuser et al., 2016); collaborative gamification, where competing users contribute collaboratively to joint results instead, is largely under-investigated (Riar et al., 2022). Filling this gap is crucial for the design and construction of high-quality human-in-the-loop systems (HITL) and engageable, that cater to the increasing needs of AI and ML.

In this context, this thesis examines the potential of collaborative gamification to improve worker performance and engagement in crowdsourcing image-labelling problems. It is based on SDT and Flow Theory and focuses on how gamified collaborative elements such as chat, feedback, scoring, leaderboard, and sharing goals, levels can allow one to fulfil the psychological needs for autonomy, competence, and relatedness while maintaining a balance between challenge and skill. By targeting teams with real-time labelling experiences, this work seeks to design, deploy, and assess collaborative gamified systems that encourage engagement, enhance skill, and provide a trusted data set for AI development. This motivation provides the foundation for the problem statement and the objectives that follow.

1.2 Research Objectives

The main goal of this thesis is to enhance worker performance, engagement, and data reliability in crowdsourcing by designing and evaluating effective task features, particularly a collaborative gamified image-labelling task, on worker engagement, performance, agreement, and perceived workload when they are used for crowdsourced image-labelling tasks. From the perspective of motivational and collaborative design, we describe how various game features (e.g., chat, feedback, scoring, leaderboard, sharing goals, and levels) potentially contribute to improving human experience while affecting quality in real-time teamwork. To meet this objective, the research is guided by specific objectives as follows:

Obj1: To explore and examine how crowd workers perform and engage in different crowdsourcing tasks.

Obj2: To classify and analyse the motivation features in the current literature.

Obj3 & Obj4: To design and implement a platform to enhance worker performance (accuracy, agreement strength), and engagement grounded in motivational theories.

Obj5: To evaluate the effect of enhancement features on worker performance, engagement, and perceived workload.

Obj6: To provide empirical and theoretical understanding about how enhancement features can be strategically used to enhance engagement and performance, thus the data quality in HITL AI workflows.

1.3 Thesis Statement

This thesis argues that the strategic integration of gamification and real-time cooperation can significantly improve workers' performance and engagement while reducing perceived workload in crowdsourced labelling tasks. Based on SDT and Flow Theory, the research illustrates, through multi-phase framework design incorporating exploratory analysis, a scoping review, and a controlled experimental study, that motivation and performance in crowd work are optimised when features like chat interaction, score feedback, level, leaderboard, and shared visibility and goal are aligned to fulfil intrinsic psychological needs. The results present a theoretical and empirical foundation for the design of effective human-centred and motivationally coherent crowdsourcing systems.

1.4 Research Questions

The dissertation attempts to address several interconnected gaps found in the literature:

- First, the lack of empirical comprehension regarding how workers perform and engage across different crowdsourcing tasks. Establishing this baseline is critical to understanding the inherent challenges, such as worker variability in performance, disengagement, and task selection preferences.
- Second, the scarcity of a cohesive synthesis for the prior studies of how different motivational features (including gamification) have been applied across different crowdsourcing archetypes. Gamification has emerged as a feasible approach; nevertheless, its implementation has been scattered and mostly focused on competitive or individual frameworks, limiting a holistic grasp of its potential.
- Third, systematic evidence supporting the process of designing, integrating, and assessing such features in crowdsourcing projects, particularly in collaborative environments, has largely been neglected in previous work despite its theoretical promise.
- Finally, there is a lack of knowledge about how these features influence the workers engagement and performance in term of reliability of outputs, specifically on worker accuracy and consensus, which are crucial for generating high-quality datasets to facilitate AI and ML applications.

This dissertation aims to address these gaps through a multi-study research programme, beginning with an exploratory baseline study of the crowd workers' engagement and their performance across crowdsourcing task types, advancing to a systematic scoping review to synthesise the use of additional features, including gamification, across crowdsourcing archetypes, and finally, the design and empirical evaluation of a collaborative gamified image-labelling task. The investigations are organised around four principal research questions (RQs), each contributing to the overall cohesive framework by establishing foundations, mapping existing knowledge, and testing interventions.

RQ1: How do workers’ performance and engagement vary across different crowdsourcing task types?

This question establishes a fundamental baseline for this dissertation. By investigating the fundamental crowdsourcing tasks, it enabled us to identify the inherent challenges, such as disengagement, the preferences of participation in crowdsourcing tasks, and performance variability. Without such a knowledge benchmark, it would be difficult to assess the added valuable design aspect interventions.

RQ2: How have the additional task features (e.g., gamification elements) been applied in existing crowdsourcing research across the main archetypes?

The question provides a structured perspective to get a full understanding of the present state of knowledge. Mapping motivational features across archetypes (crowd processing, crowd rating, crowd solving, and crowd creating) helps to reveal patterns, strengths, and deficiencies in previous studies. This is crucial for clarifying areas where such features have been excessively used and where additional investigation is needed.

RQ3: How can such features be systematically designed and integrated to improve workers’ performance and engagement in crowdsourcing image-labelling tasks?

This question addresses a significant challenge in enhancing the efficiency and engagement of crowd workers by concentrating on the design and implementation of collaboration and gamification features. The importance of such a question lies in connecting the theoretical incentive models with the actual system design, thereby guiding features that maintain worker interest in image-labelling tasks.

RQ4: How do such features affect the reliability of the crowdsourcing outcomes, particularly in terms of worker agreement?

Consensus among workers is crucial for generating high-quality datasets, which subsequently support ML and AI applications. Examining the impact of additional features on worker consensus and accuracy is critical for assessing how the task designs using such features may enhance not only engagement but also reliable and scalable outcomes, which emphasises the importance of this research question.

1.5 Summary of Key Findings

The thesis enriches our understanding of how workers’ engagement and performance vary across different crowdsourcing task types and of workers’ task-selection preferences. A theoretical review, moreover, identified the key gaps in the current studies in the crowdsourcing with mapping the motivational features across different crowdsourcing types. The thesis has made significant empirical and theoretical contributions to the literature on how the worker outcome quality, engagement and perceived workload are affected by collaborating gamification in crowdsourcing environments. Between subjects experiments of participants engaged in real-world image-labelling tasks, we found that workers who participated in collaborative task environments have significantly higher engagement and accuracy comparing to the workers in non-gamified and individual conditions. The participants in condition of multiple gamification features (i.e., collaboration with scoring, feedback, levels, chat, and leaderboards inserted) achieved the highest levels of agreement and

accuracy as well as the highest engagement scores across all conditions. This indicates that the social and feedback-based components increased attentiveness and decision-making consistency. Moreover, these workers had a smaller mental and temporal workload in comparison to those not collaborating, suggesting that providing real-time collaboration and feedback reduced perceived workload without compromising performance using NASA-TLX analyses by Hart and Staveland (1988). Teams with low levels of gamification (e.g., Chat-Only) showed average engagement and higher workloads, and individual players had the lowest engagement but higher workloads on average. Statistical analyses also indicated that there were statistically significant differences between the groups (using Kruskal–Wallis and Mann–Whitney U tests), suggesting that integrated gamified collaboration is more effective for generating motivation and data quality compared to isolated approaches.

The contributions therefore offer empirical evidence that combining features applied from motivational and collaborative perspectives drawing on the SDT (Ryan and Deci, 2000) and Flow Theory (Csikszentmihalyi, 1990) can enhance intrinsic motivation, contribute to high-quality outcomes and improve engagement in human-AI data-labelling workflows.

1.6 The Significance and Contributions of the Study

This research contributes to the theoretical, methodological, and practical aspects of crowdsourcing, human-computer interaction, and motivational design. The thesis progresses through three main studies, each one producing a distinct contribution (C1-C4); these together advance our knowledge on understanding how motivational and collaborative mechanisms influence performance and engagement in the field of crowdsourcing. Across studies, worker outcomes are consistently examined across three dimensions, performance, engagement, and perceived workload. Performance is operationalised progressively based on each study's design, baseline task accuracy in the exploratory Study 1 (C1), accuracy-related outcomes synthesised from the literature in the scoping review (C2), and labelling accuracy and worker agreement strength in the experimental Study 3 (C4).

C1: Insight into how workers behave in baseline crowdsourcing tasks. It empirically contributes to understanding worker behaviour across different task types, using 750 workers recruited through Amazon Mechanical Turk, through explaining how task characteristics shape the performance, engagement, and perceived workload they exhibit across a variety of task types, where performance at this baseline stage is captured through task accuracy.

C2: A theoretical review that discusses the motivating factors in crowdsourcing. This contribution provides an integrative understanding of motivation theory, integrating SDT and Flow Theory, and identifies through a systematic mapping, how gamification features relate to worker performance (including accuracy), engagement, and workload across archetypes, and reveals the comparative under-exploration of collaborative

motivational design. It thereby explains why gamification features drive worker motivation and provides the basis for a collaborative motivational design.

C3: Methodological Framework of Collaborative Gamified Crowdsourcing. This work provides a self-developed experimental environment for real-time, collaborative crowdsourcing with configurable gamification features and a combined evaluation pipeline that provides a reusable pipeline tool for exploring how motivational and collaborative configurations influence engagement and data trustworthiness.

C4: Empirically demonstrates that collaborative gamification significantly improves worker performance (measured as labelling accuracy, majority accuracy and workers' agreement strength), and engagement while reducing perceived workload on key dimensions, with wider application in the domains of designing human-in-the-loop AI systems and next-generation platforms for crowdsourcing to strike a balance between enjoyment, fairness and productivity. These findings carry direct implication for platform requester and crowd workers alike. In summary, these contributions:

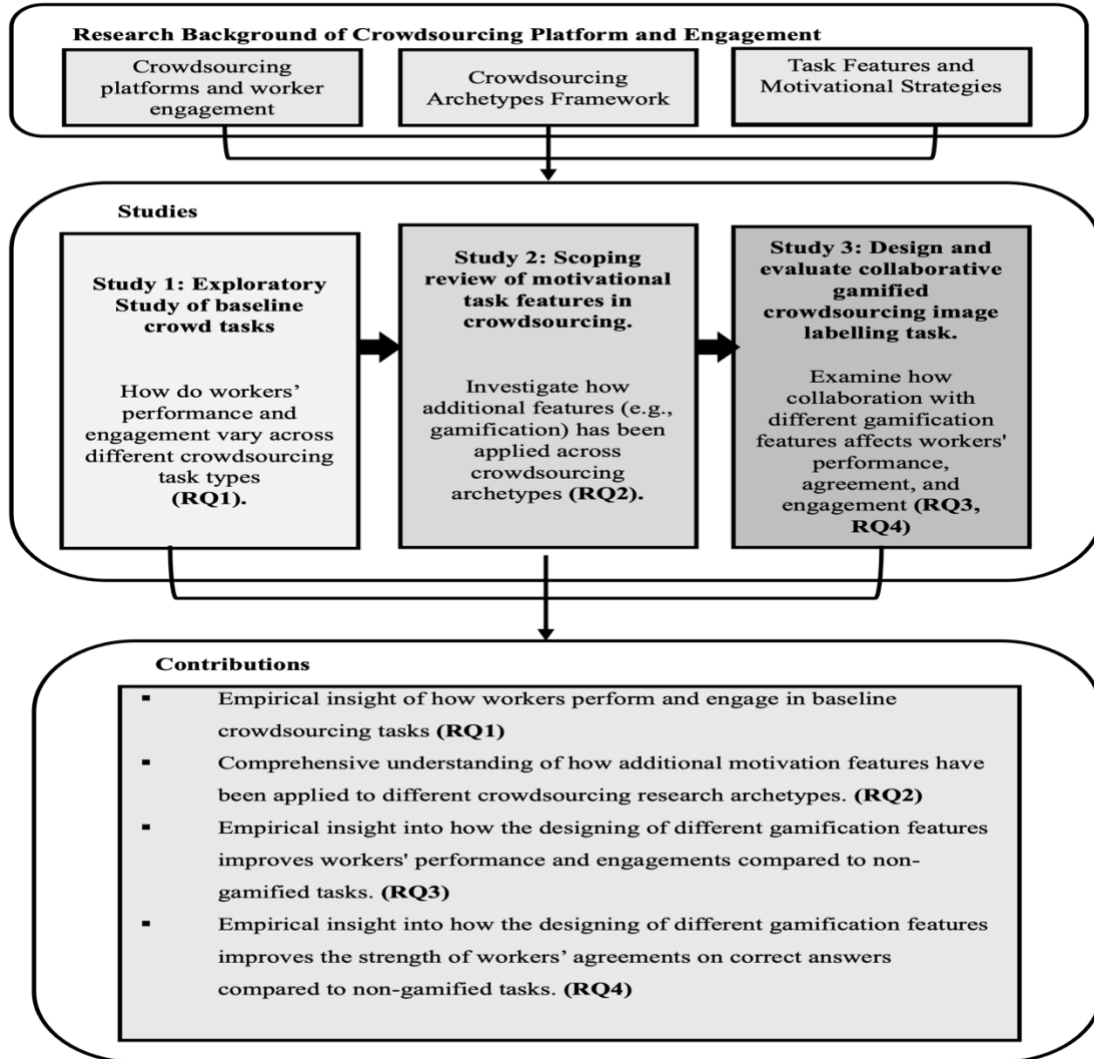


Figure 1.1: Overview of the research framework

- Advance theory on how gamification and collaboration interact to influence worker motivation in the collective digital labour market.
- Provide a validated methodological description to assess the reliability of collaborative gamified systems outcome.
- Provide practical implications for building effective, accurate, and scalable human-AI data pipelines.

These contributions lay the groundwork for next-generation gamified crowdsourcing systems that both enhance human experience and data quality. Figure 1.1 illustrates the comprehensive framework of the research, encompassing its theoretical background, consecutive studies, and resultant contributions. The figure presents how each study advances from the preceding one to tackle specific research inquiries and produce related contributions (C1–C4).

1.7 Thesis Structure

Figure 1.2 illustrates the distribution of research topics, aims, and contributions throughout the chapters of the thesis. The figure presents a roadmap that connects each thesis study to the corresponding contribution(s), helping to track how the argument and evidence build across chapters. The whole thesis is structured around nine chapters, which follow one another logically:

Chapter 1: Introduces the context, rationale, problem statement, aim and object of the research, research questions/contributions, and significance of the study.

Chapter 2: Presents the background and literature review which focuses on the theoretical bases that underlie crowdsourcing, gamification, and motivation. It reviews current taxonomies of crowdsourcing systems, motivation theories in relation to digital labour and previous empirical research on gamified crowdsourcing.

Chapter 3: The research methodology presents the general research plan, theoretical rationale, data-collecting procedures, and experiment conditions, as well as major instruments used and issues of ethics.

Chapter 4: The exploratory study describes a large-scale empirical experiment involving 750 crowd workers, analysing the engagement and performance across different types of crowdsourcing tasks. This provides the bedrock for comparative behavioural analyses in later studies. This chapter forms the basis of my Human Computation and Crowdsourcing paper titled “An exploratory study of the impact of task selection strategies on worker performance in crowdsourcing task”

Chapter 5: A scoping review of gamified crowdsourcing systems reviews existing gamification research in different archetypes of crowdsourcing systems based on their gamification elements, motivation mechanisms.

Chapter 6: System design and implementation describe the development of the collaborative gamified image-labelling prototype platform, its architecture, functionalities, and interface design as anchored in SDT and Flow Theory.

Chapter 7: Experimental results and discussion report quantitative and qualitative results of the main study, examining the accuracy, agreement strength, level of engagement, and perceived workload among experimental conditions. Statistical analyses and post hoc tests are described with a content analysis of workers' feedback.

Chapter 8: Discusses implications of the findings in relation to current theory and research and presents the strengths and the weaknesses of the study.

Chapter 9: Conclusion and Future Work.

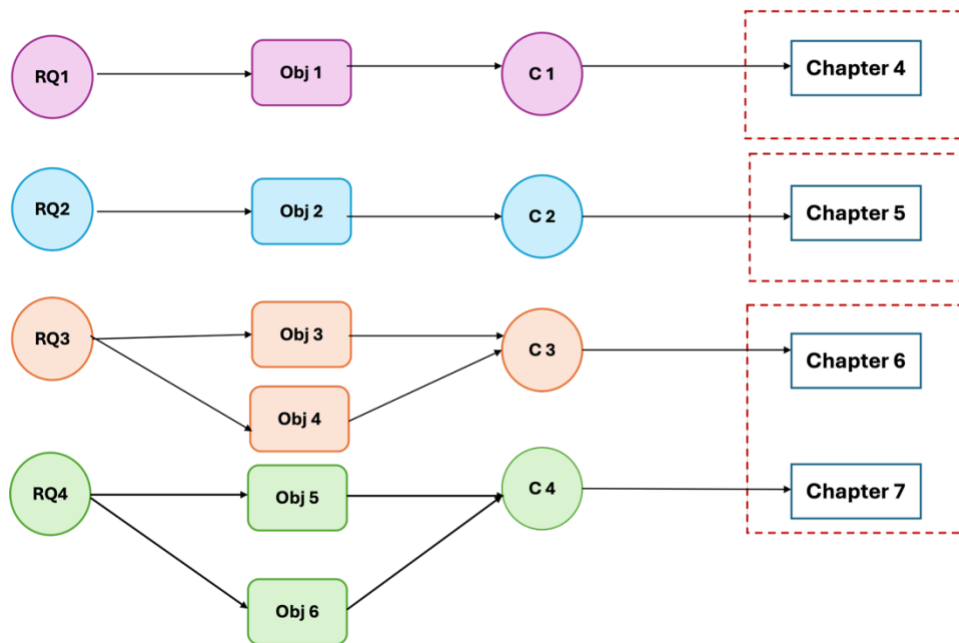


Figure 1.2: Mapping of research questions (RQ#), objectives (Obj#), contributions (C#), and corresponding chapters

Chapter 2 : Background and Conceptual Foundations

2.1 Introduction

In this chapter, we develop a theoretical and conceptual framework for contextualising the study under the wide umbrella of human computation, crowdsourcing, and gamification. Crowdsourcing and human computation have emerged as critical paradigms for resolving complex problems that are still difficult to fully automate. Such approaches are particularly important in domains that involve judgment, perception, and contextual understanding, such as data annotation, evaluation, and quality assurance in ML and AI systems. This chapter provides a comprehensive review that reviews different domains of literature in areas of collective intelligence, human computation, crowdsourcing, and its interface with ML and AI, addressing the importance of task design and worker characteristics on performance and data integrity. The literature creates conceptual foundations, elucidate essential linkages between these areas, and position crowdsourcing as a socio-technical mechanism for facilitating data-driven AI systems. The goal of this chapter is to identify remaining gaps and questions in the available literature, in particular between crowdsourcing task design decisions, worker engagement, and the performance of crowdsourcing systems. The chapter ends by emphasising also how these deficiencies make further considerations for better task design necessary. The deficiencies observed with the help of the identified tasks are further discussed in the exploratory analysis in Chapter 4, and the subsequent studies developed in later chapters in the thesis.

2.2 Literature strategy

The literature presented in this chapter was selected and organised to support the conceptual development and problem formulation in the context of crowdsourcing for AI related tasks. In this chapter, we focus on reviewing the studies that contribute towards understanding the role of human computation, the architecture and function of crowdsourcing systems and their applications in ML and AI pipelines. This review was based on the aims of the research for the thesis, prioritising research that examined task design, worker behaviour, engagement, and performance in crowdsourcing contexts relevant to data annotation and human-in-the-loop systems. Foundational studies were used to define central concepts and definitions, and more recent studies have been reviewed to capture contemporary challenges and emerging directions in the field. The studies are presented thematically to demonstrate the conceptual connections between human computation, crowdsourcing, and AI. The methodology for selecting and organising literature used in this study is illustrated in Figure 2.1.

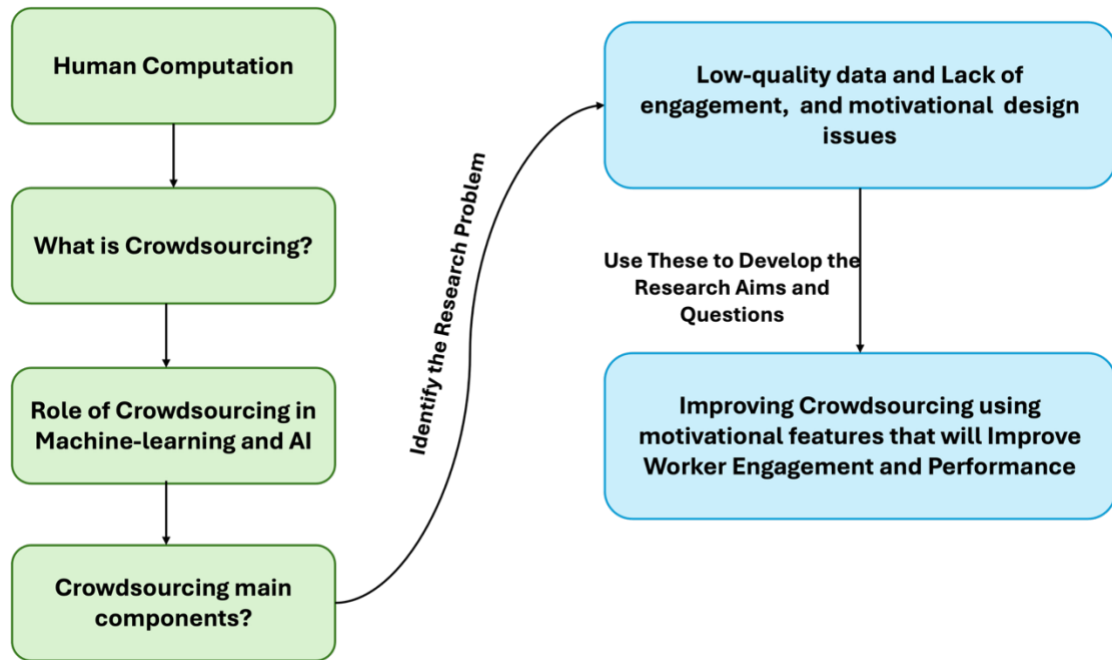


Figure 2.1: Flow diagram for literature review reading and writing

Such a structure allows the discovery of persistent issues, such as low data quality, limited worker engagement, and insufficient alignment between task design and human motivation, which directly guide the research questions and experimental approach described in subsequent chapters.

2.3 Foundation of Crowdsourcing and Human Computation in ML/AI

Crowdsourcing is generally defined as an open call that invites individuals to contribute knowledge, effort, or creativity through a digital platform (Howe, 2008, Brabham, 2008). It is a socio-technical model of mobilising distributed human intelligence to address problems and create content and to complete tasks that may exceed the capabilities of a single expert or organisation; thereby, harnessing and leveraging the collective skills and motivation of the crowd is a main concern that should be considered in such platforms (Malone et al., 2009, Brabham, 2008, Mennis, 2006).

Crowdsourcing takes many forms, ranging from data annotation for ML, creative ideation, design competitions, task processing, and finally large-scale scientific projects. Classic examples include Galaxy Zoo for astronomical image classification (Lintott et al., 2008) and Foldit for protein structure prediction (Cooper et al., 2010). These examples illustrate how human perceptual judgment and pattern recognition continue to add value in domains where computational models face ambiguity or limitations.

In recent years, crowdsourcing has evolved into complex sociotechnical systems that integrate human intelligence with algorithmic processes, often called human-in-the-loop (HITL) systems. This integration

ensures that data-driven systems remain sensitive to contextual nuances, ethical considerations, and domain-specific interpretations that automated models alone may overlook (Saeed et al., 2025, Le et al., 2025).

Human computation (HC) broadens this concept by integrating human reasoning into algorithmic processing to solve problems that computational systems cannot efficiently or reliably handle on their own (Quinn and Bederson, 2011). In AI and ML, HC provides essential human judgment on labelling, verification, and evaluation tasks.

Crowdsourcing and HC are being used more frequently as the source for supplying judgment in computer-based or data-oriented tasks that humans are still better at than machines. Recent works on human-in-the-loop learning and interactive annotation demonstrate the contributions of crowd workers in noise-label smoothing, robustness improvement, and accelerated annotation efficiency (Le et al., 2025, Saeed et al., 2025, Xu et al., 2025b). These works, as well as applications in disaster management, environmental monitoring, and social computing (Abid et al., 2025) demonstrate that HC is not limited to the simple tasks but extends into dynamic AI-learning environments and complex decision-support systems. Together, they suggest that the crowd has moved from a source of micro-labels to an active participant in responsible, transparent, and rigorous AI development.

In synthesis, crowdsourcing together with human computation thus supplies the human infrastructure for many present AI systems. Its effectiveness requires not simply high-quality architectural mechanisms but also an understanding of worker behaviour, motivation, and engagement. These subjects will be explored in later parts of this thesis.

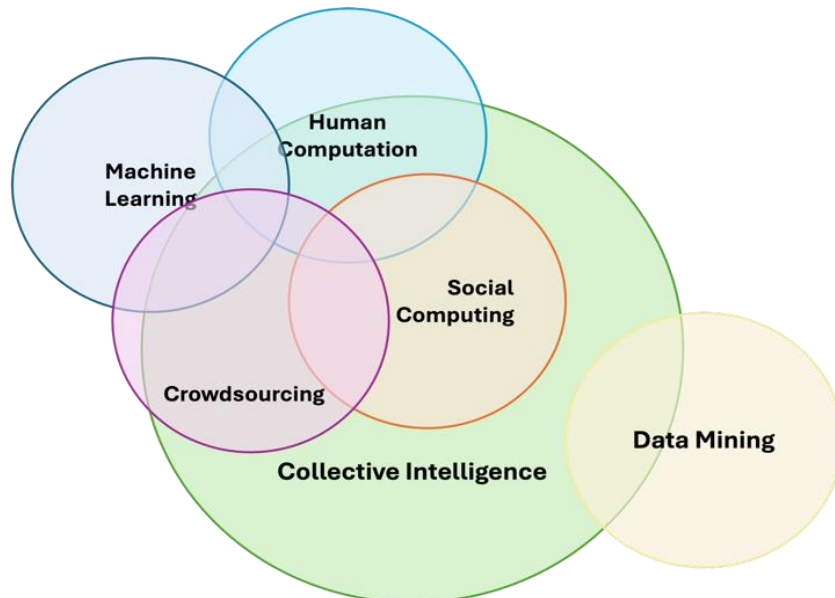


Figure 2.2: Adding Machine Learning with Human Computation as a means of solving computational problems, adapted from (Quinn and Bederson, 2011)

Crowdsourcing has gained significant popularity in recent years (Lenart-Gansiniec et al., 2022, Zuchowski et al., 2016) and has profoundly affected the processes of product and service invention, production, funding, marketing, distribution, and utilisation, resulting in substantial economic consequences (Hoque, 2024, Gong et al., 2025, Cao et al., 2023). Crowdsourcing has been significantly introduced in many sectors, including science and education (Brossard et al., 2005), the government (Hassan, 2016), journalism (Aitamurto, 2015), tourism (Sigala, 2015), software development (LaToza, 2019), manufacturing (Hutter et al., 2011), and AI and ML (Geng, 2024).

In the context of AI, as mentioned by many researchers, there is a need to engage such crowds of humans who have different intelligence and skills to lay the base for the development of the machine. Several studies in various areas have shown that there are some approaches in which human cognitive capabilities complement the computational processes to solve problems (Kim and Monroy-Hernandez, 2016, Tsvetkova et al., 2017). To define crowd intelligence, it is vital to explore the most crucial concepts and their relation. Figure 2.2 shows these concepts as overlapping circles. The essential central construct is collective intelligence, which many researchers highlight as fundamentally important to effective distributed problem-solving. A definition of collective intelligence was by authors in (Malone et al., 2009), who defined it as "a group of individuals who cooperate in a way that looks intelligent." While HC uses a few humans, crowdsourcing platforms leverage the wisdom of those needed workers by outsourcing tasks to a large number of individuals in the community. Crowdsourcing is primarily defined by Doan et al. (2011) as "enlists a crowd of humans to help solve a problem defined by the owners of the system".

The term "social computing" refers to a wide notion that encompasses all aspects of social behaviour and computing. The use of social software to support communication and collaboration is at the heart of social computing. Social network analysis, online social dynamic modelling, and social software design are all research issues in social computing. Because both phrases have their origins in the field of social science, social computing and collective intelligence are closely related.

Moreover, our research adds a ML aspect to the taxonomy of the study Quinn and Bederson (2011). There are several connections between the domains of ML, Human Computation, and crowdsourcing that provide new study opportunities. Currently, in the ML area, we are seeing a high growth in AI, robotics, and automation in various job markets. The importance of human contribution to creating entire data sets or labelling training data to drive advancements in ML has been well documented; for example, consider ImageNet's role in catalysing advances in AI (Deng et al., 2009). Therefore, crowdsourcing techniques solve the knowledge expert and annotator bottlenecks required for the high-scale use of AI research such as linked data and support the creation of ML datasets. Labelling tasks has gained extensive usage across different ML applications as a popular and cost-effective solution for performing data analysis and sentiment analysis. As a

result, the widespread adoption of ML relies on the availability of a sizable human input, which might be offered through crowdsourcing platforms.

2.4 Crowdsourcing Components

Crowdsourcing and online labour markets (e.g., MTurk) have arisen as novel sources of human labour, enabling enterprises to dynamically adjust their workforce and engage specialists, often at a relatively cheap cost or no cost at all. The neologism was coined by Howe (2006b) as “The act of a company or institution taking a function once performed by employees and outsourcing it to an undefined network of people in the form of an open call”. This definition clearly delineates the essential elements of crowdsourcing and aligns with other definitions of the term (Estellés-Arolas and González-Ladrón-de-Guevara, 2012, Hammon and Hippner, 2012). Crowdsourcing is the combination of the concepts of "crowd" with "outsourcing" (Rouse, 2010). The notion resembles traditional "outsourcing," albeit with numerous significant distinctions. Initially, in traditional outsourcing, one organisation delegates work formerly executed internally to another organisation (Lacity and Hirschheim, 1993). Consequently, it is evident to whom the corporation delegates its responsibilities (Rouse, 2010). In crowdsourcing, a firm or individual delegates tasks to an "undefined network of people", the crowd (Howe, 2006a, Rouse, 2010). Consequently, they are unaware of the entities to which they are outsourcing. Secondly, crowdsourcing constitutes an "open call" and thus the completion timeline and quality level of the tasks remain uncertain. Thus, the two main important factors in such platforms are the workforce “workers” and the outsourcing “tasks”.

2.4.1 Crowdsourcing workers

The human factors in the crowdsourcing platform depend on the two different actors in the crowdsourcing platform: the worker and the requester. Requesters are organisations that represent the requested tasks in the crowdsourcing platform, while the workers constitute the distributed workforce that does these tasks. The task’s owners are the source of financing for the crowdsourcing workers and pay labourers who deliver substantial or acknowledged high-performance output. Workers are the critical value of the crowdsourcing ecosystems (Howe, 2008). This workforce is accessible online and performs millions of tasks on digital platforms, typically in exchange for a small reward or other forms of compensation. Dependence on human capabilities raises numerous questions that have been central to academic investigation by analysts for a long time. Some studies highlight the impact of the educational background, demographic composition, and task context on the workers' performance and reliability. (Ipeirotis, 2010b) defined that approximately 60% of the solvers were from the United States and India, and about 35% possessed a bachelor's degree or higher. Similarly, a study of (Jain et al., 2017) found that around 50% of the crowd labourers originated from five nations: the United Kingdom, the United States, Venezuela, India, and Canada. Other researchers, such as (Jain et al., 2017), and (Gadiraju et al., 2015), examined psychological behaviour, satisfaction, and motivations, all of which affect the workers’ efficiency and ability to produce high-quality results. Jain et al. (2017)

demonstrated that a deeper understanding of worker behaviour, including anchoring effects, worker learning patterns, and their social interactions across different task types, can significantly improve the task design and worker reliability. Likewise, (Gadiraju et al., 2015) identified an important observation: the gold standard test questions alone are insufficient for identifying malicious workers or low-quality work. Even trusted workers who pass the qualification criteria test can provide inappropriate answers, indicating the need for robust continuous behaviour monitoring and clearer contextual guidelines during the task design. By understanding the common behaviours in the crowd, requesters can make their tasks much more effective. The study by Kazai et al. (2011) aimed to classify different worker types based on their performance and engagement patterns. They identified five types of workers: (1) spammers, (2) careless, (3) incomplete, (4) competent, and (5) hardworking. Their findings demonstrated a strong relationship between worker characteristics and the quality of submitted labels, showing that such insights are useful not only for collected data analysis but also for refining HIT designs to attract and retain competent contributors.

A study by André et al. 2014) emphasised that several contextual variables, such as the personality of the workers, task release time on the platform, and the nature of the tasks can significantly affect the workers' motivation and engagement on crowdsourcing platforms. Zhuang and Gadiraju (2019) examined the impact of workers' mood on task performance and engagement, showing that positive mood states enhanced workers' perceived achievement and satisfaction. As a result, understanding the workers' personality traits, mood, and motivational drivers remains a critical challenge for encouraging engagement and enhancing outcome quality (Doan et al., 2011). Recent empirical evidence also indicates that such behavioural insights are highly important in modern AI and human-in-the-loop systems for enhancing the accuracy, well-being, and reliability of the worker in crowdsourcing.

2.4.2 Crowdsourcing Task

Numerous efforts have been made to categorise the extensive array of crowdsourcing systems by considering the attributes of the crowdsourced tasks and the techniques through which value is generated. One of the most widely cited classifications is that of Geiger et al. (2012), who identify four archetypal methods, crowd-processing, crowd-solving, crowd-rating, and crowd-creating, based on the homogeneity of contributions and whether value directly or collectively emerges from them (Figure 2.3).

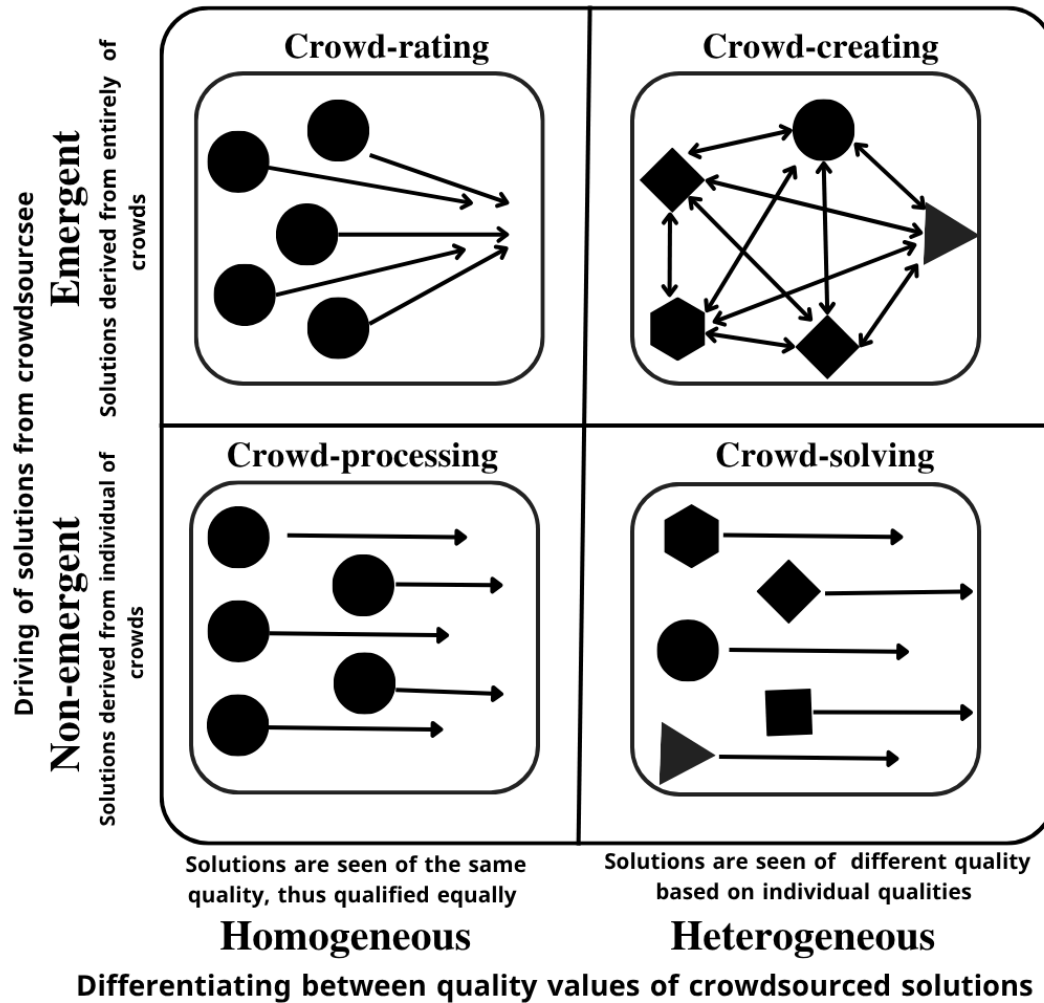


Figure 2.3: Crowdsourcing categories adapted from Geiger et al (2012).

Crowd-rating is a system that commonly takes the benefits of a crowd to collectively predict or assess the required task, where the emergent data is produced from large homogeneous data such as votes or ratings. Crowd-creating systems aim to build artefacts or content through various heterogeneous contributions, such as user-generated data or collaborative knowledge projects. Crowd-solving harnesses the wisdom of the crowd to construct heterogeneous solutions to complex problems or poorly defined problems, such as idea competitions or innovation challenges. Finally, crowd-processing involves distributing huge numbers of homogeneous micro-tasks, such as labelling or verification, where values are derived from many non-emergent individual outputs. Table 2.1 summarises the four principal archetypes of crowdsourcing, establishing a systematic framework for comprehending the variations in worker contributions and value creation across different systems.

Table 2.1: Crowdsourcing archetypes

Archetype	Contribution	Structure	Value Type	Representative Examples	Key Literature
Crowd-processing	Homogeneous microtasks performed at scale by many workers		Non-emergent (value derived directly from each contribution)	Image labelling, text transcription, click tasks	(Mahmoudi-Dehaki and Nasr-Esfahani, 2025, Le et al., 2025, Ng et al., 2025, Moreno-Alcayde et al., 2025)
Crowd-rating	Homogeneous evaluations aggregated into a collective outcome		Emergent (value arises from aggregation of votes/ratings)	Ratings, relevance judgements, ranking	(Wu et al., 2024, Zhai et al., 2025, Zhang et al., 2024, Frering et al., 2025)
Crowd-solving	Heterogeneous solutions to complex or open-ended problems		Non-emergent (solutions evaluated individually)	Ideation contests, problem-solving challenges	(Pan et al., 2024, Mesbah et al., 2023, Yang et al., 2022, Daradkeh and Atalla, 2022)
Crowd-creating	Heterogeneous contributions co-creating complex artefacts		Emergent (collective creation of knowledge or content)	Citizen science projects, collaborative datasets	(Wyss et al., 2025, Spicher et al., 2024, Meisner et al., 2024, Arellano-Verdejo and Lazcano-Hernandez, 2023)

The above categorisation framework has been widely applied in crowdsourcing research, as it places an emphasis on the value creation mechanisms underlying the different forms of crowdsourcing. Other scholars have extended this view through goal-oriented task-type classifications. Gadiraju et al.(2014) identified four taxonomies that organised crowdsourced projects based on the essential objectives, such as verification and creation, creative production, and information retrieval. Catallo and Martinenghi (2017) further refined this perspective by defining four design parameters: what tasks are outsourced, who performs them, why solvers are motivated to participate in the solution process, and how the tasks are processed. These models indicate that the systematic design of such tasks and motivational alignment is crucial to successful crowdsourcing. Motivating crowdsourcers to participate efficiently is therefore critical to the overall design of any crowdsourcing task. Poorly designed tasks reduce worker motivation and the outcome quality. Consequently, the success of crowdsourcing depends heavily on how the tasks are designed, communicated, and rewarded (Brabham, Brabham, 2008, Doan et al., 2011, Law and Ahn, 2011). Understanding the motivations of crowdsources, intrinsic and extrinsic, is essential (Zhao and Zhu, 2014a). Moreover, individuals in large-scale online communities possess diverse interests, motivational orientations, and skills, which makes it challenging for organisations to engage and coordinate crowds toward specific objectives or behaviours (Tapscott and Williams, 2008). Research studies consistently demonstrate that motivation in crowds requires a mix of appropriate incentive mechanisms and well-structured workflows (Aitamurto, 2015, Kashima et al., 2024, Soliman and Tuunainen, 2015). Intrinsic worker motivation arises from tasks that allow autonomy, competence and enjoyment, whereas extrinsic worker motivation arises from external rewards such as monetary compensation or social recognition (Hamari et al., 2016, Kaufmann et al., 2011, Straub et al., 2015). In many

contexts, intrinsic rewards can outperform the financial rewards, leading to better outcome quality and higher engagement. Despite extensive research on crowdsourcing, only a limited number of studies have thoroughly examined the motivational structures of different crowdsource types and contexts (Soliman and Tuunainen, 2015, Vanden Berghen et al., 2024, Humayun et al., 2024, Zhao and Zhu, 2014b). Moreover, crowdsourcing outcome quality is still a challenge, especially for tasks where financial incentives dominate, due to worker boredom, fatigue, inadequate design of Human Intelligence Tasks (HITs), and mismatches of task requirements and the skills of the workers (Kazai et al., 2012). The final quality of the data is affected by the profiles and abilities of the workers, the formulation of the tasks, the incentive schemes, and the reputation and interaction style of the requesters. As Daniel et al. (2018) explain, output quality can be measured with tangible criteria that embody both objective and subjective aspects of performance, and it is ultimately shaped by human, technical, and procedural factors in platform ecosystems. Hence, a fundamental key challenge in designing crowdsourcing systems is to provide motivation that aligns with the nature of the work and the diversity of the contributors. Different crowdsourcing tasks could require unique forms of participant motivations and structure approaches. Both a goal-oriented perspective of Gadiraju et al. (2014) and a values-creation perspective of Geiger and Schader (2014) showed that crowdsourcing activities can be understood across a number of levels, from isolated micro-contributions to complex and collaborative projects.

Some methodologies rely primarily on isolated worker outputs, while other types may require social motivation mechanisms or gamified techniques that promote cooperation and motivate engagement. Gamification has proven to be one of the most high-profile approaches to enhancing motivation and involvement in crowdsourcing tasks (Morschheuser et al., 2016). A growing body of work on gamification in crowdsourcing types also shows how it is being applied in crowdsourcing environments (Kavaliova et al., 2016, Mekler et al., 2017, Zheng et al., 2011), yet the existing literature lacks a comprehensive understanding the impact of specific gamification mechanisms on the workers' motivation and performance outcomes across different archetypes.

2.5 Motivational Mechanisms

Motivation is an essential factor that enables individuals to participate in many jobs and achieve their objectives across different platform community, shaping the quality, engagement and persistence in these platforms (Zhao and van der Hoek, 2015). In the context of crowdsourcing, what determines the motivation score is not only whether workers will engage but, far more importantly, how effectively they perform their tasks. Motivation is particularly complex in crowdsourcing, as it mediates between the cognitive, social, and economic dynamics that drive individuals to perform microtasks, label data, or co-create content in human–AI collaboration systems. Motivation is a multidisciplinary subject that has been examined across different domains, including business (Guo et al., 2019), management (Dzombak et al., 2016), medicine, social sciences (Shank, 2016), and psychology (Stewart et al., 2017). Recent studies in crowdsourcing also place an emphasis on the interplay between cognitive incentives (e.g., learning, curiosity), affective experiences (e.g., enjoyment,

belonging), and instrumental drivers (e.g., payment or recognition). Motivation is seen as the impetus that compels individuals to engage in actions that are either objective (goal-oriented) or subjective (mental constructs) (Matschke et al., 2014). Consequently, motivation drives both cognitive intention and the subsequent behaviour. Thus, it can steer humans toward desired outcomes while also keeping them out of undesirable circumstances (Zhou and Wang, 2018). When individuals are positively motivated, they experience satisfaction, enjoyment, and intrinsic pleasure in performing the activity. Significant challenges have been encountered by the researchers in motivating workers to complete tasks, including issues like loss of interest, withdrawal, or attrition over time. Since disengagement are persistent issues in crowdsourcing, motivation is acknowledged as a cornerstone for maintaining engagement and quality in online crowdsourcing platforms (Kittur et al., 2013). Motivation possesses various forms and meaning and its integration into crowdsourcing systems has demonstrated substantial effectiveness (Graboviy, 2016).

Several motivational mechanisms have been employed by researchers, incorporating intrinsic, extrinsic, and social elements on online platforms to improve participant engagement in crowdsourcing tasks. Research demonstrated that no single motivational approach is generally effective; instead, optimal approaches depend on the task type, platform design, and worker characteristics (Kaufmann et al., 2011, Spindeldreher and Schlagwein, 2016, Posch et al., 2019). Previous reviews have classified these mechanisms of motivation in crowdsourcing in various ways. For instance, Pinto and dos Santos (2018) categorised 25 motivational elements into intrinsic, extrinsic, and social motives, a practical taxonomy commonly utilised in information system research. Additionally, contemporary literature highlights design-centric strategies, such as priming, onboarding, conversational styles, and gamification, which integrate many motivators and can be intentionally embedded within the task workflows to improve workers' performance and engagement. Gamification is particularly useful in non-game scenarios for incorporating elements of the game experience: points, badges, leaderboards, etc. (Morschheuser et al., 2019). Current literature in human-in-the-loop (HITL) data annotation and AI training tasks indicates that motivational approaches must be context-sensitive to balance the efficiency requirements of ML workflows with human well-being and cognitive satisfaction (Jiang et al., 2024; Patel et al., 2023).

More extensive reviews, however, have discovered that gamification in crowdsourcing is under-investigated and warrants further systematic exploration grounded in theory to ensure collective participation in crowdsourcing activities (Riar et al., 2022). Although extrinsic motivators continue to be the bedrock to engage the workforce, intrinsic and social resources are gaining precedence to ensure and enhance engagement and eliminate cognitive fatigue during a repetitive or complex labelling task. These mechanisms are discussed in detail in the following subsections.

2.5.1 Extrinsic Motivation Mechanisms

The most traditional form of motivation in crowdsourcing includes extrinsic rewards, especially financial remuneration. Initial research on task marketplaces such as Amazon Mechanical Turk revealed that workers frequently choose tasks primarily according to the payment level or compensation rates (Ipeirotis, 2010a). Subsequent studies have explored how different payment systems affect participation and quality of output in data annotation settings as well as environments where people are loops for other people to make decisions. Furthermore, to enhance the quality of the outcomes, the effect of performance-based payments has been examined by subsequent research. Ho et al. (2015) discovered that performance-based payments improved the quality in "effort-responsive" tasks, where increased worker effort yielded superior outcomes. Studies of large-scale labelling and AI data collection projects, however, have recently shown that payment mechanisms still play a role in task acceptance; but not necessarily sustain high performance over time (Jiang et al., 2024, Kong et al., 2025). For instance, hybrid reward structures using base pay plus task-specific bonuses have been used to increase crowd annotation accuracy while controlling costs (Guo et al., 2023). This confirms earlier findings that rewards need to be proportional to task complexity and worker expectations, rather than be tied to the same reward in the pay-per-task model (Gallus, 2017). However, beyond a certain threshold, increased payment may also reduce the data quality, especially when involved in monotonous and repetitive small-scale task-oriented activities (Rogstadius et al., 2011). Multi-dimensional incentive systems using financial, reputation, and feedback-based mechanisms are suggested to mitigate this issue (Patel et al., 2023, Zhao et al., 2023). In addition to maintaining engagement, such systems encourage perceived fairness and trust, which are critical for retaining participation in platform work (Gallus, 2017, Yu et al., 2019). Performance-based payment approaches have been shown to improve the output quality in crowdsourcing activity where the workers' efforts directly influence the performance, specifically in the data annotation and labelling contexts (Ho et al., 2015). However, their effectiveness relies on whether the workers perceive evaluation criteria as transparent and fair, since an ambiguous or punitive evaluation system can undermine trust and involvement (He and Zhang, 2024). New algorithmic incentive models also prioritise flexibility, with the use of dynamic incentives that adjust to employee behaviour, speed, and reliability (Jiang et al., 2024, Kong et al., 2025). This marks a shift from static reward design to continuous, context-sensitive feedback loops. Extrinsic motivation basically remains foundational but insufficient to encourage participants' engagement (Kaufmann et al., 2011). Improving participation and involvement requires a balance of economic, cognitive, and social reinforcement, which naturally introduces the field of intrinsic motivation. Although extrinsic incentives are extensively examined, there is limited knowledge regarding the interaction between monetary and non-monetary extrinsic incentives and intrinsic motivations.

2.5.2 Intrinsic Motivation Mechanisms

Intrinsic motivation describes participation that is driven by internal satisfaction, curiosity, or enjoyment of the task, and it is essential to understanding why individuals engage in activities with minimal or absent extrinsic rewards (Ryan and Deci, 2000). Within the crowdsourcing platforms, intrinsic motivators have been shown to foster autonomous engagement and improve quality and attention levels in labelling and creative tasks (Raddick et al., 2010, Tinati et al., 2017).

Recent studies have enhanced this comprehension by linking intrinsic motivation with psychological constructs such as autonomy, competence, and relatedness (Siamtanidou et al., 2025, Suen et al., 2022). When platforms promote these needs (freedom to choose tasks, evidence of progress, and social contact), workers demonstrate enjoyment and persistence (Behl et al., 2021, Vergara-Borge et al., 2025).

A sense of meaningful contribution is also most important in scientific and humanitarian crowdsourcing activities in which respondents are motivated by prosocial- or altruism-oriented motives rather than profit gain (Bastanfard et al., 2024, Tinati et al., 2017). Similarly, curiosity-driven design methods, for example, using progressive task disclosure or visualisation of the feedback, have been shown to improve internal motivation without affecting the accuracy of the work (Law et al., 2016). Similarly, onboarding design elements may improve worker engagement in citizen science contexts, as shown by (Cascaes Cardoso, 2017). Human-in-the-loop annotation environments have seen an increased focus on “learning value”, where the workers perceived skills development or knowledge gain as rewarding. Furthermore, gamified cues including progress bars, proficiency-tracking dashboards, and self-assessment feedback, enable mastery and self-efficacy (Spicher et al., 2024). However, internal motivators can be weakened by poor task design, ambiguous feedback, and excessive automation that reduces workers’ sense of control over their participation (Kaufmann et al., 2011, Riar et al., 2022, Spicher et al., 2024). Thus, effective crowdsourcing systems must maintain a good balance, providing enough structure to guide the accuracy while maintaining sufficient space for freedom and curiosity to enhance interest. Recent findings also tend to support the idea that intrinsic satisfaction in the labelling tasks can enhance both participation engagement and data quality, as motivated workers exhibited lower error rates and higher attentional consistency (Bastanfard et al., 2024, Wiethof et al., 2022). Thus, intrinsic motivation drives care, depth and improve engagement, whereas extrinsic incentives primarily trigger initial participation.

2.5.3 Social Motivation Mechanisms

Social motivation in crowdsourcing research refers to the influence of recognition, reputation, collaboration, and perceived community-relatedness on worker engagement. Whereas extrinsic incentives encourage participants and intrinsic incentives deep engagement, social strategies support relatedness, fairness, and identity (Pinto and dos Santos, 2018). According to recent literature, recognition and transparency are among the most effective non-monetary motivators in human computation contexts, with evidence that symbolic

awards, despite lacking monetary value, markedly enhanced participation, highlighting the influence of social recognition (Gallus, 2017; Patel et al., 2023). Moreover, Morschheuser et al. (2019) demonstrated that collaboration and recognition of peers enhanced creativity and the calibre of tasks in crowdsourcing. Experiments on collaborative labelling platforms reveal that symbolic recognition may be as motivating as monetary rewards in aiding accuracy and involvement. Workers persist longer when they receive visible acknowledgement through badges, peer approval, or status rankings, especially when those metrics are tied to visible contribution quality rather than arbitrary measures (Kong et al., 2025). Accountability and trust are also enhanced by reputation systems and social visibility. Platforms such as Zooniverse and Foldit utilise public contribution histories and discussion boards for mutual learning and recognition (Raddick et al., 2010, Tinati et al., 2017). In data-centric crowdsourcing, gamified participation mechanisms and social feedback features can enhance workers' engagement, confidence, and perceived contribution, particularly if complemented by meaningful feedback and opportunities for progress monitoring (Siamtanidou et al., 2025, Spicher et al., 2024). However, competitive social incentives require careful implementation. For example, leaderboard-based comparative systems serve as an additional social incentive that can improve the performance for the top experienced achiever, while discouraging those at the bottom who can become frustrated through their lack of ability (Na and Han, 2023), which indicates that the nature of social incentives can either promote or hinder outcomes, depending on their design and individual differences. Likewise, studies on fairness and worker motivation suggest that perceived inequity, limited transparency, and poor design reward can negatively affect worker satisfaction, trust and workers outcome (Borromeo et al., 2017, Behl et al., 2021). Recent research also emphasises that feedback visibility, recognition strategies, and transparent contribution play a critical role in enhancing motivation in gamified human-in-the-loop systems (Vergara-Borge et al., 2025, Wiethof et al., 2022). In sum, social mechanisms work most effectively when they balance competition and cooperation; when they do, they affirm both individual worth and the crowd's collective achievement.

2.5.4 Design-Centric Motivation Mechanisms

Besides the classical extrinsic, intrinsic, and social mechanisms, motivation can also be shaped through the design of the task environment and worker experience. Such design-centred approaches include task structuring, priming, onboarding, conversational interfaces, feedback mechanisms, micro-diversions, and gamification. These approaches are not equivalent or interchangeable, as each addresses different aspects of worker motivation and experience. Gamification provides a comparatively broad motivational design space through which multiple motivational mechanisms can be incorporated within the same task environment. Rather than functioning as a single isolated feature, gamification can combine mechanisms such as feedback, progression, goal setting, achievement, recognition, social comparison, competition, and cooperation (Morschheuser et al., 2019, Riar et al., 2022). Gamification is therefore treated as a higher-level design approach through which intrinsic, extrinsic, and social motivational mechanisms can be operationalised and

coordinated, rather than as an alternative to task design itself. This broader design capacity made it particularly suitable for investigating how motivational mechanisms might be systematically integrated within repetitive crowdsourcing tasks (Morris et al., 2012). Recent empirical research has shown that gamified features might enhance worker satisfaction, time-on-task, and perceived competence in labelling and citizen-science situations (Bastanfard et al., 2024, Ernestivita et al., 2024, Spicher et al., 2024).

Moreover, systematic reviews on gamified crowdsourcing (Riar et al., 2022, Morschheuser et al., 2019) emphasise that although points and leaderboards improve engagement and it requires alignment with psychological needs such as autonomy, competence, and relatedness. These are complementary ideas with SDT and Flow Theory which is the theoretical framing of this thesis. Design-based mechanisms also extend beyond the mechanics of the game elements. Micro-diversions, brief, enjoyable breaks or contextual variations, enable the restoration of attentional recovery and enhanced memory in crowd tasks (Dai et al., 2015, Rzeszotarski et al., 2013). Priming techniques, such as quotes, inspirational messages, or visual cues, have been shown as positive or achievement-oriented stimuli that enhance engagement, creativity, and persistence (Gadiraju and Dietze, 2017). Conversational interfaces and avatar-based guidance (Qiu et al., 2020, QIU et al., 2021) have increased the motivational landscape, creating a feeling of social presence in otherwise solitary annotation work. Hybrid motivational systems, merging financial, social, and design-centric stimuli, are currently the gold standard in crowdsourcing motivation research (Jiang et al., 2024, Kong et al., 2025). Such systems can self-adjust based on worker behaviour, skill development, and engagement signals, effectively bridging the gap between efficiency and human-centred design. In summary, Table 2.2 consolidates the four principal methods employed in crowdsourcing systems and their common applications, as explained in Section 2.5.

2.6 Theoretical lens

SDT and Flow Theory are most widely recognised as complementary frameworks for motivation and engagement in human-computer interaction and digital work contexts. In a crowdsourcing platform, the relevance of these frameworks is particularly prominent, as the performance and participation of the workers are affected by a combination of intrinsic, extrinsic, social, and design-related factors, as discussed in Section 2.5.

Table 2.2: Motivational mechanisms in crowdsourcing

Motivation Type	Description	Common Mechanisms in Crowdsourcing Contexts	Supporting Literature
Extrinsic	Motivation driven by external rewards or consequences	Monetary pay, performance-based bonuses, approval ratings	(Xu et al., 2024, Abid et al., 2025, Wang et al., 2017, Fong et al., 2025)
Intrinsic	Motivation driven by interest, curiosity, or enjoyment	Interesting tasks, learning value, meaningful contributions	(Liu et al., 2025, Humayun et al., 2024, Moreno-Acevedo et al., 2024)
Social	Motivation arising from recognition and social connection	Leaderboards, peer visibility, shared spaces, cooperative tasks	(Wyss et al., 2025, Spicher et al., 2024, Meisner et al., 2024)
Design-centric	Motivation embedded in task/interface design	Gamification elements, micro-diversions, human-in-the-loop interaction	(Vergara-Borge et al., 2025, Spicher et al., 2024, Bastanfard et al., 2024, Ernestivita et al., 2024)

Prior research has indicated that while financial incentives may motivate participation, intrinsic motivations, social interactions, and meaningful task experiences are more critical in motivating engagement and guaranteeing high-quality contributions (Suen et al., 2022, Siamtanidou et al., 2025, Ryan and Deci, 2000). SDT and Flow Theory are investigated as essential frameworks for understanding motivation in crowdsourcing contexts, further elaborating on these insights. Ryan and Deci (2000) differentiate between extrinsic and intrinsic motivation and highlight the significance of three fundamental psychological needs:

- **Autonomy:** the need to experience a sense of control and agency over one's actions.
- **Competence:** the need to sense oneself as being successful and proficient in executing tasks.
- **Relatedness:** the necessity to establish connections with others and experience a sense of belonging.

If these requirements are met, individuals are more likely to internalise objectives, maintain effort, and demonstrate greater satisfaction and performance. SDT has been extensively used to analyse motivation in online work communities, citizen science, and collaborative systems, as it illuminates how contributors can remain engaged without strong extrinsic rewarding incentives. In this context, SDT offers the basic understanding of how the platform design supports autonomy (e.g., choices of tasks), competence (e.g., clear feedback and progress indicators), and relatedness (e.g., collaborative features and communication space) and can thus improve data quality, participation, and worker retention. The combined SDT demands and Flow Theory conditions offer a theoretical basis framework for understanding how particular motivational mechanisms and task-design choices are expected to affect engagement, persistence, and output quality;

Table 2.3: Alignment of SDT and Flow constructs with motivational mechanisms

Theory Construct	Motivational Mechanism	Intended Worker Effect	Supporting Literature (Scopus)
Autonomy (SDT)	Autonomy-supportive design	Supports internalisation and self-directed participation	(Xu et al., 2025a, Zhao et al., 2020, Yu et al., 2019)
Competence (SDT)	Feedback and progression	Enhances mastery, learning, and accuracy	(McNeil et al., 2023, Reinhard et al., 2023, Xu et al., 2019)
Relatedness (SDT)	Social motivation	Strengthens belonging, social presence, and collaboration	(Ustalov et al., 2023, Ye et al., 2023, Li et al., 2022)
Flow (Challenge–Skill Balance)	Design-centric challenge calibration	Enhance immersion; avoids boredom or overload	(Liu et al., 2016, Cui et al., 2017, Palma et al., 2024)
Flow (Immediate Feedback)	Instant performance cues	Improves attention, focus, and quality	(Xin et al., 2018, Liu et al., 2020, Wang et al., 2021)

Table 2.3 summarises how these theories align with common motivational mechanisms and their effects in the crowdsourcing context.

Flow Theory (Csikszentmihalyi, 1997) delineates an ideal psychological condition of immersion and pleasure that arises when an individual's abilities are appropriately aligned with the challenge posed by a task. The state of flow is identified by intense focus, a feeling of mastery, and inherent gratification from the activity. Research into digital labour increasingly demonstrates that striking a challenge–skill balance can lead to lower cognitive fatigue, improved enjoyment, and increased task persistence, particularly in repetitive image-labelling or data annotation situations. Collectively, SDT and Flow Theory provide a conceptual lens for analysing motivation and involvement in crowdsourcing. SDT highlights the psychological needs that must be fulfilled to promote ongoing participation, whereas Flow Theory elucidates how task design can foster immersive engagement when challenge and skill are harmonised. In this research paper, these two theoretical frameworks are combined to interpret worker behaviour, motivation design, and subsequent empirical patterns of engagement. The design-oriented methods, such as gamification, are conceptualised not as theoretical models but as practical means in which autonomy, competence, relatedness, and flow-affirmative situations are operationalised. For example, leaderboards may enhance competence and relatedness, while progressive levels and calibrated feedback may help maintain a challenge–skill balance that encourages flow.

2.7 Conclusion

This chapter has presented the conceptual and theoretical foundation underpinning this thesis. It began with the history of crowdsourcing and human computation, illustrating their increasing significance in the current modern artificial intelligence (AI), machine Learning (ML) and human-in-the-loop (HITL) systems. Through this discussion, crowdsourcing was framed not merely as a mechanism for outsourcing tasks but as a dynamic socio-technical process that supports large-scale data annotation and the human contributions required in AI and ML development workflows. The basic components of crowdsourcing have also been examined in this

chapter, including workers, tasks, requesters, and platforms. Together, this literature positions the crowdsourcing as a platform in which the task features, worker motivation, system design, and outcome quality requirements should be considered collectively.

A substantial section of this chapter has been devoted to motivation, which is the primary conceptual focus of the thesis. In section 2.5, we synthesised the research on extrinsic, intrinsic, social, and design-centric motivation mechanisms. This discussion emphasised that the motivation in crowdsourcing systems is multidimensional: financial rewards are frequently important for attracting participation, but worker engagement and their reliable outcome also depend on intrinsic interest, perceived competence, and social connection. Design-centric strategies like gamification, hybrid rewards, and interactive feedback have received growing attention, but the literature does not yet provide a sufficient integrated account of how these mechanisms could be combined with collaborative strategies in repetitive labelling tasks.

This chapter establishes gamification as a comparatively broad motivational design space that can integrate multiple mechanisms, including feedback, progression, achievement, social interaction, competition, and cooperation (Riar et al., 2022, Morschheuser et al., 2019). The subsequent study in Chapter 4 therefore examines how workers naturally differ in performance, engagement, workload, and task-selection preferences across non-gamified crowdsourcing tasks. These empirical observations provide a basis for narrowing the investigation in Chapter 5, where gamification is examined systematically through a scoping review to determine how its different elements have been operationalised across crowdsourcing archetypes, which motivational mechanisms have been emphasised, and where important theoretical and design gaps remain.

Building on this background and theoretical foundation, section 2.6 presented SDT and Flow Theory as the lenses that support this thesis. Both theories address how autonomy, competence, and relatedness could support motivation and engagement, and how the alignment between task difficulty and user skill promotes enjoyment immersion and attention. These theories thus provide a basis for investigating the motivational task design; however, the literature studies reviewed in this chapter indicate that they have not been consistently applied to the design and assessment in crowdsourcing environments. This chapter demonstrates four linked gaps in the literature. First, while several research in crowdsourcing have explored participation and performance, little comparative research exists on how workers' levels of engagement and performance differ across crowdsourcing task types. In the absence of this baseline knowledge, it becomes unclear which task features or motivational challenges require design intervention. That gap drives RQ1: “How do workers’ performance and engagement vary across different crowdsourcing task types?”

Second, following the exploratory investigation of worker engagement and task-selection preferences in Chapter 4, a more systematic understanding is required of how motivational design has been operationalised in existing crowdsourcing research. Gamification provides a particularly suitable focus for this synthesis because of the breadth of motivational mechanisms that can be represented through its design elements.

However, the literature remains fragmented across task contexts and crowdsourcing archetypes. Specifically, it remains unclear how different gamification elements have been applied across crowd-processing, crowd-rating, crowd-solving, and crowd-creating systems, which motivational mechanisms and theoretical perspectives they reflect, and which areas remain comparatively underexplored. This gap motivates RQ2: “How have the additional task features (e.g., gamification elements) been applied in existing crowdsourcing research across the main archetypes?”

Third, there are still inadequate evidence for how motivational features are systematically applied into the design of a crowdsourcing image-labelling task. Whilst the literature recognises the positive effects of feedback, progression, communication and social support, these are seldom combined into a theoretically derived and experimentally testable model. This design deficit inspires RQ3: “How can such features be systematically designed and integrated to improve workers’ performance and engagement in crowdsourcing image-labelling tasks?”

Fourth, the literature presents us with relatively insufficient support for how they impact the trustworthiness of the crowdsourcing findings, especially with respect to worker performance and thus the data quality. This is the problem at the heart of crowdsourced labelling, because agreement strength and accuracy in so far as their usage directly influences the use of labelled datasets in ML and AI applications that follow. This quality gap drives RQ4: “How do such features affect the reliability of the crowdsourcing outcomes, particularly in terms of worker agreement?”

Taken together, these four gaps represent the motivation for research programme in this thesis, and the methodological approach to answer these research questions is described in the following chapter.

Chapter 3 :Research Methodology

3.1 Introduction

This chapter defines the methodological framework adopted in this thesis. It outlines how the insights from Chapter 2, which reviewed the theoretical and empirical foundations of crowdsourcing, motivation, and task design, informed the overall research strategy. The methodology establishes the basis for ensuring that the research approach is philosophically consistent, empirically robust, and aligned with the thesis objectives.

The objectives of this thesis, as stated in Chapter 1, range from exploring and examining how the crowd workers perform and engage in different crowdsourcing tasks to examining the impact of collaborative task design with motivational features on engagement and performance within crowdsourcing environments. Addressing these objectives necessitated a methodological framework adept at encapsulating the intricacies of human motivation in distributed digital work while guaranteeing reliable and valid results throughout various research phases. Thus, a multi-phase mixed methods approach has been applied in this thesis, with each study built based on the findings of the previous phase study:

1. An exploratory empirical study examining baseline patterns of engagement and accuracy for different types of crowdsourcing tasks.
2. A scoping review synthesising the literature on motivational approaches in crowdsourcing, with particular attention to design features, such as gamification, that have been proposed to enhance participation and performance.
3. A collaborative, gamified experimental study testing the impact of specific gamification features (chat, leaderboards, score feedback, shared goals, and levels) on worker engagement and performance in image-labelling tasks.

The thesis sequentially organises its research, enabling not only a deepened understanding of what is known but also testing the effects of design interventions in controlled conditions.

3.2 Research paradigms and methodologies

The nature of the research problem and the research questions determine the most appropriate paradigms and methodological selection; each paradigm has various philosophical presumptions and pragmatic consequences (Kankam, 2019). Common utilised paradigms in information research are positivism, interpretivism, pragmatism, and post-positivism (Kankam, 2019). The interpretivism paradigm uses qualitative research methods to explore human behaviour, viewpoints, and relationships (Kumatongo and Muzata, 2021).

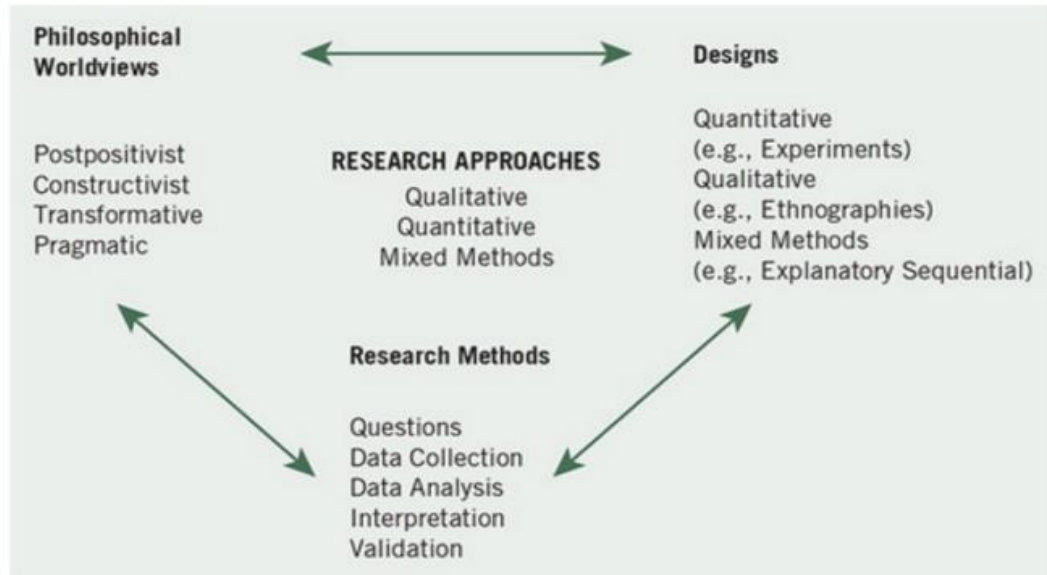


Figure 3.1: The interconnection of worldviews, design, and research method (Creswell and Creswell, 2017)

The positivist paradigm emphasizes objective measurement, hypothesis testing, and quantitative analysis. To properly deal with research difficulties, the pragmatism paradigm uses a framework combining both qualitative and quantitative approach relationships (Kumatongo and Muzata, 2021). Qualitative, quantitative, and mixed methods techniques are the three basic forms of research methodology approaches available. Mostly developed in social sciences, qualitative research emphasizes thoroughly investigating a research issue (Creswell and Creswell, 2017). In qualitative research, several techniques are applied, including grounded theory, phenomenology, ethnography, and case studies. Interviews, observations, recordings, and documents can all be part of qualitative data collection (Creswell and Creswell, 2017). To address hypotheses and problems, the quantitative design centres on variable measurement and analysis of the interactions among variables (Creswell and Creswell, 2017). One can take advantage of methods including surveys and experiments. Quantitative data collection could call for numerical records, checklists, or instrument data.

Overcoming the restrictions of applying a single methodology, the mixed methods approach combines qualitative and quantitative methodologies to offer a more complete knowledge of study issues (Creswell and Creswell, 2017). For instance, while qualitative approaches investigate fundamental causes, quantitative techniques can test the hypothesis and identify the relationships. The mixed methods approach also clarifies the whole nature of the research issue. Using rigorous approaches for data collecting, analysis, and interpretation, it compiles both qualitative and quantitative data (Creswell and Creswell, 2017). A mixed methods approach is suitable for explaining quantitative results, including personal viewpoints, and assessing programmes and policies as well as other aspects (Creswell and Creswell, 2017). Furthermore, three different kinds of mixed methods design exist: the convergent, the explanatory sequential, and the exploratory sequential

(Creswell and Creswell, 2017). Figure 3.1 shows the relation between the research methods, paradigms, and the design. For research that focuses on understanding complex human–technology interactions, pragmatism is a well-suited paradigm, as it emphasises the generation of actionable insights rather than strict adherence to a single philosophical doctrine (Morgan, 2014). Thus, pragmatic research is the paradigm that guides the methodological choices underpinning this thesis. In the context of crowdsourcing, this paradigm offers the adaptability to include many approaches, where the workers’ engagement and performance are affected by several motivational factors and social technological elements

From an epistemological perspective, a pragmatic position means recognising that knowledge of crowd work is both constructed and quantifiable. Qualitatively, we investigate how individuals understand and execute crowd tasks and on a more general level, behavioural measures and survey responses offer supplementary and confirmatory quantitative evidence (Creswell and Creswell, 2017, Johnson et al., 2007). This perspective is consistent with the mixed methods framework utilised in our research by blending experimental methods and systematic literature synthesis to achieve an integrative understanding of the research problem (Morgan, 2007).

Ontologically, this thesis suggests that crowdsourcing platforms are dynamic, socio-technical systems, where the worker behaviour emerges through the ongoing interaction between human and technological infrastructures. Such systems cannot be adequately explained through purely deterministic models nor through exclusively subjective accounts. Epistemologically, meaningful understanding requires different viewpoints and data sources (Kittur et al., 2013, Orlikowski and Iacono, 2001). Thus, pragmatism allows the focus of the research to be amended through certain iterations and insights from one study as a guide to the further development and design of the subsequent study without sacrificing the philosophical coherence. This flexibility renders pragmatism a highly fitting basis for multi-phase explanatory sequential mixed methods research in which evidence accumulates progressively and integrates across phases (Creswell and Creswell, 2017, Maxwell and Loomis, 2003). This makes it a consistent base for explanatory sequential mixed-design methods, as we shall see in subsection 3.3.

3.3 Methodological Framework: Multi-Phase Mixed Methods Research.

This thesis adopts multi-phase mixed methods framework guided by the explanatory sequential logic, which is a widely used approach that requires integrates quantitative rigour with qualitative contextual understanding. The design consists of three interrelated studies, each study developed based on the findings of the preceding study, allowing evidence generated in earlier phases to inform the design, focus, and interpretation of subsequent phases.

The use of multi-phase explanatory sequential mixed methods framework is motivated by the complexity of crowdsourcing contexts, where the worker performance and engagement are influenced by technical, social, and motivational factors. A focus on quantitative results solely would mask the context of

implementation, whereas totally qualitative accounts may lack generalisability. The multi-phase mixed methods research combines the strengths of the two research styles, enabling an in-depth exploration and iterative refinement of the research questions. The framework is organised in the following manner:

Exploratory Empirical Study (Study 1): A mixed methods analysis to exploring the baseline patterns of engagement, accuracy, and task selection with different criteria across various types of crowdsourcing tasks. The results facilitated the identification of critical factors affecting worker performance, motivation, and engagement.

Scoping Review (Study 2): This study used qualitative coding supported by descriptive quantitative summaries of current studies regarding the use of motivational and design strategies in crowdsourcing across various archetypes. This phase integrated knowledge from current research, identified theoretical and empirical gaps, and established the conceptual foundation for the concluding experimental study.

Collaborative Experimental Study (Study 3): A controlled, mixed methods analysis of motivational factors in collaborative image-labelling tasks. The quantitative measures recorded behavioural and performance-based impact; qualitative data (e.g., worker feedback, chat transcripts) provided insight about workers' perception and experiences.

The design facilitates a gradual alignment of theory, evidence, and practice by structuring the research in consecutive studies. Every study refines the inquiry's focus, establishing a structured pathway that progresses from understanding the baseline operations to incorporating previous information and knowledge, leading to the systematic evaluation of various interventions. This framework aligns directly with the dissertation research questions, including descriptive, exploratory, and evaluative dimensions.

3.4 Methods for Data Collection and Analysis Instruments

This thesis' strategy for gathering material intended to establish robust evidence based on engagement and performance in crowdsourcing. According to the multi-phase mixed methods research approach guiding by explanatory sequential logic, three different studies were carried out, each using appropriate datasets and instruments and yet keeping the major indicators consistent across studies to make comparisons more meaningful.

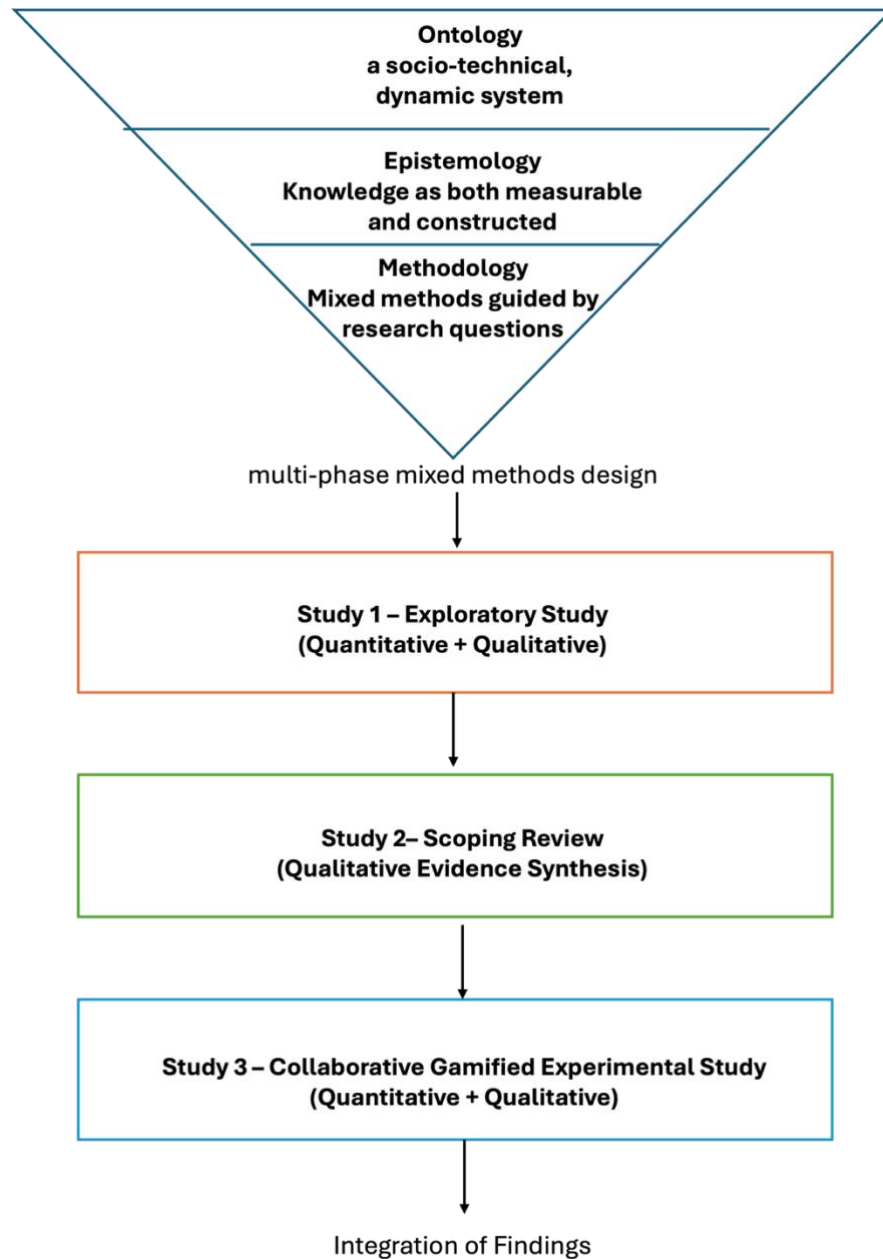


Figure 3.2: Integration of the pragmatic paradigm with the multi-phase mixed methods research design adopted in this thesis.

3.4.1 Study 1: Exploratory Study of Different Tasks

In the first study, we investigated baseline engagement and performance for three widely used crowdsourcing task types—text classification, image classification, and surveys—implemented on Amazon Mechanical Turk. We selected these task types because they represent real-world task markets (Gadiraju et al., 2014) and cover both content annotation (text, image) and reflective survey work, which differ in cognitive demand and worker propensity.

Data sets:

Text classification: CrisisNLP collection of disaster-related tweets (Imran et al., 2016). This choice was made because CrisisNLP is concerned with humanitarian response, where accurate tagging of crisis information can make a real difference. It is also a classic area for public "voluntarism" or donated effort. CrisisNLP's nine-class schema is well enough for our purposes to apply for untrained workers. Image classification: Stanford Dogs collection (Khosla et al., 2011), small because fine distinctions are very demanding on the worker's mind and so fit such benchmarking tasks as crowd-sourced annotation of ML training sets. Its eminent verity also makes it good for measuring engagement levels and accuracy, as underlined by our piece on behavioural data. Survey tasks: E-learning survey on experience during the COVID-19 Pandemic. Included here for comparison with annotation content tasks is a reflective self-reporting task type largely found in online venues.

Quantitative instruments:

Accuracy: direct comparison with gold-standard tags where available gave an impartial measure of performance on the task (Sammut and Webb, 2011, Snow et al., 2008).

User Engagement Scale – Short Form (UES-SF): This scale has been chosen for its efficacy in capturing the degree to which people engage with digital environments and includes such variable factors as focused attention, perceived usability, and reward (O'Brien et al., 2018).

NASA Task Load Index (NASA-TLX): This widely recognised scale measures perceived workload across six dimensions (Hart and Staveland, 1988). Here it is applied to varying task types.

Big Five Personality Inventory (BFI-10): This is a validated, concise test measuring personality traits (Rammstedt and John, 2007). The reason for including it in the study is to see whether stable individual differences modulate engagement and accuracy.

Qualitative instrument:

Open-ended feedback: This allowed the workers to discuss difficulties, make suggestions, or suggest preferences, so contextual understanding of quantitative scores could be added.

3.4.2 Study 2: Scoping Review of Motivational Features in Crowdsourcing

The second part of the thesis conducted a systematic scoping review of existing literature to synthesise what is known about using motivational resources in crowdsourcing systems. The review followed (Arksey and O'Malley, 2005) and methodology framework, which provides a systematic way to search, screen, and synthesise appropriate studies.

Data sources and search: Searches were conducted in Scopus, Web of Science, ACM Digital Library, and IEEE Xplore to provide coverage for both computer science-based research as well as social science. Inclusion and exclusion criteria were established to generate coverage of studies describing motivational or gamified crowdsourcing task design.

Gamification elements: The framework that our extraction and classification approach is based on is the taxonomy identified by Koivisto and Hamari (2019) which offers a comprehensive methodology for understanding gamification as a motivational design framework, allowing us to aggregate insights from the overall findings. This taxonomy differentiates between social, achievement/progression, immersion and economic aspects, and allows for a uniform categorisation of game design elements in the studies analysed.

Analytical methods: The extracted data were analysed to identify the patterns in all CS task archetypes, motivational engines, theoretical framework underpinnings, and reporting of the outcomes. This mapping enabled the detection of gaps in research, including the under-exploration of collaborative motivational elements in the context of crowdsourcing.

Therefore, the scoping review acted as a bridge between our preliminary work (Study 1), which identified intrinsic interest as being key for task engagement and interest-related motivational features to possibly be used, and the co-designed experiment (Study 3) that systematically examined specific motivational properties in a more controlled setting.

3.4.3 Study 3: Collaborative Experimental Study with Motivational Elements

In this phase, we implemented a custom-built crowdsourcing prototype system with collaborative image labelling and integrated motivational features drawn from Studies 1 and 2.

Data set:

The Galaxy Zoo astronomical image data set (Lintott et al., 2008). We selected this data set for three reasons: (1) it provides a typical example of large-scale citizen science annotation to compare with our method; (2) classification is based on human judgment rather than any specialist training, so Amazon Mechanical Turk workers can do it, and (3) it aligns well with previous research in collaborative labelling methods that make annotation feasible based on large numbers of participants.

Quantitative instruments:

Accuracy and Agreement Strength: These are standard performance measures for crowd-sourced annotation tasks (Snow et al., 2008). We applied them for consistency among workers using majority voting and Cohen's kappa.

Engagement: The assessment of engagement in this study, unlike Study 1, which employed the UES-SF as a general engagement measure, Study 3 adopted a theory-informed engagement framework based on Self-Determination Theory and Flow Theory. These constructs were:

- Autonomy (e.g., “I did things in the game because they interested me”),
- Competence (e.g., “I am satisfied with my performance for the task”),
- Relatedness (e.g., “When I visit the online community, I feel supported by other users”),
- Enjoyment (e.g., “I enjoyed this task very much”),
- Perceived usability (e.g., “I found this task confusing to do”).

Items were primarily adapted from the Intrinsic Motivation Inventory (Deci and Ryan, 1985, McAuley et al., 1989), and previous engagement studies in the HCI and gamification (O'Brien et al., 2018). This adaptation ensured that the gamified experiment of engagement measures was in line with the needs of motivation embedded in the design of the system (chat, leaderboard, scoring, and levels); it reflects not only SDT-related needs such as autonomy, competence, and relatedness, but also HCI-related aspects, enjoyment, and perceived usability.

NASA-TLX: This is the measure of perceived effort of workload across six dimensions (Hart and Staveland, 1988). Here it is applied to varying task types.

Qualitative instruments:

Chat transcripts with open-ended feedback responses were analysed to capture patterns of collaboration and provide supplementary contextual information.

3.5 Integration and Consistency

In the data collection method, two general methodological principles were used to guide us for the three research studies: triangulation and continuity. These two principles ensured that our findings were methodologically robust and conceptually coherent across all research study designs.

First, triangulation of data and instruments was used in improving the credibility and validity of the findings. By using the principles of the methodological triangulation (Denzin, 2017), multiple forms of data were collected in each study, including objective performance logs, self-report survey measures, and qualitative feedback. A variety of data sources (e.g., task accuracy and completion records), perceptual measures (e.g., engagement and workload scales), and open-ended participant responses allowed cross-validation of findings and reduced the risks associated with relying on one data source or measurement perspective. This sort of triangulation is especially critical in research that investigates complex human–technology interactions where no one method can capture users' behaviours, perceptions, and experiences (Creswell and Clark, 2017). The use of qualitative evidence paired directly with quantitative data allowed us to corroborate and explain these outcomes across different types of data sets and provide a richer and deeper analysis of worker engagement and performance in crowdsourcing contexts.

Second, the continuity of the key constructs was maintained across the research programme of the thesis to facilitate meaningful comparisons and cumulative knowledge building. The constructs of user engagement, perceived workload, and individual differences were examined across the empirical studies; based on validated measures. The scoping review extended these constructs by exploring how motivational mechanisms and design strategies have been studied in previous crowdsourcing work. Meanwhile, study-specific instruments were also added whenever needed to take into account the unique goals of the individual study. For example, we used coding in a systematic manner in the scoping review to categorise motivational design components; chat log analysis and post-task qualitative feedback were integrated in the collaborative

gamified experiment to document social interactions and coordination dynamics, etc. This trade-off between consistent core measures and context-relevant instruments is also common in mixed methods research, where stability in key components of the study is combined with a methodological flexibility to tackle research questions (Creswell and Creswell, 2017, Maxwell and Loomis, 2003). Overall, employing triangulated sources of data and a consistent set of measurement instruments ensured that the insight gained in the earlier studies would inform later investigations studies without compromising methodological coherence. This approach supports the multi-phase mixed-method nature of the research design, increasing the credibility of the cross-study synthesis in subsequent chapters.

3.6 Ethics Consideration

Ethics were granted via department approval for an exploratory study into different crowdsourcing tasks (Application ID: 1399). Approval was granted and received from the Department of Computer and Informatics Sciences Ethics Committee. Collaborative gamified image-labelling task (Application ID: 2548). For both empirical studies, all participants provided informed consent electronically before tackling the tasks. Participation was voluntary, with no requirement to finish any part of the study and subjects reserved the right to withdraw at any point before they actually joined the task. Workers were recruited via Amazon Mechanical Turk, where tasks were posted with clear descriptions of the requirements, expected duration, and rewards. The compensation rates were set to align with the recommended ethical practices for crowdsourcing research. To protect individual privacy, no personally identifiable information was collected other than worker IDs issued by the platform. Those were replaced by random identifiers during analysis as anonymisation measures. All information was kept secure on institutional servers and survey platforms, only available for access to the researcher or supervisors. Furthermore, it has been processed in compliance with the General Data Protection Regulation (GDPR). Aggregated results are reported to protect individual anonymity, and the datasets are reserved only for academic research and dissemination.

3.7 Conclusion

This chapter has provided the methodological framework for this thesis and presented a philosophical position, research design processes, data collection methods, integration and consistency, and ethical considerations. The study was based on a pragmatic paradigm, allowing methodological flexibility and coherence from the beginning until the end of the process. A multi-phase mixed methods design guided by explanatory sequential logic that combines qualitative and quantitative data in three consecutive studies—an exploratory study across different crowdsourcing tasks, a scoping study of motivational features in crowdsourcing work, and a cooperative gamified experiment investigating characteristics of task design—was followed. Data were gathered during these studies through performance metrics, validated instruments, and qualitative sources such as open-ended feedback and communication logs. This triangulation approach guaranteed the inclusion of not

just behavioural consequences and subjective experiences but also a holistic view of what drives workers' performance and engagement. This research received ethical clearance for all human subjects' studies and procedures ensured that all participants provided informed consent, were offered total privacy and were treated fairly. This methodology is implemented throughout the thesis research via the same metrics of engagement, workload, and performance, facilitating systematic comparison across the experimental settings.

Overall, the approach that has been followed provided a robust and ethically defensible foundation for answering the research questions. The next chapter (Chapter 4) describes our first empirical study on engagement and performance across different crowdsourcing task types to serve as a baseline for further comparison.

Chapter 4 :Exploratory Study of Worker Engagement in Different Crowdsourcing Tasks

4.1 Introduction

Based on the methodological framework explained in Chapter 3 of this thesis which details the general research design and mixed method approach employed, this chapter describes the first exploratory empirical study. This first exploratory investigation serves as the initial study of the thesis' follow-up research process, thereby providing an evidence-based understanding of how workers perform and engage in different crowdsourcing tasks.

Crowdsourcing is a well-established platform for performing large-scale tasks online by dividing the work among a large pool of people over the Internet. Various factors such as type of task, task design, complexity, and clarity of instructions affect workers' performance, engagement, and motivation. Studying the behaviour of workers when dealing with different tasks is important for optimising crowdsourcing systems and building cost-effective task environments. The research serves as a base for later experimental and design-oriented studies through identifying worker motivational and behavioural patterns that may affect their outcome quality and their willingness to participate in the task. Although there have been substantial studies that compared task properties such as task cost, duration, or payment rate, few studies have investigated how these properties interact with worker preferences and perceived interest during task selection and how the effect of multiple selections or ordering of the tasks affects their performance and engagement.

4.2 Study Objective

Accordingly, the aim of this study is not to assess or evaluate a specific system but first to gain preliminary insights regarding the performance differences across task types and about the workers' task selection behaviour and the possible reasons that make some tasks more attractive than others. In particular, this study focuses mainly on exploring how crowdsourcing workers perform and engage across different task types, including image classification, text classification, and survey tasks. The primary aspects of this examination include:

- Examining the relationship between workers' demographic features or personality traits and their performance outcomes and engagement.
- The influence of undertaking multiple diverse tasks on workers' workload is also a crucial consideration.
- The preferences that workers exhibit when selecting tasks on a crowdsourcing platform

Taken together, these features respond to the main research question (RQ1) and the sub-research questions:

RQ1: How do workers' performance and engagement vary across different crowdsourcing task types?

The core research question is further refined into the following sub-research questions (SRQ) to help break down the exploratory aspect of this study and enable us to explore different dimensions of worker behaviour.

SRQ1.1: How does worker engagement, performance, and perceived workload differ across different crowdsourcing task types?

SRQ1.2: Do worker personality traits correlate with their accuracy and engagement across different crowdsourcing tasks?

SRQ1.3: How does multiple-tasking behaviour affect worker engagement, performance, and perceived workload on crowdsourcing platforms?

These sub-research questions enable the systematic investigation of task-, individual-, and behavioural-level influences on crowd work outcomes. The most critical finding from this study was that workers always rated “task interest” as one of their central concerns (second only to monetary reasons). This result emphasises the role of task design characteristics in making tasks interesting and enjoyable. This motivational insight supports the theoretical emphasis on intrinsic interest found in SDT and Flow Theory and is the key driver for the subsequent next phase of the thesis. Accordingly, this chapter not only provides the behavioural baseline to comprehend engagement and performance within crowdsourcing but also motivates the subsequent scoping review (Chapter 5) that examines how these motivational characteristics have been operationalised across different types of crowdsourcing. This helps to identify the optimisation strategies and serves as an important diagnostic step in the overall thesis, bridging basic behavioural insights with practical design methods with the goals of enhancing performance and motivated participation, thereby improving data reliability.

4.3 Study Design

To test and answer the research question, we explored the crowdsourcing platforms and the worker performance and contributions across different task types by running three experiments as follows:

The study did not feature an experimental manipulation but used a descriptive and comparative approach with the intention of capturing natural variations of behaviour (e.g., performance, workload, engagement) among different task modalities collected from 750 crowd workers. The three experiments were intended to sample different task modalities: text-, image-, and survey-based tasks that have been frequently used on crowdsourcing platforms.

The theory behind the choice of different task modalities was also influenced by SDT and Flow Theory. These theories highlight autonomy, competence, intrinsic interest, and attentional absorption as major drivers of engagement. The aim of the experiment was to study how different types of crowdsourcing tasks, each involving distinct cognitive demands of worker interaction, affect the levels of engagement and performance. Understanding these baseline differences was critical for identifying motivational challenges associated with

various forms of crowd work. This laid the empirical groundwork for the literature synthesis presented in Chapter 5 and directly informing the development of the thesis research programme.

4.3.1 Crowdsourcing Platform Selection

The extensive utilisation of Amazon Mechanical Turk in academic and industry investigations is due to its simplicity in task deployment and its access to a diversified worker participant pool, which was the principal reason for selecting the Amazon Mechanical Turk as the platform for this research. Amazon Mechanical Turk enables the researchers to broadly publish the Human Intelligence Tasks while implementing qualification filters to identify certain worker profiles. Numerous studies have demonstrated the platform's suitability for behavioural tests, data annotation tasks, and gamification-based evaluations, particularly in research related to human computation and collaborative systems (Buhrmester et al., 2016, Chandler et al., 2014, Peer et al., 2017).

The use of Amazon Mechanical Turk also ensured methodological consistency with succeeding chapters. This consistency facilitated an equitable comparison of worker behaviour across the exploratory trial and subsequent intervention-based investigation studies. Moreover, Amazon Mechanical Turk was chosen as it provides a naturalistic work environment where workers freely choose tasks depending on different preferences necessary for observing intrinsic and extrinsic motivators outlined in the SDT (Ryan and Deci, 2000).

4.3.2 The Selection of the task types

Crowdsourcing task-type taxonomy defined by Gadiraju et al (2014), employed hierarchical levels of categorisation: the first categorisation was based on the purpose of the design, and the second classed tasks based on how they were performed or the job's workflow. According to Difallah et al., (2015), the authors mentioned that the most popular form of work over the last few years had been content creation. In contrast, the use of surveys, interpretation, and analysis has expanded fast since the 2009 (Difallah et al., 2015). Based on that, the tasks employed in our experiments were text and image categorisation and survey tasks. These tasks were chosen not just because they were simple but also because they could be replicated with the same GUI task.

Each task type was also chosen as a representation of a distinct cognitive modality that is commonly used in crowdsourcing studies: (1) linguistic interpretation for text classification e.g., (Imran et al., 2016, Snow et al., 2008), (2) perceptual discrimination for image classification e.g., (Deng et al., 2009, Khosla et al., 2011). and (3) reflective judgement for survey tasks e.g., (Chandler et al., 2014, Difallah et al., 2018, Peer et al., 2017), correspond to different aspects of cognitive workload (Hart and Staveland, 1988), and present diverse opportunities for intrinsic interest and competence, the fundamental aspects of SDT and Flow Theory.

Fine-grained image classification tasks, such as dog breed identification, are extensively employed to examine perceptual difficulties and worker annotation behaviour in crowdsourcing (Khosla et al., 2011, Welinder et al., 2010).

Tweet classification is a common form of microtasking in crowdsourcing (Snow et al., 2008). Particularly for crisis tweets, a wide range of such datasets are readily available to examine annotation reliability under ambiguity and linguistic variability (Imran et al., 2016). This complexity facilitates investigation of SDT's competence dimension, since annotators have to make sense of short, noisy texts in the presence of uncertainty.

Survey tasks were added, as they form a major segment of paid microtasks (Paolacci and Chandler, 2014) and are often more mentally demanding. This allows researchers to assess workers perceived workload and level of frustration, foundational constructs in NASA-TLX and SDT's competence and autonomy dimensions. Another feature to be seen in survey tasks, alongside the perceptual and linguistic demands of the other two task types, is that they capture workers' reflective decision-making behaviour. To ensure comparable task complexity across various task types, we adhered to the task complexity parameters delineated by Aipe and Gadiraju (2018). The dimensions included task completion time, task batch size, and task completion reward. To quantify the design and degrees of difficulty of those tasks based on task completion time, give a sense of the likely response pattern, and it has been well established in a pilot of this study on our college colleagues was carried out with 10 sample sizes on the Amazon sandbox, followed by 25 real Amazon Turk workers per task. The result showed the average time to complete a task with the same batch size (ten) and was used to estimate the reward per minute in the main study. We used Amazon Mechanical Turk to provide equitable compensation for crowd-workers; we remunerated them according to the projected duration at the standard minimum pay of \$8.50 per hour (Hara et al., 2019). The tasks were designed to possess a moderate level of difficulty, and the workloads were assessed to ensure they did not induce excessive stress. This judicious balance of difficulty, reward and batch length ensured that all task types could be compared meaningfully and without environmental confounders of fatigue or insufficient payment.

This study complied with the ethical guidelines approved by the Department of Computer and Information Sciences Ethics Committee (Application ID: 1399). Participants were given an information sheet and consent form at the start of each task. Data storage and processing followed the ethical guidance for research at the University of Strathclyde as well as GDPR data protection regulations.

4.4 Experiments Setup

In our experiments, we used Amazon Mechanical Turk's core interface templates to construct and implement the task batches to maintain a consistent design aesthetic. This involved an HTML template customisation supplied by Amazon Mechanical Turk to integrate the consent, links, and instruction forms into the tasks. The interface of the task provided explicit instructions for participation criteria, including the quantity of questions

or images to be tagged and the link for the post-task survey. The HITs were designed with defined completion time constraints based on the pilot study on our university colleagues with the average of 10 min completion rate. As discussed, the choice to implement Amazon Mechanical Turk was bolstered by its capability for extensive job execution, enabling us to oversee recruitment, monitoring, and compensation via a centralised and meticulously documented structure. One batch for each of the three task types has been published. The three tasks are:

1. **Text classification:** workers are asked to read ten tweet sentences and then select the proper label from the nine annotations (described in Section 4.6.1) Figure 4.1.
2. **Image classification:** workers were shown ten different dog images and asked to select the proper breeds (described in Section 4.6.2) of the dogs that appeared in the images, Figure 4.2.
3. **Survey tasks:** the workers will answer a multiple-choice question regarding the E-Learning experience during the pandemic of Covid-19, Figure 4.3.

At the end of each task, the workers were asked to fill out the Qualtrics survey and copy the given code back to the Amazon Mechanical Turk task. We set up the number of assignments in batches (number of workers assigned for each task type) to be 250 assignments with ten minutes per time allocated for each, and we paid every worker 1.5 \$ per ten-minute assignment. The time and payment were decided based on the time estimations gathered from our pilot study and the highest state minimum wage of the United States of \$8.5/h (Hara et al., 2018). A total of 750 workers finished the tasks. The unified design of all three task experiments enabled cross-comparison and offered a comprehensive perspective on the impact of task type on worker experience. The implementation of a post-task questionnaire immediately following each activity facilitated

Classification of Tweeter Contents

[View instructions](#)

The goal of this task is to categorize each tweet into one of the nine categories given below. These tweets were collected during some crisis emergency situation. Please, read the text carefully and decide only based on the content of the tweet, you do not need to access any of the URLs in tweets

☐ I confirm that I have read, understood and agree the above policy and procedure for enrollment in this task

Choose the right category for the following tweets:

Tweet 1

* RT @Faiz_Baluch: #BalochistanEarthquake Pakistan army is terrorizing the people by aerial firing in #Awaran <https://t.co/R3R5ph8zSS> <http://>

-- Please select the correct category for the Tweet --

Tweet 2

#Earthquake 2013-09-28 02:39:43 (M5.0) EAST OF THE SOUTH SANDWICH ISLANDS -59.5 -19.1 (70fa9) <http://t.co/uBN98fFmNj> notice

Showing Task 1 of 10 [Next HIT](#)

Figure 4.1: Tweet classification task

Image Classification

[View instructions](#)

The goal of this task is to view the images and categorize them each based on the dog breeds

☐ I confirm that i have read, understand and agree the above policy and procedure for enrollment in this task

Choose the right breed for the following dogs in the images:

image 1



Showing Task 1 of 10 [Next HIT](#)

Figure 4.2: Dog classification task

Students perception of E-learning during pandemic Survey

[View instructions](#)

The main objective of this survey is to assess the perception of higher educational students towards E-Learning during Covid-19 pandemic.

☐ I confirm that i have read, understand and agree the above policy and procedure for enrollment in this survey

Question 1

Age in years

☐ Less than 25

☐ Greater than or Equal to 25

Question 2

Type of residence

☐ Urban

☐ Rural

Figure 4.3: Survey task

the acquisition of immediate perceptions regarding workload and involvement, hence mitigating recollection bias. Moreover, since the deployment of tasks was uniform, differences in behaviour could be linked to the task types, rather than the interface or environment. This design also provides an empirical baseline for the subsequent studies of motivational mechanisms, as I discovered which types of tasks the workers found more or less engaging.

4.5 Participants and Procedures

All workers in the Amazon Mechanical Turk platform could participate in the experiments of this study, with no restrictions on whether they had a particular degree or level of expertise. No information could be used to identify any workers. Worker participation was purely voluntary, and they were able to leave at any time. In the experiments, we released a distinct batch of HITs for each task type previously described, allowing workers to select them from the Amazon Mechanical Turk worker dashboard interface. The workers could choose any task type of interest, and to mitigate memory bias, we allowed them to engage in only one instance of each task type. Upon completion, the workers were mandated to finalise a brief post-task Qualtrics survey and return the provided code to Amazon Mechanical Turk to incentivise survey completion. This open participation structure mirrored the natural working environmental model of crowdsourcing platforms, where workers select the tasks to perform depending on what they find interesting, challenging and rewarding. Allowing the workers' self-selection for the tasks was methodologically premeditated that allowed the identification of the behaviour pattern and task preference, which guided the later scoping review on motivation features in Chapter 5. The identity of each participant was connected across the Amazon Mechanical Turk task and Qualtrics survey using an anonymised code to protecting confidentiality. Demographic information was gathered from the post-task survey to situate analyses of engagement and performance across the worker groups. The demographic data were gender, age, education level, the device used in tasks, the years of experience on Amazon Mechanical Turk. This data was subsequently compared to established benchmarks from prior related studies e.g. (Difallah et al., 2018, Hara et al., 2018, Martin et al., 2017) to verify representativeness.

A total of 750 workers took part in the study, performing tasks from each type. This sample size was representative of diverse participants for comparison and to guarantee a reasonable level of statistical power in detecting performance and engagement disparities across tasks. Similarly, this sample size allowed for stable non-parametric inference across demographic segments and is well-known in human computation research, where comparisons of performance can be achieved multiple times (Kittur et al., 2013).

4.6 Datasets

The datasets used in this study included annotations by experts (serving as the truth labels), allowing us to compare worker labels to the ground truth standard to determine differences in performance. As stated, the study focused on the type of categorisation work (text and image) with nine and ten classes. In the selection of the datasets, three principles were followed: (1) The most common crowdsourcing scenarios, (2) The availability of the ground truth standard annotations, and (3) They all have varying cognitive loads and levels of end user involvement.

These data sets were suited for not only analysing objective results (accuracy) but also the subjective experience of worker efforts (workload and engagement). This also aligned with motivational theories such as

SDT and Flow Theory, as each dataset represented a different modality through which autonomy, competence, and intrinsic interest may naturally vary.

4.6.1 Text-based classification Dataset

The tweet dataset (Dataset 1) consisted of tweets collected during a crisis or disaster (Imran et al., 2016). This dataset contained roughly 52 million disaster-related texts and were collected from 19 different disasters between 2013 and 2015. The scheme of annotation was used to categorise each tweet content into one of nine possible categories:

1. Injured or dead people: Reports of injured people due to the crisis.
2. Missing, trapped, or found people: Reports of questions about found or missing people.
3. Displaced people and evacuations: People who have relocated due to the crisis.
4. Infrastructure and utilities damage: Reports of damaged bridges, buildings, roads, or interrupted services.
5. Donation needs or offers or volunteering services: Reports of urgent needs or donations of shelter and supplies such as food, clothing, water, money, blood, medical supplies, and volunteering services.
6. Caution and advice: Reports of warnings issued.
7. Sympathy and emotional support: Thoughts, emotional support, and prayers.
8. Other useful information: Other useful information that helps understand the situation.
9. Not related or irrelevant: Unrelated to the situation or irrelevant.

This dataset was chosen since text classification is among the more common forms of microtasks in crowdsourcing. Short-text labelling tasks such as tweets are commonly employed to investigate annotation behaviour because of their brevity, ambiguity, and linguistic variability (Snow et al., 2008). It also involves cognitive interpretation and linguistic comprehension; hence the study could capture subtle aspects of worker effort and focus. Moreover, crisis tweet datasets are often used to study crowdsourcing researcher handling of complex semantic cues (Imran et al., 2016, Olteanu et al., 2014).

4.6.2 Image-based classification dataset

This dataset comprises images of 120 dog breeds worldwide included in the Stanford Dogs dataset (Khosla et al., 2011). For the aim of fine-grained picture categorisation, this dataset was created utilising images and annotations from ImageNet. It was initially gathered to aid with fine-grain image categorisation. We choose ten categories from the dog breeds to have the same label numbers of the tweet labels and avoid task abundant, which are Labrador Retriever, Golden Retriever, Yorkshire Terrier, German Shepherd, French Bulldog, Standard Poodle, Beagle, Doberman, Boxer, and Pug.

This dataset allowed the examination of the visual-perceptual classification tasks, which is cognitively different from the text interpretation of the tweet dataset. So, the high-quality and distinctive visual images provided a good balance between difficulty and recognisability, ensuring an equal platform for accuracy comparison against text-based tasks. Tasks to classify fine-grained images, for instance, dog-breed

classification, stand as a standard benchmark in human computation research, requiring concentrated visual attention and perceptual discrimination (Welinder et al., 2010). This cognitive transition from linguistic reasoning to perceptual processing creates a stark counterpoint for analysing engagement and workload, underpinning the multidimensional design of this exploratory study.

4.6.3 Survey Dataset

The third dataset (Dataset 3) surveys about the E-learning experiences over the 2019-2022 pandemic with related global lockdown restrictions. Survey-based tasks were included to represent analytical and reflective task types which are more common on crowdsourcing platforms. The engagement of the workers in these tasks will be different from other classification tasks, will bring different skills into plays for workers as self-reporting and reflection instead of direct labelling accuracy. The survey dataset therefore created a third dimension of intellectual involvement that complemented the text and the image classification tasks. It allowed workload and engagement in task types to be measured in greater detail. Survey tasks are also a big component (Paolacci et al., 2010, Chandler et al., 2014) of Amazon Mechanical Turk and therefore are ecologically valid for investigating crowd worker perceived load and decision making. By designing a reflective and question-based dataset, the study was able to access the greatest intensity of mental demand and temporal pressure (Hart and Staveland, 1988) and thus, the contrast that was achieved with perceptual and linguistic tasks.

To summarise, the combination of the above datasets offers a multidimensional model in studying the effects of different information modalities on worker performance, engagement, and perceived workload by integrating linguistic, visual, and analytical modalities. These datasets facilitated explorations of the influences of task design upon motivation, cognitive difficulty, and interest in task elements that are subsequently highlighted in Chapter 5, where we identified gaps for motivational stimulation.

4.7 Evaluation metrics

Over the last decade, various human computation assessment methods have been created, each of which serves a particular function. Our work was evaluated using a combination of these and other methods. In this section, we specify the metrics and analytics to measure performance, workload, engagement and personality-based aspects of workers across various task types. Quantitative and qualitative studies were

Table 4.1: Confusion matrix for positive and negative labels

		Predicted Labels of Workers	
		Positive	Negative
Gold Standard (original class)	Positive	TP	FN
	Negative	FP	TN

conducted to obtain a complete understanding of worker behaviour. The methodologies adopted are aligned with established crowdsourcing evaluation frameworks and provide assurance of the reliability and validity of the results.

4.7.1 Performance measures: 1- Accuracy

We employed statistical methods that have been used with binary and multi-classification crowdsourcing projects to evaluate worker performance. Accuracy is one such metric. Workers predicted labels were compared to the labels in the gold standard; this measure was utilised to calculate their score. Accuracy is alternatively expressed as the ratio of correct predictions to errors, as to accuracy = 1.0 - error rate. The confusion matrix is used in the classification problem to calculate the accuracy, where the gold standard label is mapped to the anticipated class, as illustrated in Table 4.1. Where True Positive (TP) is the number of correctly labelled items as positive classes by workers, and True Negative (TN) is the number of correctly marked items as negative classes by workers. False Positive (FP) is the number of items incorrectly labelled as positive by workers, whereas False Negative (FN) is the number of items incorrectly labelled as negative by workers. The accuracy score was obtained using the confusion matrix as the true annotated items proportion (including true negatives and true positives) in the total number of evaluated items, as shown in Equation 4.1.

$$Accuracy = \frac{(TP+TN)}{(TP + TN + FP + FN)} \quad \text{Equation 4.1}$$

We picked accuracy, since it provides a simple and intuitive measure for the quality of labelling which can be used in both image and text classification similarly. Accuracy has also been a widely adopted metric in crowdsourcing literature due to its interpretability, which allows comparisons of worker outputs to expert annotations (Snow et al., 2008, Gadiraju et al., 2017).

4.7.2 NASA Task Load Index (NASA-TLX) evaluation

The workers were required to fill out a questionnaire on the Qualtrics platform about some demographic information at the end of each task. They perceived performance using the NASA-TLX tool to assess workers perceived workload (Hart and Staveland, 1988), where the worker assesses their feeling toward the task workload. This test also allowed us to assess the impact of batch design on task difficulty. A total of six subjective questions were included in this survey. These six dimensions of NASA-TLX, Mental Demand, Physical Demand, Temporal Demand, Performance, Effort and Frustration, enabled the quantification of workers perceived workload among the three task types. These measurements have given clear subjective information in addition to objective accuracy data, revealing how the workers experienced each task's cognitive and effective load. NASA-TLX was chosen because it is one of the most widely validated subjective workload tools in Human Computer Interaction (HCI) and crowdsourcing research, allowing evaluation of the perceived load, fatigue, and task complexity.

4.7.3 User Engagement Evaluation

In the field of HCI, measuring is becoming increasingly important. The User Engagement Scale (UES) is a tool that has been used in a variety of digital domains to quantify UE. We assess worker involvement in this study using a standardised questionnaire UES, which was developed by a group of researchers in (O'Brien et al., 2018). This version's UES items were suitable to be adopted in the particular crowdsourcing context to capture effective and behaviour indicator of the engagement. The UES encompasses experiential characteristics including focused attention and perceived usability, which relate to core elements of Flow Theory and explain how and why some tasks lead to greater levels of intrinsic engagement compared to others.

4.7.4 Personality traits evaluation

As part of the questionnaire, we included short-scale questions from (Rammstedt and John, 2007) to collect personality trait information on the workers. We used a standard ten-question test used to measure the five personality traits which are, Openness, the desire to be inventive, autonomous, and interested in variation (high score) instead of practical, conforming, and interested in routine (low score). Conscientiousness is the inclination to be self-disciplined, act responsibly, and be organised, cautious, and disciplined (high score) instead of being disorganised, thoughtless, and impulsive (low score). Extraversion is the inclination to be gregarious, fun-loving, and affectionate (high score) as opposed to retiring, solemn, and quiet (low score). Agreeableness is the ability to be empathetic, cooperative, trusting, and helpful (high score), while the low score indicates someone who is highly competitive, analytical, who generally prioritizing their own needs above group harmony. Neuroticism is the inclination to be calm, secure, and self-satisfied (low score) vs. nervous, insecure, emotionally unstable, and self-pitying (high score). By including personality data, we can deepen our understanding of why individuals differ in task engagement and accuracy. Analysis later verified it further, with relational and regression studies questioning whether inner dispositions predict performance across different types of tasks. Personality data also endorse motivational interpretations according to the effect of dispositional tendencies on the likelihood of workers achieving competence, experiencing flow, or intrinsic motivation (Kazai et al., 2011).

4.7.5 Statistical Analyses

To rigorously test the main research question and its sub-questions in a very orderly manner, depending upon the type of data distribution and variables, we used a range of non-parametric tests as well as parametric tests.

- Multinomial Logistic Regression: used to predict class membership or probability of class membership for a dependent variable based on multiple explanatory variables (Starkweather and Moske, 2011). This has been used in our study to predict the probability of membership of each category of the independent variable to dependent variables such as task types and multitasking behaviour.

Table 4.2: Mapping of Sub_Research questions to results sections

Sub-Research Question	Measures Used	Corresponding Results Sections
SRQ1.1	Individual accuracy performance, workload, and engagement	4.8.2 - 4.8.3
SRQ1.2	Correlation between personality traits and worker outcomes	4.8.3
SRQ1.3	Impact of multitasking behaviour and task sequencing on engagement, performance, and perceived workload NASA-TLX	4.8.5

- Mann-Whitney U Test: used to determine if there are statistically significant differences between the two groups. The Mann–Whitney U-Test (MacFarland and Yates, 2016), is used with nonparametric data, typically ordinal data, to define the significant difference between the personality traits of the workers and engagement level between each pair of task types and the two groups of multitasking and non-multitasking groups.
- Kruskal-Wallis Test: a nonparametric statistical test that assesses the differences among three or more independently sampled groups on a single, non-normally distributed continuous variable (McKnight and Najab, 2010). This was used to examine the statistically significant difference between the three task types based on engagement and personality traits.
- Chi-square Test: this applied to test the categorical variables to identify the significant associations between the demographic features and the task type (Howell, 2011).

All analyses were implemented using SPSS and Python statistical libraries. Bonferroni corrections were applied to guard against Type I errors in multiple comparisons, ensuring the interpretation of significance levels was both transparent and meaningful. Together these statistical strategies provided a strong and dependable analysis for behavioural differences among task types, and contributed to the study's mixed methods research design

Overall, this mixed-methods evaluation model guaranteed methodological triangulation. Objective measures (accuracy), subjective self-reports (NASA-TLX and UES), dispositional characteristics (personality), and open-ended reflections collectively provided a comprehensive, multi-faceted insight into worker performance and engagement across various task types. This is in line with best practice in HCI and crowdsourcing research, which aims for a combination of depth and breadth in interpreting behavioural results.

4.8 Results

After collecting the results from the Amazon Mechanical Turk platform and to assure data analysis reliability, in the survey task, we only filtered out the workers who failed to finish the entire document. For the classification tasks, we excluded the worker assignments if the worker failed to achieve at least one-quarter of the items correctly, corresponding to an accuracy threshold below 0.3. This quality control criterion was communicated to participants in the consent form (see Appendix A). Based on this filter, 139 workers were approved for text-based classification tasks, while 169 were approved for image-based classification tasks, and all 250 were approved for the survey task. Consequently, our study encompassed 558 valid worker submissions. Tasks were disseminated at approximately three-minute intervals. Each assignment was designed to last approximately ten minutes. Despite the absence of restrictions, we examined submissions and verified that worker executing several tasks had an average interval of roughly three minutes between jobs. The frequent issuance of tasks and their short duration suggest that employees presumably focused on these tasks sequentially, without involvement in extraneous tasks; however, this cannot be entirely assured.

4.8.1 Worker Demographics Features and Experience

In each particular task, we found the contributing workers' demographic features were in line as the result of previous studies (Difallah et al., 2018, Hara et al., 2019, Ipeirotis, 2010b, Ross et al., 2010, Martin et al., 2017). As shown in Figures 4.4 – 4.7, most participants were male of 25-34 years old, and held at least a bachelors' degree, and the demographic profile is consistent with the prior Amazon Mechanical Turk samples that have been reported in the literature of crowdsourcing. Approximately 70% of the workers were employed for wages; that indicates that the crowdsourcing platform is not their primary source of income. The married or in a domestic partnership with white and Asian workers were the highest proportion groups of the workers.

Based on the working environment, the majority (60%) of the workers participated from home, which was also reported by Martin et al. (2017), and secondly from their offices. Moreover, 50% of workers characterised themselves as having an intermediate level of 3-4 years of Amazon Mechanical Turk experience. Interestingly, most of the workers preferred to do the tasks using laptops or desktops rather than tablets or smartphones across all tasks, indicating the workers preferred working on larger interfaces in labelling and survey tasks. These stable demographics further confirm the sampling representativeness of the sample compared to prior research, supporting generalisation to similar microtask populations. Further demographic distribution, including the workers ethnicity, work environment and others are provided in Appendix A.4.

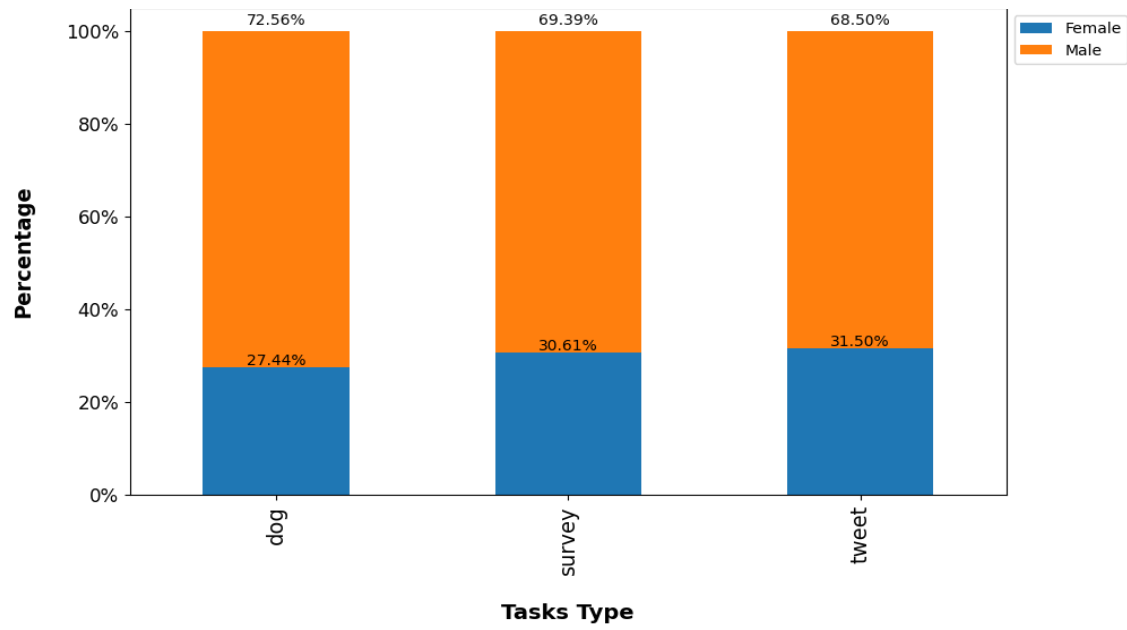


Figure 4.4: Gender distribution in all tasks

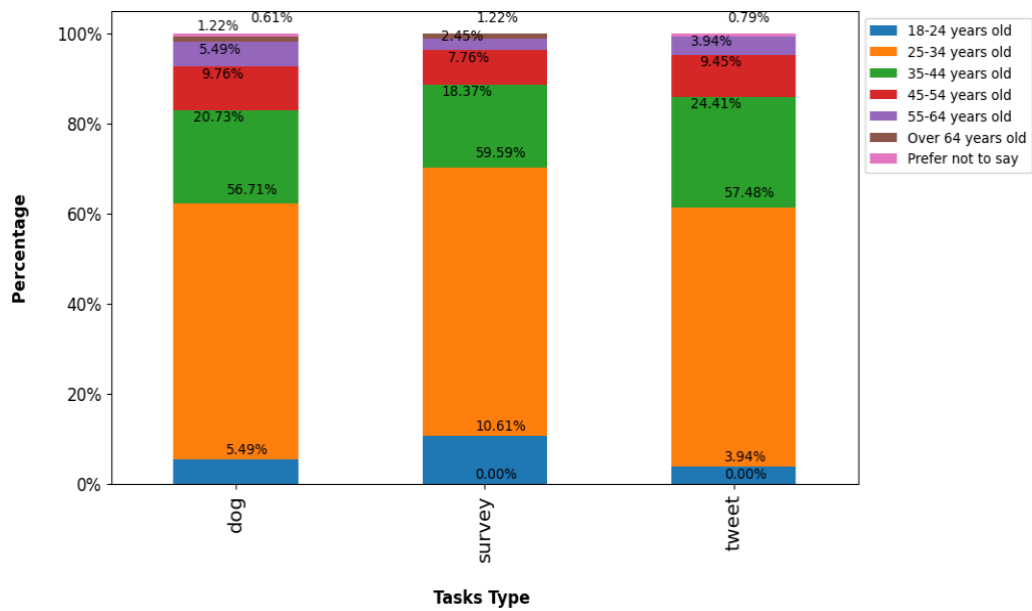


Figure 4.5: Age distribution in all tasks

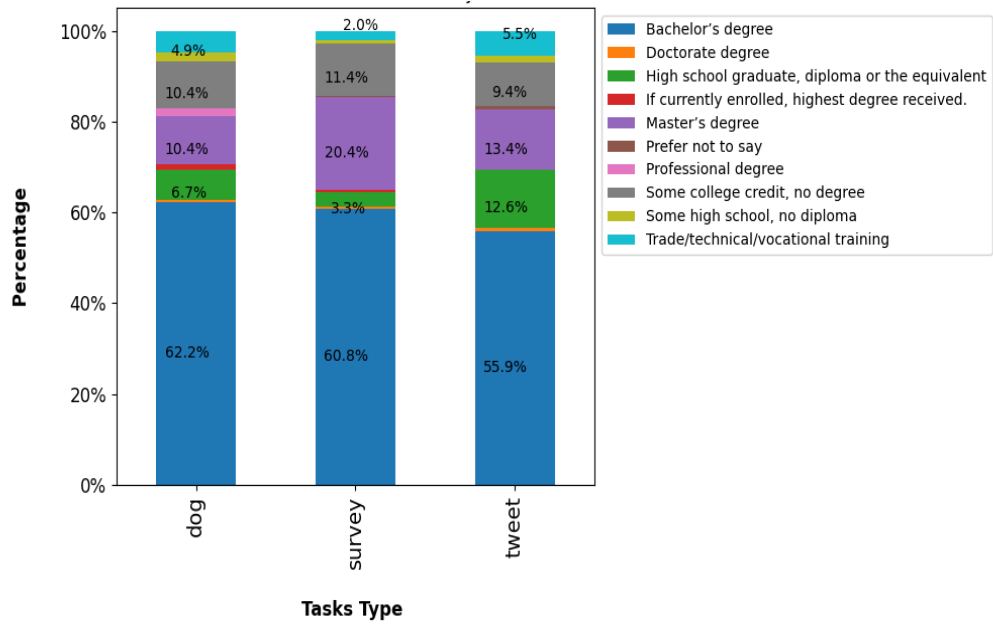


Figure 4.6: Education level distribution in all tasks

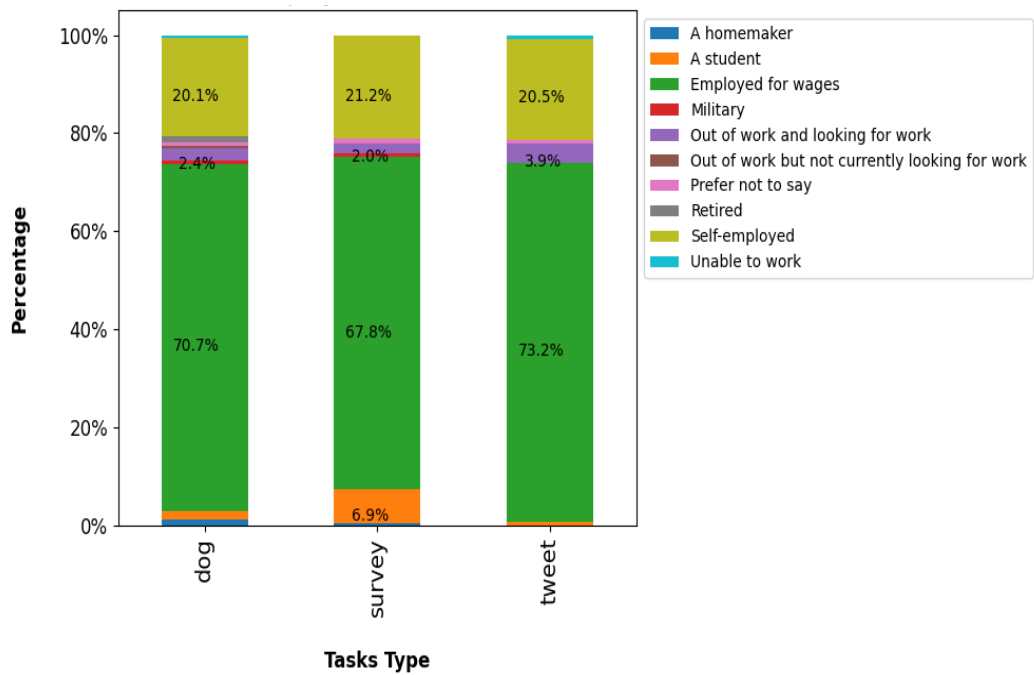


Figure 4.7: Employment status distribution across all tasks

4.8.2 Performance and Workload Result

In relation to SRQ1.1 (Task types of effects on performance and workload), we compared worker performance and perceived workload across different crowdsourcing task types using objective accuracy and subjective workload measures. The results revealed that participant performance varied slightly by task type. Image-labelling tasks gained the highest average accuracy rate, followed by the tweet task. The average accuracy for the dog-classification task ($M \approx 0.65$) exceeded that of the tweet-classification task ($M \approx 0.58$). A significant difference has been identified by Mann–Whitney in the accuracy ($U = 6,772.0$, $p < .001$), with a moderate effect size ($= -.35$) suggesting improved performance on the dog task.

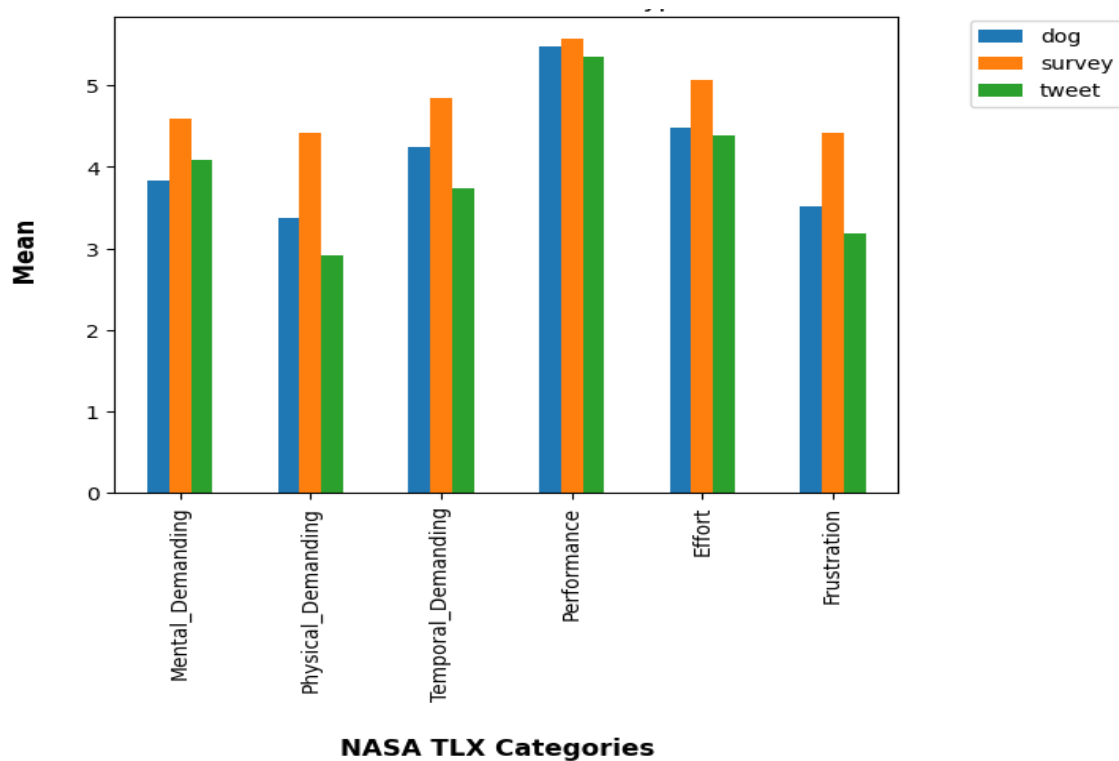


Figure 4.8: Distribution of task load

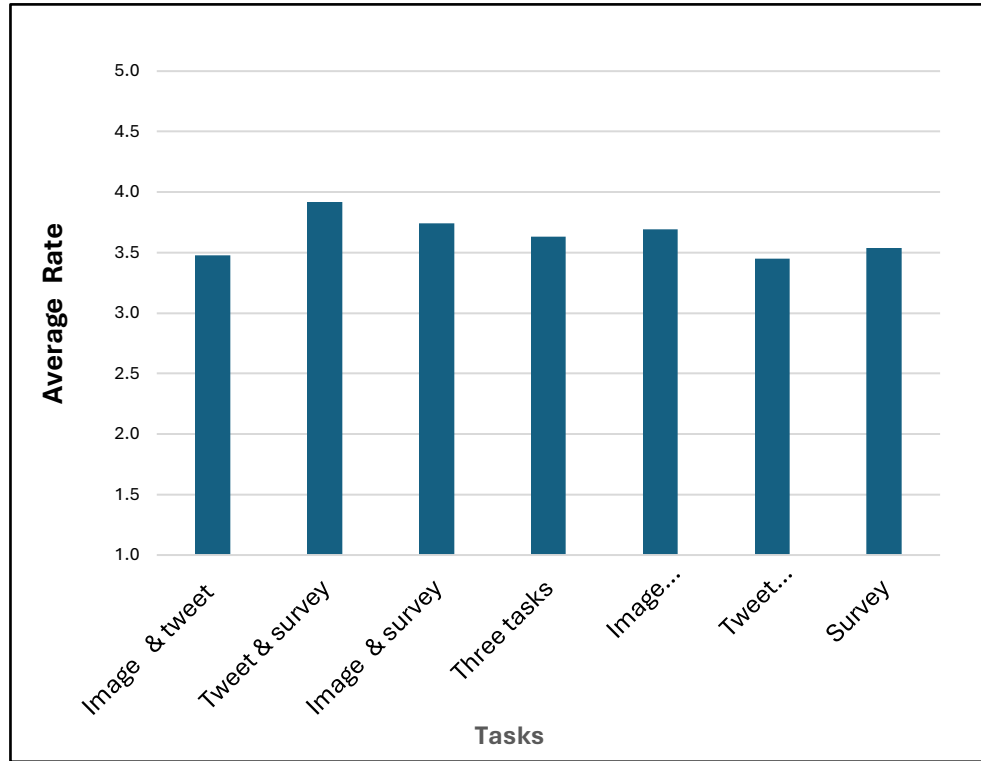


Figure 4.9: Engagement level distribution

Based on the task load, the NASA-TLX score revealed that the survey task was perceived as more demanding than the other two tasks. Survey workers reported higher mean in Mental Demand ($M=4.62$, $SD=1.66$), Physical Demand with ($M=4.46$, $SD=1.97$), and higher Temporal Demand ($M= 4.88$, $SD=1.86$) dimensions as shown in Figure 4.8. These results indicated that the analytical survey task impose a heavier cognitive load compared to classification task. Dog image-labelling jobs, which are less mentally demanding, were more time-consuming because of detailed visual differences. Frequently, workers mentioned that many breeds were “looking alike” and asked for more examples as observed in the theme feedback (Appendix A.5). Thus, the results directly answer SRQ1.1 by revealing that the specific task modality strongly affects accuracy and perceived workload, as image classification yields higher performance and survey tasks impose greater cognitive demand.

4.8.3 Engagement and Personality Traits

This part of the results addresses SRQ1.1 and SRQ1.2 specifically by discussing the differences in engagement for different task types and personality traits and how they relate to engagement and performance. Based on using the User Engagement Scale (UES-SF) (O’Brien et al., 2018), the results showed that the workers who carried out image labelling had a higher level of engagement compared to those doing the two other tasks (Figure 4.9). A Mann–Whitney U test revealed that image classification elicited higher levels of engagement than text classification $U = 11,977.50$, $p = .014$. Survey tasks, on other hand, while it showed higher mental

and temporal demands, had a lower engagement level. According to SRQ1.1, the engagement level varied considerably by task type, where we found image-based tasks to be the most engaging compared with text and survey tasks.

Depending on personality traits, we observed that workers who performed the tweet tasks had a greater level of Agreeableness with ($M=3.44$, $SD=0.88$), which indicates that the worker was generally reliable and helpful. They also had higher levels of Conscientiousness with ($M=3.58$, $SD=0.84$) than workers in the other two tasks, as shown in Figure 4.10, this indicates that the workers did the job thoroughly. Survey workers compared to tweet workers reported a higher level of Extraversion $U = 17,700.00$, $p = .010$, while there were higher levels of Agreeableness for tweet task workers $U = 18,178.50$, $p = .003$ and dog workers $U = 23,802.50$, $p < .001$, comparing to survey tasks. Personality had no significant association to accuracy, but interest was positively associated with Extraversion and Conscientiousness and negatively correlated with Neuroticism.

These results directly address SRQ1.2 by showing that personality traits do not significantly predict objective correctness, but they do correlate with differences in subjective involvement and task interest. This result suggests that individual preferences could have a moderating effect on crowdsourcing participation.

4.8.4 Preferences in Task Selection

An important and key result related to our thesis trajectory was our investigation of workers' self-reported preferences in selecting tasks. Figure 4.11 shows (in the order of importance to workers) which factors influenced whether a worker would take up any given task. Money remained the first and paramount determinant, with “interest in the task” coming close on its heels.

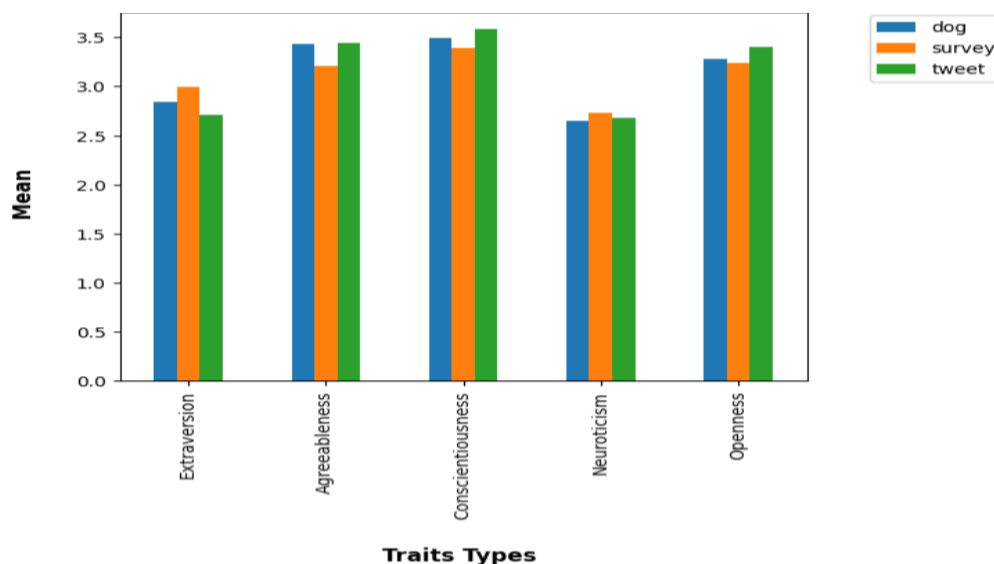


Figure 4.10: Distribution of personality traits

4.8.5 Multiple-Tasking Behaviour

Moreover, to assess the sub-research question, SRQ1.3, we examined the effects of selecting multiple task types on workers' level of engagement, performance, and perceived workload. Group clustering was used in each task type; the first group, called the multiple-tasker group, involved workers who tended to do more than one type of task, and the second group, single-taskers, was for those workers who only participated in one particular task. This clustering was done irrespective of the order in which the tasks are selected; the pairs we found were (image and tweet classification), (tweet classification and survey), (image classification and survey), (all three tasks), and (single task). To guarantee that our investigation possessed sufficient power to identify significant effects, we performed a power analysis based on the sample sizes utilised in these clusters. Power analysis was conducted utilising a medium effect size (Cohen's $d = 0.5$) and a high effect size (Cohen's $d = 0.8$), with a significance threshold (α) set at 0.05. The sample sizes were as follows: 82 participants did the Image and Tweet Classification task, 63 participants did the Tweet Classification and Survey task, 60 participants did the Image Classification and Analysis task, and 33 participants worked on all three task types.

The results indicate that the sample sizes in the first three groups (82, 63, 60) provided sufficient numbers to identify medium effect sizes; however, though the group of 33 individuals was small, it was still capable of detecting a large effect.

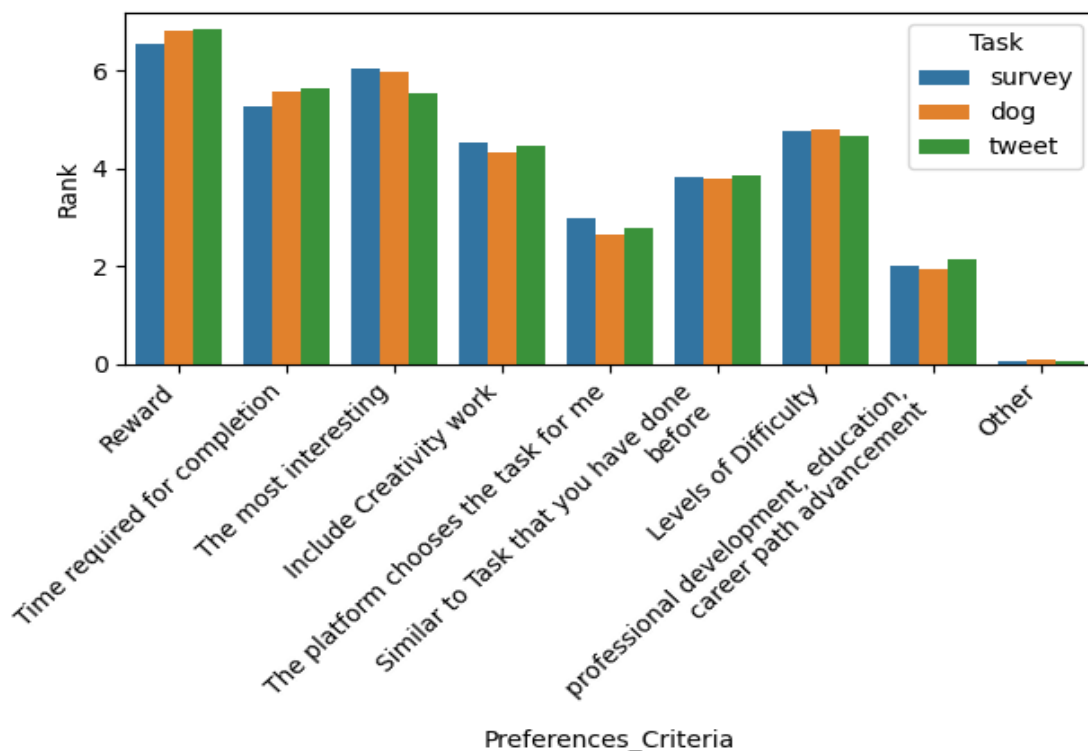


Figure 4.11: Rank of workers' task selection criteria

Table 4.3 illustrates that workers who completed all three tasks perceived them as more physically and temporally demanding compared to those who engaged in any pair of tasks ($U = 330.50, p < .01$) and ($U = 347.10, p < .01$), respectively. Furthermore, the workers who completed all three tasks expressed greater annoyance than those who performed only image and tweet classification tasks ($U = 502.50, p = .001$), and more than those who did image classification and survey ($U = 260.50, p < .001$). Although the frustration for the three-task workers was also higher than the workers who carried out the tweet and survey tasks, the difference was not statistically significant.

The workers' engagement level in doing only the image classification task was higher than the other two individual tasks as shown in Figure 4.9. The workers' feedback explains this, as they reported that they found the images interesting. Moreover, the results show that there were statistically significant differences ($p = .008$) between the groups in terms of engagement and perceived workload. The workers who did (tweet & survey) tasks were more engaged than those who did the (tweet & image) tasks more than those who did all three tasks with ($U = 330.00, p = .001$) and ($U = 210.50, p = .004$), respectively.

These results answer SRQ1.3 by showing that workers engaging in multiple-tasking behaviour reported higher workload and frustration and reduced engagement in many cases, even when the difference is small in accuracy. Based on the accuracy of the workers in the two classification tasks, see Figure 4.12 and Figure 4.13, around (20%) of workers who did only one classification task (in red colour) had a higher accuracy rate than those who participated in more than one task. We also analysed the relationship between crowdsourcing workers' personality traits and their task workload, performance, and engagement levels when doing multiple tasks.

The results showed, as in Table 4.4, that in the tweet classification and survey tasks pair, the workers had higher levels of Openness with an average rate (3.85), Agreeableness (3.62), and Conscientiousness (4.23) compared to those who did all three tasks or the other pair of tasks. Workers who did all three tasks exhibited

Table 4.3: Workers' workload in different tasks

	Sample Size	Mental Demand	Physical Demand	Effort	Perceived Performance	Temporal Demand	Frustration
Image & tweet	82	4.03	3.66	4.47	4.94	4.19	3.50
Tweet & survey	63	3.85	1.77	4.23	6.00	3.00	1.85
Image & survey	60	3.59	3.59	4.12	5.24	4.06	3.47
Three tasks	33	4.95	5.19	5.48	5.43	5.85	5.24

elevated Neuroticism ratings, namely 3.0. In the task of tweet classification, there was a negative relationship between higher levels of Extraversion and accuracy, with a coefficient = -0.29 and a $p = .130$. On the other hand, Agreeableness and Conscientiousness had a slight positive effect on accuracy, with coefficients = 0.16 and 0.17, respectively, and $p = .480$ and $p = .440$. However, these correlations were not statistically significant. Neuroticism had a coefficient = -0.06 with a $p = .709$, indicating a negative relationship. Moreover, Openness had a coefficient = 0.048 with a $p = .853$, suggesting a positive relationship. However, neither of these relationships was statistically significant. In the task of classifying images, a higher level of Conscientiousness was found to possibly lead to an increase in accuracy (coefficient = 0.3158, $p = .160$). However, higher Extraversion levels (coefficient = -0.2603, $p = .199$), Agreeableness (coefficient = 0.0586, $p = .770$), Neuroticism (coefficient = -0.0662, $p = .655$), and Openness (coefficient = 0.1758, $p = .437$) did not show statistically significant correlations with accuracy.

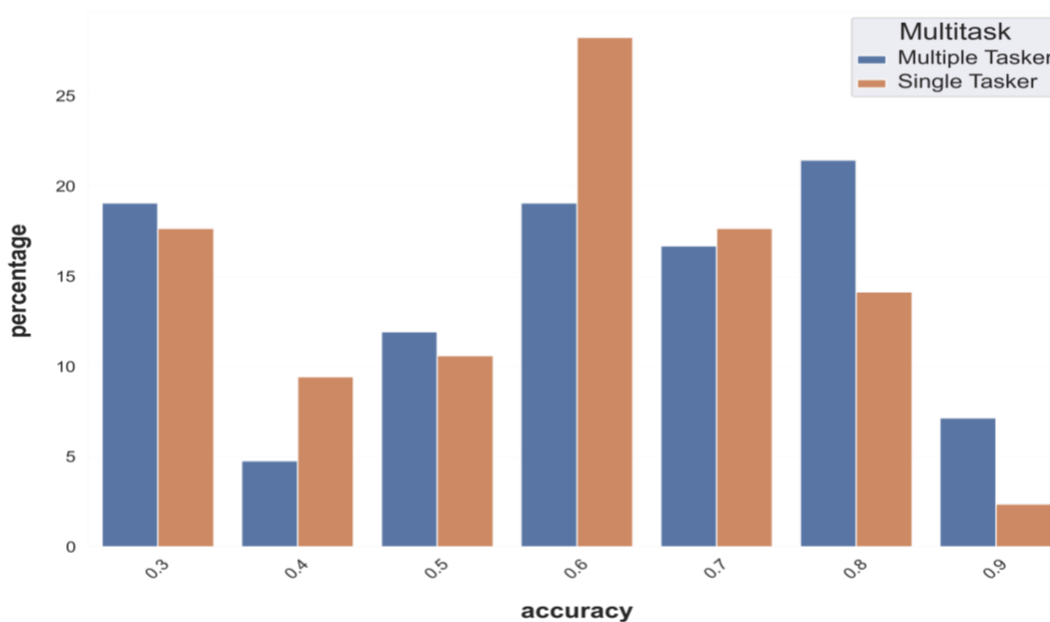


Figure 4.12: The percentage of workers with a certain accuracy rate in the tweet classification task

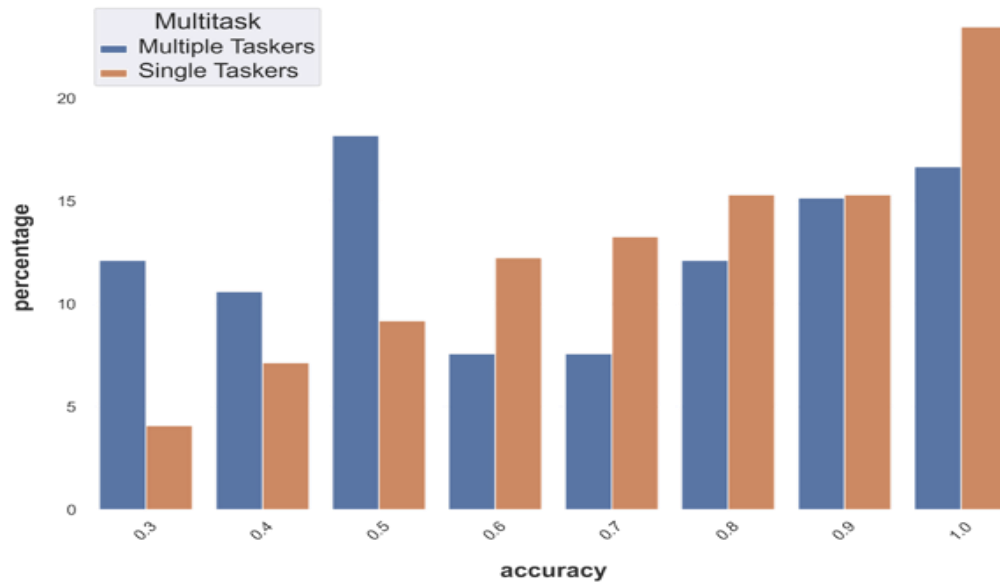


Figure 4.13: The percentage of workers with a certain accuracy rate in the dog classification task

In general, even though personality traits appeared to have an effect on accuracy; those outcomes did not attain statistical significance, which suggests further analyses using larger sample sizes or other analysis methods. For further information and based on the workers' personality traits, we analyse workers' demographic features for those who did multiple tasks and their correlation to their personality. The result showed that the male workers outnumbered female ones, with around (72%) male participants. The distribution of ages is similar in all tasks, with the majority in the age group 25 to 34 (above 50%) and around (50%) with a university degree or higher. Around (70%) of the workers were employed for wages, so the crowdsourcing platform was not their only source of income.

In line with previous studies (e.g., (Martin et al., 2017)), the majority of workers (60%) participated from home and (50%) characterised themselves as having an intermediate level of 3-4 years of experience in crowdsourcing work.

Table 4.4: Workers' personality traits in different tasks

	Openness	Extraversion	Agreeableness	Conscientiousness	Neuroticism
Image & tweet	2.96	2.81	3.33	3.36	2.67
Tweet & survey	3.85	2.58	3.62	4.23	2.12
Image & survey	3.50	3.21	3.62	3.50	2.47
Three tasks	3.09	2.88	3.12	3.21	3.0

These findings addressed SRQ1.3 by demonstrating that multiple-tasking behaviour correlated with the highest perceived workload, reduced interest, and dissatisfaction despite a minor variation only in accuracy. This confirmed that behavioural features, such as task switching, significantly influence worker experiences on crowdsourcing platforms.

Furthermore, in the second approach of analysing multiple tasking behaviours, we consider the ordering of the tasks' selection for the workers who did two and three tasks. We found that there were six groups of ordering selections of the tasks for the workers who did two tasks, which were

1. Image classification and then a survey task.
2. Image followed by tweet classification.
3. Survey, then image classification task.
4. Survey then tweets classification task.
5. Tweet classification then survey task.
6. Tweet then image classification task.

While the groups who did all three tasks were:

1. Survey – image classification – tweet classification.
2. Survey – tweet classification – image classification.
3. Tweets classification – image classification – Survey.

There weren't any statistically significant differences in engagement levels, accuracy rate, or perceived workload for any of the task orders. The lack of statistically significant order effects also supports SRQ1.3, as the multitasking effects are not explained by task sequence, learning transfer, or fatigue. These results suggest that multitasking behaviour is associated with increased strain and reduced the task focus, indicating the need to create a clearer design for the task structure and support for keeping attention.

4.9 Discussion

This section discusses the results presented earlier relative to RQ1 alongside the theoretical lenses introduced in Chapter 2. It synthesises the results across the sub research questions SRQ1.1-SRQ1.3 to explain how the task types, individual variations, and multiple-tasking behaviour shape worker performance and engagement. Moreover, it reflects on the implications of the multiple tasking and the task-order effects for the validity of the study. First, the results showed that the task type meaningfully contributes to both performance and perceived workload. Image classification yielded higher accuracy compared to text-based classification, and survey tasks required the highest mental, physical, and temporal demand. From an SDT perspective and flow approach, this would seem to reflect the fact that the visual image task may have been more explicit with clearer, immediate cues for competence, and the survey task required more workload effort with lower levels of intrinsic enjoyment. Second, engagement outcomes reflected the same pattern, workers indicated a higher level of engagement for the image-labelling task and a lower level of engagement for the survey task, despite

having a heavier workload in the latter. This suggests that high workload cannot automatically ensure engagement, but that engagement occurs when challenge, clarity and intrinsic interest are adequate, as highlighted in the Flow Theory. Third, personality traits accounted for some of the variability of experiential outcomes. Workers who scored high in Agreeableness and Conscientiousness selected and completed tweet tasks, while respondents who scored high in Extraversion responded more on the survey. While the relationship between personality and accuracy was not statistically significant, interest positively correlated with Extraversion and Conscientiousness, and negatively with Neuroticism. This is consistent with the findings of previous research such as (Kazai et al., 2011) and indicates that personality moderates how readily motivated workers can move into and maintain motivation in various task states. Fourth, the analysis of task-selection preferences showed that “interest in the task” was consistently ranked second to monetary reward as an important determinant of participation. This finding is significant to the trajectory of the thesis because it indicates that crowd workers are not motivated exclusively by financial compensation; the experiential and motivational qualities of the task itself also influence participation and engagement. From the perspective of Self-Determination Theory, this observation is consistent with the importance of intrinsic motivation, perceived interest, and task value. Importantly, however, this study provided empirical evidence that worker motivation extends beyond payment and therefore strengthened the rationale for examining motivational task design more systematically. This empirical insight informed the next phase of the research in Chapter 5, where a scoping review was conducted to investigate how gamification has been operationalised across crowdsourcing contexts, which motivational mechanisms and theoretical foundations have been used, what outcomes have been reported, and where important design gaps remain. Finally, the multitasking and order-effects analysis have two key implications. On the one hand, workers that performed more than one type of task reported higher workload, frustration, and, in many instances, less engagement, which corroborates concerns in cognitive load theory that increased task switching often reduces focused activities, thereby negatively affecting performance (Sweller, 1988). On the other hand, this is a good methodological finding as there is no statistically significant difference between task-order groups in engagement, accuracy, and workload. Multi-condition experiments raise an area of concern regarding asymmetrical learning or asymmetrical fatigue: if one takes Task A and does Task B, Task A may increase task B performance or decrease it, which is incompatible with the assumption they are independent of one another. By specifically testing for order effects, and observing no effect, the internal validity of the task type, multiple-tasker comparison with multiple-tasking group is improved. Statistically, the results collectively provide clear answers to the three sub-research questions that direct this chapter. SRQ1.1 showed that task modality strongly affects the workers' performance, workload, and engagement, with image-based tasks indicating higher engagement than text or survey tasks. SRQ1.2 revealed that some personality may shapes how the workers perform and engage with tasks, despite the accuracy remaining unchanged. Finally, SRQ1.3 demonstrated that

multiple-tasking behaviour introduces additional cognitive-load and emotional expenses, underscoring the significance of task concentration and behavioural design factors in crowdsourcing systems. Statistically, the results collectively provide clear answers to these three sub-research questions that direct this chapter. To sum-up, the findings of RQ1 are summarised by the following statement: workers' engagement and performance vary according to task type, multiple-tasking behaviour, and that the differences are closely linked to the perceived workload and task interest. Taken together, these insights strengthen the thesis focus on motivational task design and provide the empirical basis for the scoping review presented in Chapter 5, where gamification is examined systematically as one broad design space for supporting worker motivation and engagement.

4.10 Conclusion

This exploratory study has offered fundamental insights into worker engagement, performance, workload, and personality attributes across three distinct non-gamified task categories: text classification, image classification, and survey assignments. Moreover, the study has emphasised the impact of multiple-tasking behaviour and personality factors on worker performance and engagement. The results provide an essential understanding of how task characteristics shape the objective and the subjective aspect of crowd work. The findings also, confirm that worker demographics in this sample are broadly representative of the Amazon Mechanical Turk population reported in previous literature, this supports the reliability and generalisability of these findings.

Performance exhibited variability across task types, with accuracy and engagement being superior for image-based labelling compared to text-based classification. whereas survey activities were linked to increased reported effort. Taken together, these patterns showed that task modality and the cognitive complexity directly affect the level of worker performance, satisfaction, and workload.

Personality traits helped to further illuminate the understanding of the individual differences: higher levels of engagement were reported by workers who scored high in Extraversion and Conscientiousness. Conversely, engagement and accuracy often run in the opposite direction when workers possess a great deal of Neuroticism. These relationships underline that intrinsic psychological variables interact with design of the task and should be taken into consideration in future motivation-focused research. A significant contribution to this study relies in its analysed multitasking behaviour. Workers who undertook more than one type of task experienced a heavier perceived workload, frustration, and lower engagement. This viewpoint underlines that the switching between the tasks undermines the engagement and the performance. This in turn, emphasises the importance of designing the tasks that foster focus and intrinsically motivate workers rather than create an excessive cognitive load. Most importantly, a look at employees' self-reported preferences for tasks shows that next to money, "interest in the task itself" ranks highest among major reasons people choose one task over another. This validated the central role of intrinsic interest in workers' decisions and provided an empirical link to the motivational focus of the rest of this thesis. In relation to RQ1: How do workers' performance and

engagement vary across different crowdsourcing task types and its associated sub-research questions of chapter 4, SRQ1.1 (how workers' engagement, performance, and perceived workload differ across task types), SRQ1.2 (whether personality traits correlate with accuracy and engagement), and SRQ1.3 (how multitasking behaviour affects engagement, performance, and perceived workload), the chapter concludes that:

- Workers' engagement and performance differ systematically across different types of tasks, and image tasks are associated with greater precision and engagement than text and survey task types.
- A survey task, when combined with other tasks, results in a more considerable workload and frustration, underscoring the challenges of increased cognitive load.
- Personality and multitasking affect how workers experience such tasks, as it may not always equate to meaningful accuracy differences.
- Beyond being paid, task interest also plays a significant role in determining whether workers decide to select a task and participate.

Collectively, the statistical analyses described in this chapter provide clear answers to SRQ1.1, SRQ1.2, and SRQ1.3, ensuring that each analysis is theoretically motivated, empirically grounded, and directly tied to the study's research questions. SRQ1.1 was answered by showing that an individual task led to various performance, engagement, and workload patterns; also, SRQ1.2 was answered by demonstrating that personality traits are associated with engagement and interest, but not with accuracy; SRQ1.3 has been answered by showing that multiple-tasking behaviour increases workload and frustration while undermining engagement. Collectively, this evidence suggests that crowd work results are determined by task design but are also influenced by task interest, quality design and cognitive loading. In summary, this establishes the baseline for behavioural and psychological investigations that are needed for the following subsequent investigation. It has shown that worker performance and motivation are not simply based on pay or task difficulty. Thus, these lessons lead us into the next chapter, Chapter 5, which builds on the findings of this exploratory study directly by systematically reviewing the literature on motivational and gamified aspects in crowdsourcing systems. This chapter explores how design elements such as feedback, progression, social interaction and rewards have been utilised to bolster intrinsic motivation and engagement and which of these mechanisms are most promising to be able to address the engagement and workload challenges.

Chapter 5 :Scoping Review of Gamified Tasks in Crowdsourcing

5.1 Introduction

Building on the conceptual foundations developed in Chapter 2 and the exploratory findings reported in Chapter 4, this chapter narrows the investigation from motivation in crowdsourcing more broadly to gamification as a specific motivational design space. Chapter 2 established that worker motivation is multidimensional and may be influenced through extrinsic, intrinsic, social, and design-centred mechanisms. Study 1 provided empirical evidence that workers' experiences and preferences vary across task types, with task interest emerging as one of the most important considerations in task selection after monetary reward. Together, these findings reinforced the need to examine how motivational qualities can be incorporated more systematically into crowdsourcing task design.

Gamification was selected as the focus of the scoping review because it offers a comparatively broad design space through which multiple motivational mechanisms can be operationalised, including feedback, progression, achievement, goals, recognition, competition, cooperation, and social interaction (Morschheuser et al., 2017, Riar et al., 2022). The focus on gamification does not imply that other motivational design approaches, such as task structuring, onboarding, priming, or interface design, are unimportant. Rather, gamification provides a useful higher-level framework for examining how several motivational affordances can be configured individually or in combination within crowdsourcing environments.

Accordingly, this chapter conducts a scoping review to systematically map how gamification has been applied across the principal crowdsourcing archetypes, which game elements and motivational mechanisms have been emphasised, which theoretical frameworks have been used, and what behavioural and performance outcomes have been reported. The review also identifies areas that remain comparatively underexplored, particularly collaborative and socially oriented forms of gamification. These findings subsequently provide the evidence base for selecting and integrating the motivational and collaborative features implemented in the experimental system presented in Chapter 6.

5.2 Scoping Review Objective and Research Questions

The main objective of this scoping review is to systematically examine gamification as a motivational design space in crowdsourcing by mapping how its elements, underlying motivational mechanisms, theoretical foundations, and reported outcomes vary across the principal crowdsourcing archetypes. This review makes a significant contribution to the field of crowdsourcing by clarifying and explaining how the mechanisms of gamification can effectively be implemented to improve participation, performance and motivation in crowdsourcing environments. The main research question guiding this review is

RQ2: How have the additional task features (e.g., gamification elements) been applied in existing crowdsourcing research across the main archetypes?

Addressing this research question allows us to understand how gamification research within the field of crowdsourcing has been developed, what forms of knowledge have been gained, and where the significant gaps are that remain unexplored. Taking an overarching look at the literature can offer valuable insights that will guide the subsequent empirical research in this thesis. The synthesis presented here builds the necessary theoretical foundation for the experimental strategy used in later chapters in this thesis. In particular, the outcomes of this review provided details on what gamification tools, mechanisms, and evaluation criteria would be chosen to later use in empirical research. This ensures that the thesis naturally builds directly upon, and contributes significantly, to establish the body of research in gamified human computation and crowdsourcing.

5.3 Scoping Review Methodological Approach

5.3.1 Methods

The scoping review process follows the methodological framework which was initially proposed by Arksey and O'Malley (2005) and subsequently improved by Levac et al (2010). This framework was widely selected for its flexibility and systematic structure, which is suitable for mapping a broad and heterogeneous research field such as gamification and crowdsourcing. The review was conducted following five stages which were:

1. Identifying the research topic.
2. Locating pertinent studies,
3. Choosing the studies based on inclusion and exclusion criteria.
4. Mapping and charting the data.
5. Synthesising themes and interpreting findings.

5.3.2 Databases Searched.

Based on the systematic approach of Boell et al. (2015), the review started with an extensive search for scholarly research publication in Scopus, selected as the primary database owing to its broad coverage of multidisciplinary academic literature, including AIS, ACM, IEEE and ScienceDirect repositories, among many others (B.V, 2022). Scopus also was selected to ensure the reliability and the wide range of the sources, given its inclusion of the highly cited papers in Computer Science, Human Computer Interaction, and Information Systems. To guarantee a wide and comprehensive coverage of the scholarly literature, a complementary search was conducted in Google Scholar, which is known to produce a greater number of citations across all disciplines (Martín-Martín et al., 2018).

The period covered in this search was 2011-2024, which captures the period that the utilisation of the term crowdsourcing and gamification gained an increasing significance in academic research. This also reflects the emergence of crowdsourcing as a socio-technical phenomenon, following the productive creators of the topic who published between 2006 and 2011(Howe, 2006b).

5.3.3 Search Terms

The main search query that has been used in Scopus was:

TITLE-ABS-KEY (Gamification and Crowdsourcing) AND (EXCLUDE (DOCTYPE,"cr") OR EXCLUDE (DOCTYPE, "bk") OR EXCLUDE (DOCTYPE, "no") OR EXCLUDE (DOCTYPE, "tb")) AND (EXCLUDE (LANGUAGE," Portuguese") OR EXCLUDE (LANGUAGE, "Russian") OR EXCLUDE (LANGUAGE, "Slovenian") OR EXCLUDE (LANGUAGE, "Chinese") OR EXCLUDE (LANGUAGE, "Croatian") OR EXCLUDE (LANGUAGE, "French")).

The search aimed to identify the papers which containing both concepts that appeared in the title, abstract, or keywords, thus guaranteeing conceptual relevance. Exclusions were implemented for non-English and non-peer-reviewed papers to ensure consistency and quality in the collected dataset.

5.3.4 Selection Criteria

The Scopus search query resulted 378 papers, which were imported in the scoping database for the screening step. Additionally, the first 100 Google Scholar results were manually screened, followed by saturation examination of the next ten results; confirmed that the inclusion of more records produced diminishing returns, aligning with the principles of saturation theory (Saunders et al., 2018). The literature searches were first conducted in September 2022 and subsequently updated in December 2024 to incorporate the latest developments in the field. Purposefully, the searches were restricted to the metadata rather than looking for the phrases across the whole text to reduce the number of false positives resulting from identical mentions of gamification or crowdsourcing. Based on the cross-disciplinary nature of crowdsourcing, we did not limit the search to certain publications or academic domain. In total, 478 unique papers were screened according to the following inclusion criteria:

1. The full text of the paper was available and accessible.
2. The paper was written in English and had been published in a peer-review international outlet.
3. The paper focused on gamification tasks in crowdsourcing instead of just terms being mentioned in the metadata of the papers.
4. The paper presented unique empirical findings (duplicate or re-reports of the same study were excluded).
5. The study paper contained a measurable construct of gamification effectiveness on participants. (e.g., performance, engagement, or motivation).

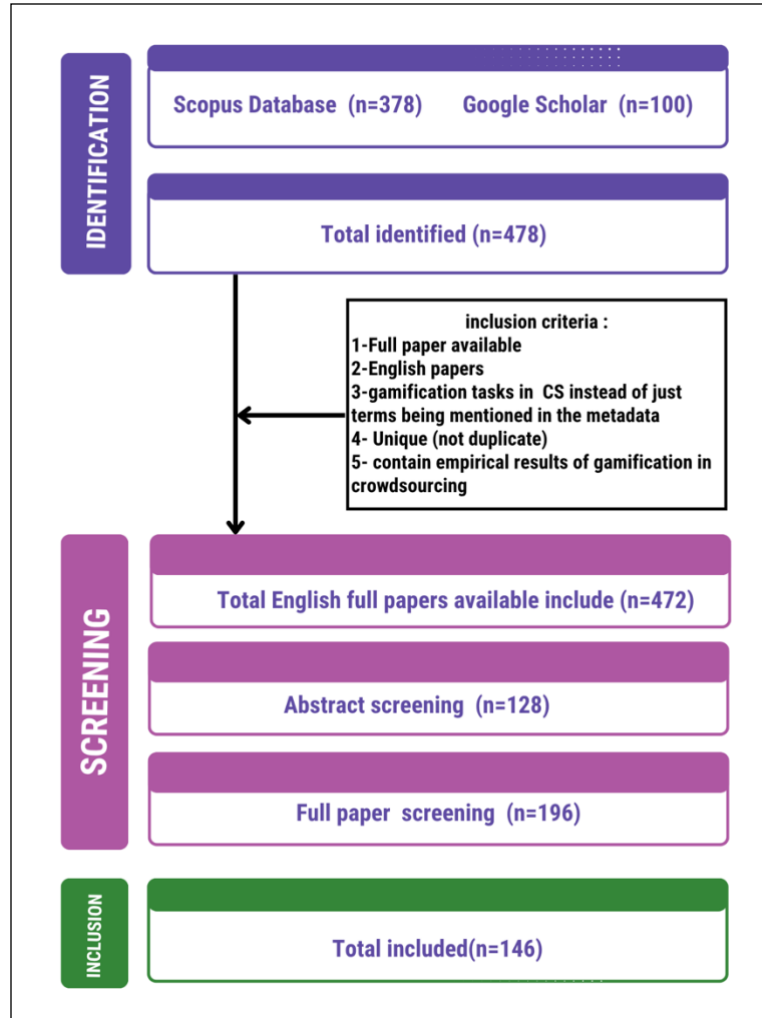


Figure 5.1: Scoping review PRISMA flow-chart process

As shown in Figure 5.1(PRISMA) flow chart, the screening process led to the exclusion of six non-English papers, two papers without a full text, 128 papers were also left out of the review since the substance of those studies did not truly involve gamified crowdsourcing. Furthermore, we found 196 papers that did not meet the inclusion criteria after reading the full text. Ultimately, 146 papers were chosen to be included in the final literature evaluation.

5.4 Scoping Review Results

5.4.1 Bibliometric Information

The broader set of records retrieved during the literature search showed that research on gamification and crowdsourcing spans a wide range of disciplinary fields, reflecting the interdisciplinary nature of the topic. Among the retrieved search results, 286 records were associated with computer science and 69 with social sciences field. The remaining published articles were distributed across other domain fields, including Engineering, Mathematics, Business, Management and Accounting, Decision Sciences, Arts and Humanities,

Medicine, Earth and Planetary Science, Psychology, Materials Science, Physics and Astronomy, Economics, Econometrics and Finance, Environment, Energy, Multidiscipline, Agricultural and Biological Sciences, Biochemistry, Genetics and Molecular Biology, Health Professions, Immunology and Microbiology, Neuroscience and Nursing. This distribution also reveals the dissemination of gamification concepts into non-technical areas and confirms that academia increasingly recognises its potential for motivating and enhancing performance and engagement across a variety of applications.

5.4.2 Papers Classifications

The included reviewed papers were classified based on a crowdsourcing archetype framework developed by Geiger et al. (2012), that were discussed in Chapter 2 and applied in this thesis for its robust conceptual distinction between the nature of crowd contributions and how the task value solutions are derived from crowdsources. This archetype framing is extensively employed to distinguish between systems where contributions are homogeneous versus heterogeneous, systems where value is aggregated or emergent and is appropriate in the context of the study of design patterns for gamified crowdsourcing systems (Geiger et al., 2012, Geiger and Schader, 2014). To ensure the validity of the author's classification of the resulted paper according to crowdsourcing framework of Geiger et.al (2012), an inter-rater agreement analysis was conducted between the author and an external human expert (Entisar Alkayal) ¹. A random subset of 30 papers, representing approximately 20.5% of the total included set of studies, was selected for validation, consistent with established guidelines for reliability assessment. Each paper in the subset dataset was independently categorised by the expert according to one of the four archetypes of the Geiger crowdsourcing framework, crowd-processing, crowd-rating, crowd-creating, or crowd-solving, with justification of the selection. In many research studies where inter-rater reliability or expert validation is involved, selecting a subset of around 20-30% of the total dataset is a common approach. A Cohen's Kappa score of 0.46 was obtained, indicating moderate agreement between the author and the expert classification. The confusion matrix analysis (Figure 5.2) shows that the highest agreement has been occurred in the category of crowd-creating, while the crowd-rating and the crowd-processing have the greatest overlap because both rely on aggregated, homogenous contributions, where the value is directly derived from the individual inputs. The results validate the dependability and clarity of the implemented categorisation process.

According to this validated methodology, most gamified crowdsourcing research was classified as crowd-creating, followed by crowd-rating and crowd-processing. Crowd-solving research was minimally represented (see Table 5.1 for comprehensive classification). This distribution matters for the thesis argument

¹ **Entisar Alkayal**

Assistant Professor in Computer Sciences Department, Kau
Cloud Computing IOT Optimization Distributed Systems Edge Computing

with respect to the extent to which most gamified systems work with either emergent collective artefacts (crowd-creating) or aggregated judgements (crowd-rating), the very contexts in which coordination support and shared performance awareness have the potential to plausibly impact on experience and also output quality.

5.4.3 Literatures Characteristics

The core component of any crowdsourcing system depends on the nature of the task delegated to the crowd community. Among the reviewed studies papers, a wide range of activities was identified, reflecting the flexibility of the crowdsourcing framework and the width of their application domains. The majority of the reviewed gamified crowdsourcing study systems were designed to motivate participants to perform computational work that would otherwise be difficult or inefficient for computers to complete autonomously without human assistance (Von Ahn, 2008). Examples include the annotation of medical texts, video transcription, or the recognition of visual objects such as animals, plant species, or environmental waste or rubbish. Other studies focused on the proofreading of text scanned with OCR technology, geolocation-based reporting information, and user-generated feedback collection, all aimed to harness the human judgement for validation of the data or enrichment.

Moreover, to such tasks, other research studies have examined creative production and problem-oriented crowdsourcing activities, for instance, ideation challenges, content creation and design optimisation problems tasks that leverage the collective intellect and creativity of the crowd workers. Also, a number of implementations targeted location-based mobile crowdsourcing application, where crowd participants contributed information via sensor devices or platforms. This variety indicates that gamification is not confined to a single type of crowdsourcing task, but is used to enhance the output quality, motivation and engagement across a continuum from simple labelling to creative productions. Specific corpus creations have been found as the objective of other research studies. Furthermore, Table 5.1 summarises the distribution of studies across the four crowdsourcing archetypes:

- Crowd-rating tasks were the most common gamified systems, where the participants voted, ranked or predicted with no substantial manual effort, but advantages could be drawn from aggregation across large numbers of users.
- Crowd-processing research focused on large volumes of homogeneous tasks, emphasising efficient processing and throughput rather than creativity.
- Crowd-creating studies involved participants in open-ended and tasks with the output often being an emergence of collective artefacts.

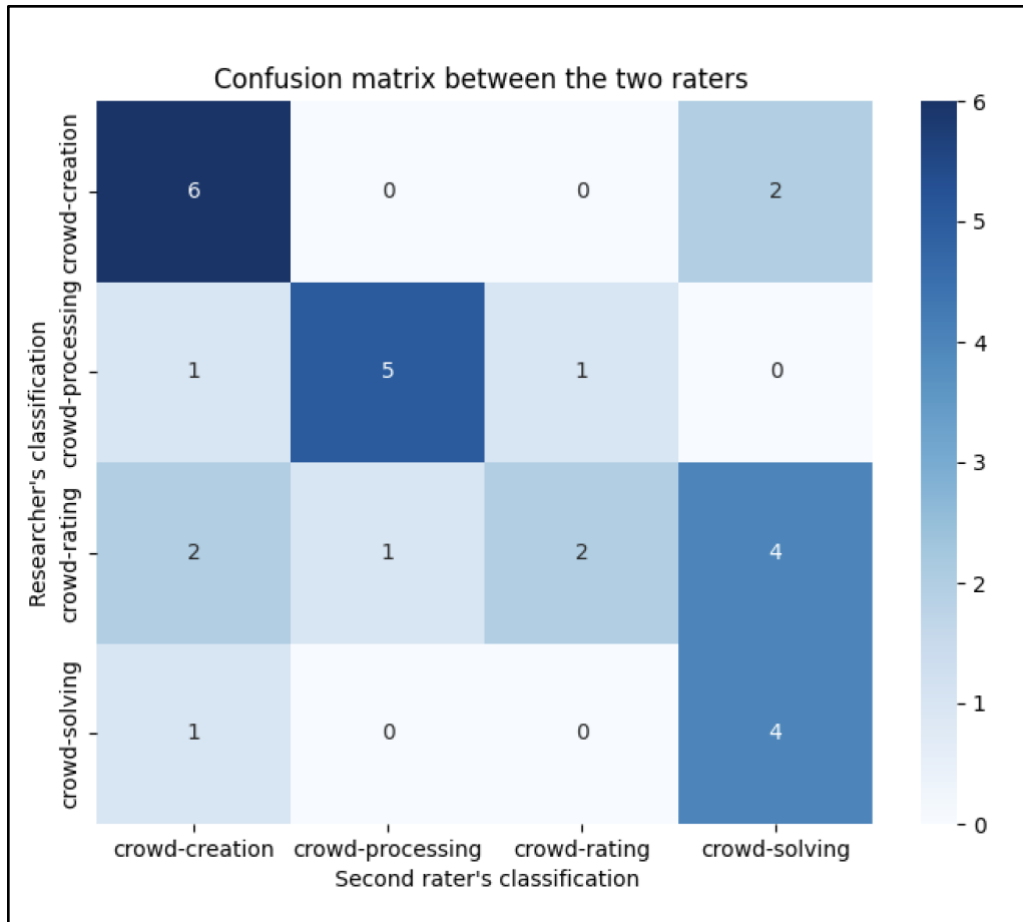


Figure 5.2: Confusion matrix between raters

- Crowd-solving studies were rare and also usually focused on difficult, intellectually challenging problems which needed individual input and a greater cognitive load

5.4.4 Types of Crowdsources

Accordingly, studies reviewed with regard to crowd-targeted populations were similarly varied (Table 5.2.). Most systems were designed for open (or public) participation, where there was no need to possess domain-specific expertise. This is consistent with the general ethos of crowdsourcing in leveraging large, diverse participant pools. A significant proportion of the studies specifically targeted students as participants, particularly in projects with an experimental or educational focus. Experimenters often depend on students because they can provide a convenient sample for testing the usability of game-like task interfaces and collaborative design environments; this trend may limit generalisability of the results.

Table 5.1: Classification of reviewed studies

Crowd Type	Reasons	Papers references
Crowd-rating	In all these studies the individuals perform an integrative aggregation, a vote or specific predication for a given topic. Generally nominal manual exertion required. No individual differences required for answers.	(Macakoğlu et al., 2022, Hong et al., 2013, Morschheuser et al., 2019, Hajarian and Hemmati, 2021, Harris and Srinivasan, 2019, Aljanaki et al., 2016, Gutiérrez Páez et al., 2021, Linares et al., 2019, Brovelli et al., 2018, Guy et al., 2011, Good et al., 2014, Sadeghianasl et al., 2021, Rinner et al., 2020, Bujari et al., 2016, Olinsky et al., 2021, Sadeghianasl et al., 2020, Fort et al., 2020, Hallifax et al., 2019, Kirkham et al., 2015, Chen et al., 2018, Harris, 2017, Lumsden et al., 2016, Cáceres et al., 2019, Pinto and Viana, 2019, Eryiğit et al., 2021, Bin Hussein et al., 2019, Cordova Mancheno, 2017, Kratchanov and Luchev, 2024, Yang et al., 2024, Bastanfard et al., 2024, Amorim et al., 2024, Duggan et al., 2024, Lin and Ding, 2023, Kalokyri et al., 2023, Leal, 2023, Duhaima et al., 2023, Stamatiadou et al., 2023, Harris, 2014, Mason et al., 2012, Massung et al., 2013, Chamberlain, 2014, Dumitrache et al., 2013, Eickhoff et al., 2012, Goncalves et al., 2014, Preist et al., 2014, Saito et al., 2014)
Crowd-processing	In these papers, the non-emergent homogenous tasks are easily enumerable. Generally nominal manual exertion required. No individual differences required for answers.	(Malatini et al., 2017, Purnomo W.P et al., 2019, Carlier et al., 2015, Ipeirotis and Gabrilovich, 2014, Feyisetan et al., 2015, Prestopnik et al., 2017, Spatharioti, 2020, Rokicki et al., 2014, Walter et al., 2022, Yacobson et al., 2021, Lichtenberg et al., 2020, Zhou et al., 2017, Lendak, 2019, Tinati et al., 2017, Packham and Suleman, 2015, Bin Hussein et al., 2019, Cordova Mancheno, 2017, Sadeghianasl et al., 2024, Spicher et al., 2024, Hajarian et al., 2024, Paulino et al., 2024,

		Maddalena et al., 2023, Rönneberg and Kettunen, 2023, Na and Han, 2023, Jacob et al., 2022, Nose and Hishiyama, 2013, Itoko et al., 2014, Brenner et al., 2014, Uzun et al., 2013, Kawajiri et al., 2014, Lee et al., 2013b, Remelhe et al., 2022, Riegler et al., 2015, Dziedzic, 2016, Deng et al., 2016, Dergousoff and Mandryk, 2015)
Crowd-creating	In these studies, the problem can have a loose definition, often creative in nature and requiring uniqueness of contribution.	(Martella et al., 2015, Pothineni et al., 2014, Köse et al., 2019, Apostolopoulos et al., 2016, Kashian et al., 2014, Wang and Lee, 2021, Prandi et al., 2016, Wernbacher et al., 2022, Gorjup et al., 2021, Martelaro et al., 2021, Ogawa et al., 2020, Gamage et al., 2020, Michail and Gavalas, 2019, Prandi et al., 2018, Tahmasbi and Sternberg, 2019, Hantke et al., 2018, Spitz et al., 2018, Tsai and Lee, 2017, Kapenekakis and Chorianopoulos, 2017, L'Heureux et al., 2017, Xu and Zheng, 2016, Laso Bayas et al., 2016, Ling et al., 2021, Hao et al., 2019, Albarqouni et al., 2016, Cáceres et al., 2019, Packham and Suleman, 2015, Pinto and Viana, 2019, Eryigit et al., 2021, Puritat, 2019, Bär, 2022, Uhlgren et al., 2024, Mendoza et al., 2024, Tocchetti et al., 2024, Yang et al., 2024, Ortiz and Tang, 2024, Duggan et al., 2024, Chalwe et al., 2024, Nummenmaa et al., 2025, Lin and Ding, 2023, Jaworski et al., 2023, Kalokyri et al., 2023, Leal, 2023, Stamatiadou et al., 2023, Weber et al., 2023, Baer and Purves, 2022, Sheng, 2013, Miyata et al., 2022, Snijders et al., 2015, Roengsamut et al., 2015, Kacorri et al., 2015, Biegel et al., 2014, Wei and Anwar, 2017)
Crowd-solving	Here the problems were complex and intellectually demanding, requiring considerable effort and the uniqueness of	(Cooper et al., 2010, Haworth et al., 2021, Bowser et al., 2013, Mairittha et al., 2021, Liu et al., 2011, Matin et al., 2020, Seong and Kim,

	contribution, possibly high individual Effort intensity and impact.	2020, Rojas et al., 2017, Allahbakhsh et al., 2018, Weng et al., 2019, Aldwairi and Shuhaiber, 2019, Bounov et al., 2018, Puritat, 2019 , Song et al., 2017, Molasy et al., 2023, Lu et al., 2021, Sakamoto and Nakajima, 2014, Vasilescu et al., 2014, Yakushin and Lee, 2014, Lee et al., 2013a, He et al., 2014, Choi et al., 2014, Nagai et al., 2014, Altmeyer et al., 2016, Melenhorst et al., 2015, De Franga et al., 2015)
--	---	---

References in bold refer to studies that relate to more than one archetype.

A smaller number of studies were designed for experts or employees, often in specialised areas, often in particular fields such as development of software, environmental monitoring, or medical annotation work, where the task accuracy and professional judgement are critical and could not tolerate errors. Teachers, older participants and citizen scientists also figured, especially where learning was involved or where social contribution motives were emphasised. Interestingly, only a handful of studies dedicated substantial space to discussing how participant type affected the effectiveness of gamification methods. This highlights a methodological gap which currently exists in academic literature: the lack of comparative analysis of whether game techniques function differently with experts, students or indeed ordinary users as their target. In summary, although gamified crowdsourcing has been applied to a wide spectrum of worker populations, future work should explore what role the participant demographics, and motivations play in the success of different gamification strategies.

Table 5.2: Participant Types in the Reviewed Studies

Crowdsources	Crowd-processing	Crowd-rating	Crowd-creation	Crowd-solving
Students	(Malatini et al., 2017, Carlier et al., 2015, Prestopnik et al., 2017, Zhou et al., 2017, Packham and Suleman, 2015, Spicher et al., 2024, Hajarian et al., 2024, Paulino et al., 2024, Kawajiri et al., 2014, Remelhe et al., 2022).	(Kirkham et al., 2015, Cáceres et al., 2019, Yang et al., 2024, Amorim et al., 2024, Lin and Ding, 2023, Kalokyri et al., 2023, Leal, 2023, Massung et al., 2013)	(Packham and Suleman, 2015, Cáceres et al., 2019, Yang et al., 2024, Lin and Ding, 2023) (Kalokyri et al., 2023, Leal, 2023, Martella et al., 2015, Prandi et al., 2016, Ogawa et al., 2020, Gamage et al., 2020, Prandi et al., 2018, Spitz et al., 2018, Kapenekakis and Chorianopoulos, 2017, Xu and	(Cooper et al., 2010, Haworth et al., 2021, Bowser et al., 2013, Liu et al., 2011, Martin et al., 2020) (Rojas et al., 2017, Allahbakhsh et al., 2018, Aldwairi and Shuhaiber, 2019, Song et al., 2017, Molasy et al., 2023, Lee et al., 2013a)

			Zheng, 2016, Hao et al., 2019, Uhlgren et al., 2024, Mendoza et al., 2024, Tocchetti et al., 2024, Ortiz and Tang, 2024, Chalwe et al., 2024, Jaworski et al., 2023, Miyata et al., 2022, Roengsamut et al., 2015)	
Teachers	(Yacobson et al., 2021, Spicher et al., 2024)		(Martella et al., 2015, Prandi et al., 2018, Mendoza et al., 2024)	
Researcher			(Tocchetti et al., 2024)	(Molasy et al., 2023, Yakushin and Lee, 2014)
Public citizen	(Purnomo W.P et al., 2019, Ipeirotis and Gabrilovich, 2014, Tinati et al., 2017, Rönneberg and Kettunen, 2023, Nose and Hishiyama, 2013)	(Macakoğlu et al., 2022, Gutiérrez Páez et al., 2021, Linares et al., 2019, Stamatiadou et al., 2023, Goncalves et al., 2014)	(Stamatiadou et al., 2023, Kashian et al., 2014, Tahmasbi and Sternberg, 2019, Tsai and Lee, 2017, L'Heureux et al., 2017, Laso Bayas et al., 2016, Puritat, 2019, Bär, 2022)	(Puritat, 2019)
Expert	(Lendak, 2019, Sadeghianasl et al., 2024)	(Sadeghianasl et al., 2021, Sadeghianasl et al., 2020, Mason et al., 2012, Dumitrache et al., 2013)		(Lu et al., 2021, Melenhorst et al., 2015)
Employee	(Jacob et al., 2022, Lee et al., 2013b)	(Guy et al., 2011)	(Pothineni et al., 2014, Snijders et al., 2015, Kacorri et al., 2015)	
Elderly people	(Itoko et al., 2014)			(Seong and Kim, 2020, Nagai et al., 2014)
Not specified	(Na and Han, 2023, Brenner et al., 2014, Uzun et al., 2013) (Dziedzic, 2016, Bin Hussein et al., 2019, Cordova Mancheno, 2017, Dergousoff and Mandryk, 2015)	(Hong et al., 2013, Morschheuser et al., 2019, Hajarian and Hemmati, 2021, Aljanaki et al., 2016, Brovelli et al., 2018, Good et al., 2014, Rinner et al., 2020, Bujari et al., 2016, Fort et al., 2020, Chen et al., 2018, Pinto and Viana, 2019,	(Pinto and Viana, 2019, Eryiğit et al., 2021, Duggan et al., 2024, Köse et al., 2019, Apostolopoulos et al., 2016) (Wang and Lee, 2021, Wernbacher et al., 2022, Gorjup et al., 2021, Martelaro et al., 2021, Michail and Gavalas, 2019,	(Weng et al., 2019, Sakamoto and Nakajima, 2014, He et al., 2014, Altmeyer et al., 2016, De Franga et al., 2015)

		Eryigit et al., 2021, Bin Hussein et al., 2019, Cordova Mancheno, 2017, Kratchanov and Luchev, 2024, Bastanfard et al., 2024) (Duggan et al., 2024, Duhaime et al., 2023, Chamberlain, 2014) (Eickhoff et al., 2012, Preist et al., 2014)	Hantke et al., 2018, Ling et al., 2021, Albarqouni et al., 2016, Nummenmaa et al., 2025, Weber et al., 2023, Baer and Purves, 2022, Sheng, 2013, Biegel et al., 2014, Wei and Anwar, 2017)	
Figure Eight	(Feyisetan et al., 2015)	(Hallifax et al., 2019)		
Amazon Turk	(Spatharioti, 2020, Rokicki et al., 2014, Lichtenberg et al., 2020, Deng et al., 2016)	(Harris and Srinivasan, 2019, Olinsky et al., 2021, Harris, 2017, Harris, 2014)		(Mairittha et al., 2021, Bounov et al., 2018)
Microworkers	(Walter et al., 2022, Riegler et al., 2015)			
Prolific	(Maddalena et al., 2023)			

5.4.5 Design Elements in Game based crowdsourcing Tasks

The body of reviewed literature revealed that a wide range of different types of Gamification Design Elements (GDEs) have been applied across different types of crowdsourcing platforms. According to Segundo Díaz et al., (2022), more than thirty different GDEs have been used across general gamification domains. For the sake of conceptual consistency and methodological coherence, this review categorises gamification features according to the taxonomy presented by Koivisto and Hamari (2019), which has been introduced as one of the most comprehensive and empirically grounded classifications for gamification resources. Koivisto and Hamari's model distinguishes four main gamification dimensions:

- Achievement-oriented features (e.g., points, badges, leaderboards).
- Social features (e.g. cooperation, competition, teams and chat).
- Immersion elements (as narratives, avatars, aesthetics).
- Real-world-related aspects (e.g., rewards, feedback, progress).

We selected this framework because it is consistent with the motivational principles of SDT (Ryan and Deci, 2000) in terms of satisfying competence, autonomy and relatedness needs. It effectively enables the grouping of design components in terms of their psychological and behavioural outcomes (e.g., motivation, engagement, performance), linking technical designs with motivational theories, which is a necessary step for establishing a theoretical basis for this dissertation.

Usage Overview of Gamification Elements

Across the reviewed literature, scoring, points and leaderboards were the most commonly used game elements, often combined to create a competitive environment among workers (see Table 5.3). Badges and levels were the next common elements that have been deployed in crowdsourcing game tasks, typically used to track people's progress and provide reward milestones. These resources have been proven effective in maintaining engagement and providing visible degrees of achievement and advancement. This focus on achievement-centric "PBL" mechanics fits with larger gamification literature examining the prevalence of points, badges, leaderboards, partly due to their simplicity to deploy, user friendliness and because they were quantifiable (Koivisto and Hamari, 2019, Sailer and Homner, 2020). Contrastingly, collaborative and social elements like chatting, common goals and teamwork were not as broadly utilised even though they have great potential for assisting communication and group problem-solving. This imbalance reflects a general orientation toward competition within gamified platforms for crowds. From an SDT viewpoint (Ryan and Deci, 2000), it suggests that competence signalling is more often endorsed (the use of points or feedback) and relatedness supportive structures (shared goals, interdependence, team interaction) are less consistently designed into the system. Figure 5.3 (Cooperative vs. Competitive Distribution) shows that the majority of designs were of competition-driven designs across most archetypes, with notable variations in the balance between the two approaches:

- Crowd-processing tasks are relied on predominantly for competition-based elements (approximately 86%), emphasising productivity and efficiency.
- Crowd-rating tasks show small inclusion of cooperative features (about 37%), basically through the feedback or mechanisms of peer comparison.
- Crowd-creating tasks are about 30% cooperative: often through shared creative goals or ways of peers voting.
- Crowd-solving tasks show a strong preference for cooperative operations (around 64%), where cooperation is necessary to achieve emergent outcomes.

This distribution reflects that, in prior research, implemented patterns were typically gamified through competitive, individualistic elements across most archetypes, while collaborative features remained underexplored, despite evidence of their benefits in complex and wide-ranging problem domains.

This review reinforces a compelling motivation to evaluate integrated configurations rather than singular measures: if previous systems have tended to support PBL in isolation, then the question becomes whether combining social interaction with progress feedback and shared performance cues leads to stronger and more reliable results than either element alone?

Table 5.3: Gamification elements distribution across reviewed studies

GDE		Crowd-processing	Crowd-rating	Crowd-creating	Crowd-solving
Achievements /Progressions	Points/scores	(Malatini et al., 2017, Purnomo W.P et al., 2019, Carlier et al., 2015, Rokicki et al., 2014, Walter et al., 2022, Lichtenberg et al., 2020, Zhou et al., 2017, Lendak, 2019, Tinati et al., 2017, Packham and Suleman, 2015, Bin Hussein et al., 2019, Cordova Mancheno, 2017, Sadeghianasl et al., 2024, Spicher et al., 2024, Hajarian et al., 2024, Paulino et al., 2024, Maddalena et al., 2023, Rönneberg and Kettunen, 2023, Na and Han, 2023, Jacob et al., 2022, Nose and Hishiyama, 2013, Uzun et al., 2013, Kawajiri et al., 2014, Lee et al., 2013b, Remelhe et al., 2022, Riegler	(Macakoğlu et al., 2022, Hong et al., 2013, Hajarian and Hemmati, 2021, Harris and Srinivasan, 2019, Aljanaki et al., 2016, Gutiérrez Páez et al., 2021, Linares et al., 2019, Brovelli et al., 2018, Guy et al., 2011, Good et al., 2014, Morschheuser et al., 2019, Sadeghianasl et al., 2021, Rinner et al., 2020, Bujari et al., 2016, Olinsky et al., 2021, Sadeghianasl et al., 2020, Fort et al., 2020, Hallifax et al., 2019, Kirkham et al., 2015, Chen et al., 2018, Harris, 2017, Lumsden et al., 2016, Cáceres et al., 2019, Pinto and Eryiğit et al., 2021, Bin Hussein et al., 2019, Cordova Mancheno, 2017,	(Martella et al., 2015, Pothineni et al., 2014, Köse et al., 2019, Apostolopoulos et al., 2016, Kashian et al., 2014, Wang and Lee, 2021, Prandi et al., 2016, Wernbacher et al., 2022, Gorjup et al., 2021, Ogawa et al., 2020, Gamage et al., 2020, Michail and Gavalas, 2019, Prandi et al., 2018, Tahmasbi and Sternberg, 2019, Hantke et al., 2018, Spitz et al., 2018, Tsai and Lee, 2017, Kapenekakis and Chorianopoulos, 2017, L'Heureux et al., 2017, Xu and Zheng, 2016, Laso Bayas et al., 2016, Ling et al., 2021, Hao et al., 2019, Albarqouni et al., 2016, Cáceres et al., 2019, Packham and Suleman, 2015, Pinto and Viana, 2019, Eryiğit et al.,	(Cooper et al., 2010, Haworth et al., 2021, Bowser et al., 2013, Mairittha et al., 2021, Liu et al., 2011, Matin et al., 2020, Rojas et al., 2017, Allahbakhsh et al., 2018, Weng et al., 2019, Aldwairi and Shuhaiber, 2019, Bounov et al., 2018, Puritat, 2019, Song et al., 2017, Lu et al., 2021, Vasilescu et al., 2014, Yakushin and Lee, 2014, Lee et al., 2013a, He et al., 2014, Deng et al., 2016, Altmeyer et al., 2016, Melenhorst et al., 2015, Dergousoff and Mandryk, 2015, De Franga et al., 2015)

Achievements /Progressions		et al., 2015, Dziedzic, 2016, Deng et al., 2016, Dergousoff and Mandryk, 2015)	Kratchanov and Luchev, 2024, Yang et al., 2024, Bastanfard et al., 2024, Duggan et al., 2024, Lin and Ding, 2023, Leal, 2023, Duhaime et al., 2023, Stamatiadou et al., 2023, Harris, 2014, Mason et al., 2012, Massung et al., 2013, Chamberlain, 2014, Dumitrache et al., 2013, Eickhoff et al., 2012, Goncalves et al., 2014, Preist et al., 2014, Saito et al., 2014, Eryigit et al., 2021)	2021, Puritat, 2019, Bär, 2022, Uhlgren et al., 2024) (Yang et al., 2024, Ortiz and Tang, 2024, Duggan et al., 2024, Nummenmaa et al., 2025, Lin and Ding, 2023, Jaworski et al., 2023, Leal, 2023, Stamatiadou et al., 2023, Baer and Purves, 2022, Sheng, 2013, Miyata et al., 2022, Snijders et al., 2015, Roengsamut et al., 2015, Kacorri et al., 2015, Biegel et al., 2014, Wei and Anwar, 2017)	
	Challenges	(Purnomo W.P et al., 2019)	(Eickhoff et al., 2012)	(Weber et al., 2023)	(Molasy et al., 2023)
	Badges	(Malatini et al., 2017, Feyisetan et al., 2015, Walter et al., 2022, Lichtenberg et al., 2020, Lendak, 2019, Bin Hussein et al., 2019, Cordova Mancheno, 2017, Sadeghianasl et al., 2024, Spicher et al., 2024,	(Macakoğlu et al., 2022, Brovelli et al., 2018, Sadeghianasl et al., 2021, Bujari et al., 2016, Sadeghianasl et al., 2020, Hallifax et al., 2019, Cáceres et al., 2019, Pinto and Viana, 2019, Eryigit et al., 2021, Bin Hussein et al.,	(Martella et al., 2015, Köse et al., 2019, Wang and Lee, 2021, Michail and Gavalas, 2019, Prandi et al., 2018, Tahmasbi and Sternberg, 2019, Hantke et al., 2018, Spitz et al., 2018, Hao et al., 2019, Cáceres et al., 2019, Pinto and Viana,	(Bowser et al., 2013, Puritat, 2019, Song et al., 2017, Lu et al., 2021, Vasilescu et al., 2014, Melenhorst et al., 2015, De Franga et al., 2015)

Achievements /Progressions		Hajarian et al., 2024, Maddalena et al., 2023, Itoko et al., 2014, Uzun et al., 2013, Lee et al., 2013b, Remelhe et al., 2022)	2019, Cordova Mancheno, 2017, Yang et al., 2024, Bastanfard et al., 2024, Amorim et al., 2024, Lin and Ding, 2023, Stamatiadou et al., 2023) (Mason et al., 2012, Massung et al., 2013, Preist et al., 2014, Eryigit et al., 2021)	2019, Eryigit et al., 2021, Puritat, 2019, Yang et al., 2024, Ortiz and Tang, 2024, Chalwe et al., 2024, Lin and Ding, 2023, Jaworski et al., 2023, Stamatiadou et al., 2023, Sheng, 2013)	
	Progress	(Feyisetan et al., 2015, Spatharioti, 2020, Yacobson et al., 2021, Lichtenberg et al., 2020, Zhou et al., 2017, Rokicki et al., 2014, Cordova Mancheno, 2017, Rönneberg and Kettunen, 2023, Itoko et al., 2014) (Lee et al., 2013b, Remelhe et al., 2022)	(Macakoğlu et al., 2022, Sadeghianasl et al., 2021, Sadeghianasl et al., 2020, Hallifax et al., 2019, Kirkham et al., 2015, Cordova Mancheno, 2017, Yang et al., 2024, Lin and Ding, 2023)	(Hantke et al., 2018, Hao et al., 2019, Yang et al., 2024, Lin and Ding, 2023)	(Aldwairi and Shuhaiber, 2019, Song et al., 2017, Vasilescu et al., 2014, Lee et al., 2013a) (Nagai et al., 2014)

GDE		Crowd-processing	Crowd-rating	Crowd-creating	Crowd-solving
Achievements /Progressions		(Purnomo W.P et al., 2019, Ipeirotis and Gabrilovich, 2014, Feyisetan et al., 2015, Rokicki et al., 2014, Walter et al., 2022, Lichtenberg et al., 2020, Zhou et al., 2017, Lendak, 2019, Tinati et al., 2017, Packham and Suleman, 2015, Bin Hussein et al., 2019, Cordova Mancheno , 2017, Sadeghianasl et al., 2024, Spicher et al., 2024, Hajarian et al., 2024, Maddalena et al., 2023, Rönneberg and Kettunen, 2023, Na and Han, 2023, Uzun et al., 2013, Kawajiri et al., 2014, Lee et al., 2013b, Remelhe et al., 2022, Dziedzic, 2016, Dergousoff and Mandryk, 2015)	(Hong et al., 2013, Harris and Srinivasan, 2019, Morschheuser et al., 2019, Gutiérrez Páez et al., 2021, Brovelli et al., 2018, Guy et al., 2011, Sadeghianasl et al., 2021, Rinner et al., 2020, Bujari et al., 2016, Fort et al., 2020, Kirkham et al., 2015, Harris, 2017, Pinto and Viana, 2019, Eryiğit et al., 2021, Bin Hussein et al., 2019, Cordova Mancheno , 2017, Kratchanov and Luchev, 2024, Yang et al., 2024, Bastanfard et al., 2024, Duggan et al., 2024, Leal, 2023, Duhaime et al., 2023, Harris, 2014, Massung et al., 2013, Chamberlain, 2014, Dumitrache et al., 2013, Eickhoff et al., 2012, Goncalves et al., 2014, Preist et al., 2014, Saito et al., 2014)	(Martella et al., 2015, Kashian et al., 2014, Wang and Lee, 2021, Gorjup et al., 2021, Martelaro et al., 2021, Gamage et al., 2020, Michail and Gavalas, 2019, Prandi et al., 2018, Tahmasbi and Sternberg, 2019, Hantke et al., 2018, Spitz et al., 2018, Kapenekakis and Chorianopoulos, 2017, L'Heureux et al., 2017, Xu and Zheng, 2016, Albarqouni et al., 2016, Packham and Suleman, 2015, Pinto and Viana, 2019, Eryiğit et al., 2021, Puritat, 2019, Bär, 2022, Yang et al., 2024, Ortiz and Tang, 2024, Duggan et al., 2024, Chalwe et al., 2024, Jaworski et al., 2023, Leal, 2023, Baer and Purves, 2022, Snijders et al., 2015, Roengsamut et al., 2015, Kacorri	(Cooper et al., 2010, Haworth et al., 2021, Bowser et al., 2013, Liu et al., 2011, Matin et al., 2020, Allahbakhsh et al., 2018, Aldwairi and Shuhaiber, 2019, Puritat, 2019, Song et al., 2017, Molasy et al., 2023, Vasilescu et al., 2014, Yakushin and Lee, 2014, Lee et al., 2013a, He et al., 2014, Altmeyer et al., 2016, Melenhorst et al., 2015, Dergousoff and Mandryk, 2015)
	Leaderboard				

Achievements /Progressions				et al., 2015, Biegel et al., 2014, Wei and Anwar, 2017)	
	Levels	(Malatini et al., 2017, Purnomo W.P et al., 2019, Spatharioti, 2020, Lendak, 2019, Bin Hussein et al., 2019, Cordova Mancheno , 2017, Sadeghianasl et al., 2024, Spicher et al., 2024, Hajarian et al., 2024, Brenner et al., 2014, Lee et al., 2013b)	(Linares et al., 2019, Brovelli et al., 2018, Sadeghianasl et al., 2020, Chen et al., 2018, Cáceres et al., 2019, Eryiğit et al., 2021, Bin Hussein et al., 2019, Cordova Mancheno , 2017, Kratchanov and Luchev, 2024, Yang et al., 2024, Bastanfard et al., 2024, Lin and Ding, 2023, Leal, 2023, Stamatiadou et al., 2023, Dumitrache et al., 2013, Eickhoff et al., 2012, Saito et al., 2014)	(Martella et al., 2015, Gorjup et al., 2021, Gamage et al., 2020, L'Heureux et al., 2017, Laso Bayas et al., 2016, Ling et al., 2021, Hao et al., 2019, Albarqouni et al., 2016, Cáceres et al., 2019, Eryiğit et al., 2021, Puritat, 2019, Yang et al., 2024, Lin and Ding, 2023, Leal, 2023, Stamatiadou et al., 2023, Sheng, 2013)	(Bounov et al., 2018, Puritat, 2019, Song et al., 2017, Yakushin and Lee, 2014, Nagai et al., 2014)
Achievements /Progressions	Feedback	(Ipeirotis and Gabrilovich, 2014, Prestopnik et al., 2017, Cordova Mancheno, 2017, Brenner et al., 2014, Jacob et al., 2022, Dziedzic, 2016)	(Harris and Srinivasan, 2019, Aljanaki et al., 2016, Gutiérrez Páez et al., 2021, Sadeghianasl et al., 2021, Fort et al., 2020, Cordova Mancheno , 2017, Bastanfard et al., 2024, Amorim et al., 2024, Duggan et al., 2024, Lin and Ding, 2023)	(Kashian et al., 2014, Albarqouni et al., 2016, Mendoza et al., 2024, Lin and Ding, 2023, Duggan et al., 2024, Weber et al., 2023)	(Mairittha et al., 2021, Liu et al., 2011, Weng et al., 2019, Bounov et al., 2018, Song et al., 2017, Yakushin and Lee, 2014, Lee et al., 2013a)

GDE		Crowd-processing	Crowd-rating	Crowd-creating	Crowd-solving
Achievements /Progressions	Goal	(Yacobson et al., 2021)	(Hong et al., 2013, Hallifax et al., 2019)		(Sakamoto and Nakajima, 2014, Lee et al., 2013a)
	Timer	(Riegler et al., 2015)	(Hallifax et al., 2019, Yang et al., 2024, Kalokyri et al., 2023)	(Yang et al., 2024, Kalokyri et al., 2023)	(Matin et al., 2020)
Social	Team		(Hong et al., 2013, Morschheuser et al., 2019, Cáceres et al., 2019)	(Xu and Zheng, 2016, Cáceres et al., 2019)	(Aldwairi and Shuhaiber, 2019, Molasy et al., 2023)
	Cooperation	(Carlier et al., 2015, Spatharioti, 2020)	(Hong et al., 2013, Morschheuser et al., 2019, Aljanaki et al., 2016, Pinto and Viana, 2019, Yang et al., 2024, Kalokyri et al., 2023, Leal, 2023)	(Pothineni et al., 2014, Ogawa et al., 2020, Spitz et al., 2018, Xu and Zheng, 2016, Pinto and Viana, 2019, Yang et al., 2024, Kalokyri et al., 2023, Kacorri et al., 2015)	(Cooper et al., 2010, Haworth et al., 2021, Seong and Kim, 2020, Aldwairi and Shuhaiber, 2019, Song et al., 2017, Yakushin and Lee, 2014)
	Competition	(Malatini et al., 2017, Purnomo W.P et al., 2019, Ipeirotis and Gabrilovich, 2014, Feyisetan et al., 2015) (Spatharioti, 2020, Rokicki et al., 2014, Walter et al., 2022, Lichtenberg et al., 2020, Zhou et al., 2017, Lendak, 2019, Tinati et al., 2017, Packham and Suleman, 2015, Bin Hussein et al., 2019, Cordova Mancheno, 2017, Sadeghianasl et al., 2024, Spicher et al., 2024, Hajarian et al., 2024, Maddalena et al., 2023, Rönneberg and Kettunen, 2023, Na and Han, 2023, Lee et al., 2013b, Dziedzic, 2016)	(Morschheuser et al., 2019, Harris and Srinivasan, 2019) (Gutiérrez Páez et al., 2021, Linares et al., 2019, Brovelli et al., 2018, Guy et al., 2011, Sadeghianasl et al., 2021, Rinner et al., 2020, Bujari et al., 2016, Sadeghianasl et al., 2020, Fort et al., 2020, Hallifax et al., 2019, Kirkham et al., 2015, Harris, 2017, Cáceres et al., 2019, Bin Hussein et al., 2019, Cordova Mancheno, 2017, Kratchanov and Luchev, 2024, Bastanfard et al., 2024, Amorim et al., 2024, Duggan et al., 2024, Lin and Ding, 2023, Duhaime et al., 2023, Mason et al., 2012, Massung et al., 2013, Dumitrache et al., 2013)	(Martella et al., 2015, Köse et al., 2019, Apostolopoulos et al., 2016, Kashian et al., 2014, Wang and Lee, 2021, Prandi et al., 2016, Wernbacher et al., 2022, Gorjup et al., 2021, Martelaro et al., 2021, Gamage et al., 2020, Michail and Gavalas, 2019, Prandi et al., 2018, Tahmasbi and Sternberg, 2019, Hantke et al., 2018, Kapenekakis and Chorianopoulos, 2017, L'Heureux et al., 2017, Hao et al., 2019, Albarqouni et al., 2016, Cáceres et al., 2019, Packham and Suleman, 2015, Puritat, 2019, Bär, 2022, Ortiz and Tang, 2024, Duggan et al., 2024, Chalwe et al., 2024, Lin and Ding, 2023, Jaworski et al., 2023, Leal, 2023, Baer and Purves, 2022, Miyata et al., 2022, Wei and Anwar, 2017)	(Cooper et al., 2010, Haworth et al., 2021, Bowser et al., 2013, Liu et al., 2011, Matin et al., 2020, Rojas et al., 2017, Allahbakhsh et al., 2018, Puritat, 2019, Molasy et al., 2023, Lu et al., 2021)
	Customization	(Spatharioti, 2020)		(Xu and Zheng, 2016, Hao et al., 2019)	

	Social networking	(Lee et al., 2013b)	(Aljanaki et al., 2016, Hallifax et al., 2019)	(Martella et al., 2015, Pothineni et al., 2014, Michail and Gavalas, 2019, Snijders et al., 2015)	(Weng et al., 2019, Lee et al., 2013a)
	Chat	(Tinati et al., 2017) (Maddalena et al., 2023)	(Cáceres et al., 2019)	(Martelaro et al., 2021, Cáceres et al., 2019)	(Cooper et al., 2010)
Immersion	Virtual world	(Paulino et al., 2024)	(Kalokyri et al., 2023)	(Martella et al., 2015, Prandi et al., 2016, Ogawa et al., 2020, Tsai and Lee, 2017, Tocchetti et al., 2024, Nummenmaa et al., 2025, Kalokyri et al., 2023, Sheng, 2013)	
	Avatar	(Maddalena et al., 2023, Paulino et al., 2024, Dergousoff and Mandryk, 2015)	(Morschheuser et al., 2019, Olinsky et al., 2021)	(Martella et al., 2015, Prandi et al., 2016, Wernbacher et al., 2022, Ogawa et al., 2020, Tsai and Lee, 2017, Ling et al., 2021, Hao et al., 2019, Mendoza et al., 2024, Miyata et al., 2022)	(Lee et al., 2013a, Yakushin and Lee, 2014, Nagai et al., 2014)
	Storyline			(Tocchetti et al., 2024, Weber et al., 2023)	(Seong and Kim, 2020, Sakamoto and Nakajima, 2014)
	Role Playing				(Nagai et al., 2014)
	Aesthetics				
Other	Sound	(Walter et al., 2022)			
	Light	(Walter et al., 2022)			
	Board	(Yacobson et al., 2021)			
Rewards	Monetary	(Feyisetan et al., 2015, Spatharioti, 2020, Rokicki et al., 2014, Walter et al., 2022, Packham and Suleman, 2015, Deng et al., 2016)	(Hallifax et al., 2019, Harris, 2017, Lumsden et al., 2016, Bin Hussein et al., 2019, Massung et al., 2013, Chamberlain, 2014, Eickhoff et al., 2012, Preist et al., 2014, Eryigit et al., 2021)	(Köse et al., 2019, Apostolopoulos et al., 2016, Kashian et al., 2014) (Martelaro et al., 2021, Xu and Zheng, 2016, Packham and Suleman, 2015, Kacorri et al., 2015)	(Matin et al., 2020, Bounov et al., 2018, Choi et al., 2014, Deng et al., 2016, Melenhorst et al., 2015)
	Voucher		(Hajarian and Hemmati, 2021, Massung et al., 2013)	(Prandi et al., 2016, Laso Bayas et al., 2016)	

Overall Analysis of the Design Patterns

- Competitive elements, such as points, leaderboards and badges were consistently associated with higher task throughput and responses accuracy. In all archetype we could see in Figure 5.3 and Figure 5.4, that the competitive elements including achievement have a higher rate than the collaborative

elements. However, it is also known from multiple sources that excessive competition may also risk overshadowing intrinsic motivation or dishearten slower participants, a limitation observed in several studies (see representative examples in Table 5.3). This trade-off is well recognised in motivation studies, in which controlling or overly comparative reward structures have the ability to inhibit autonomous motivation for some participants (Ryan and Deci, 2000)).

- Immersion-related elements, such as avatars, storytelling and aesthetic feedback, appeared in a few subset studies but demonstrated a notable impact on emotional engagement and worker enjoyment. For instance, narrative structures were more common in crowd-solving tasks, where framing challenges such as “quests” or “missions” increased perceived purpose and persistence (see Figure 5.5).
- Social elements as shown in Figure 5.6, including other social shared goals, chat, and team structures, were mainly realised in crowd-creating and crowd-solving, and few crowd-creating systems, where they were used to foster communication and collective ideation. However, channels for their implementation remain limited, revealing notable a research gap in exploring how collaboration resource might be utilised to enhance engagement. Significantly, even with social elements, many papers refer to them as an add-on (e.g., shared state, structured feedback loops, and mutual awareness), not closely correlated coordination mechanisms, creating uncertainty about when communication results in higher-quality collective decisions.
- Real-world elements, specifically financial rewards and progress feedback, were implemented to reinforce people's extrinsic motivations and provide tangible incentives. However, a small number of papers have touched on how these rewards interact with intrinsic motivation or collective dynamics.

Overall, the reviewed studies showed that most gamified crowdsourcing systems remain task-achievement-oriented, prioritising quantity indicator of success over social or immersion dimensions. On the other hand, balanced integration between competitive and cooperative modes was not really evident in any of the studies reviewed for this thesis, even though theory and preliminary findings both argue that hybrid approaches may better engage people and produce optimal performance. These findings build a foundation for this thesis' later design decision to integrate progress, feedback and scoring structures with team interaction in real time, and thus the study will be testing whether coordination gains relevance for performance when teams are able to see and interpret shared performance information rather than only exchange messages.

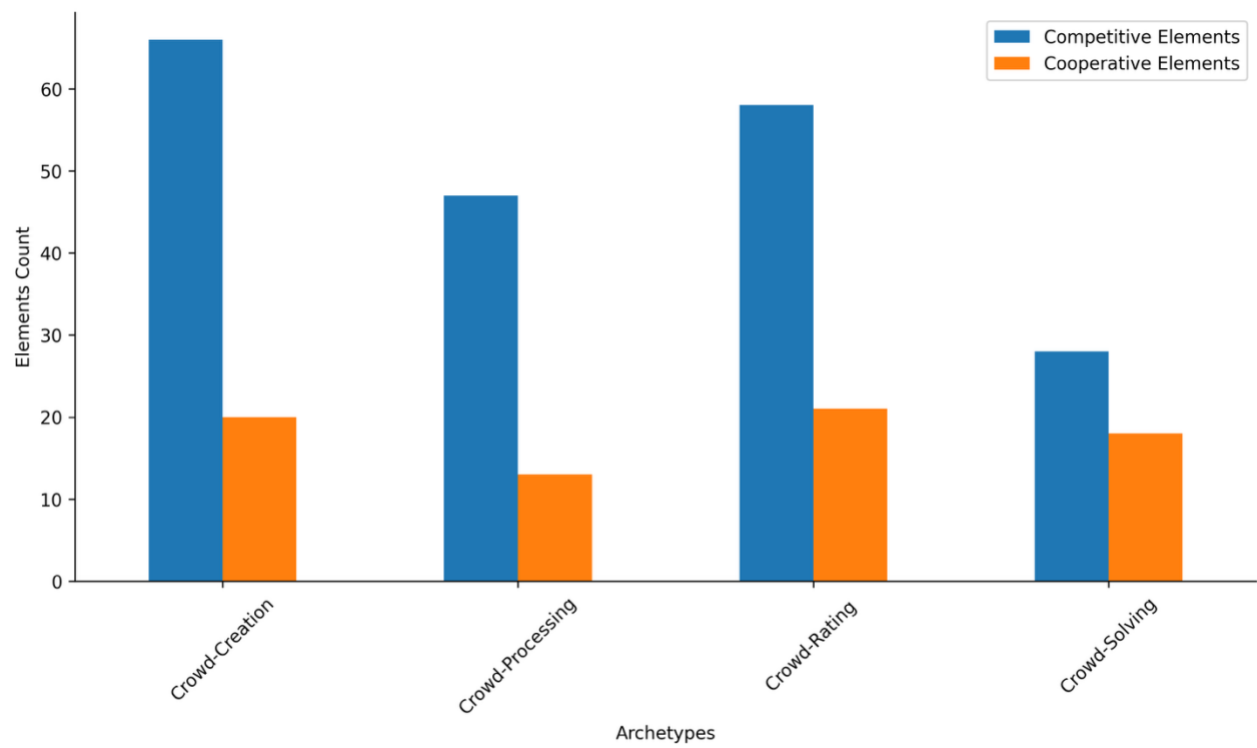


Figure 5.3: Cooperative vs competitive across all archetypes

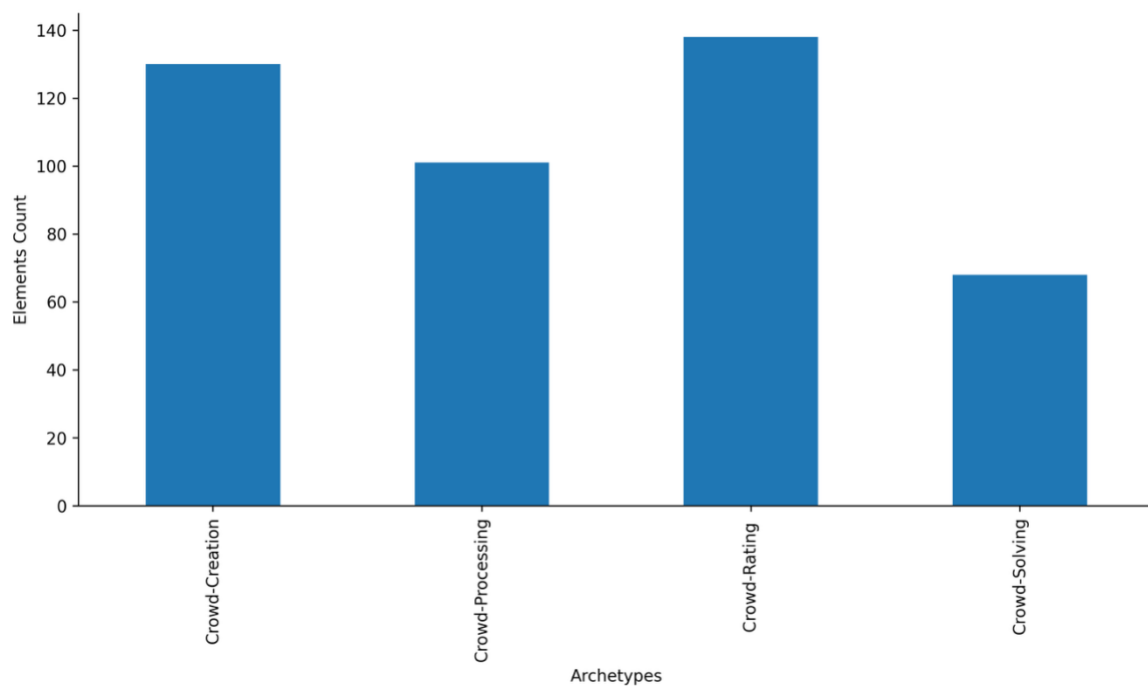


Figure 5.4: Achievement element usage across all archetypes

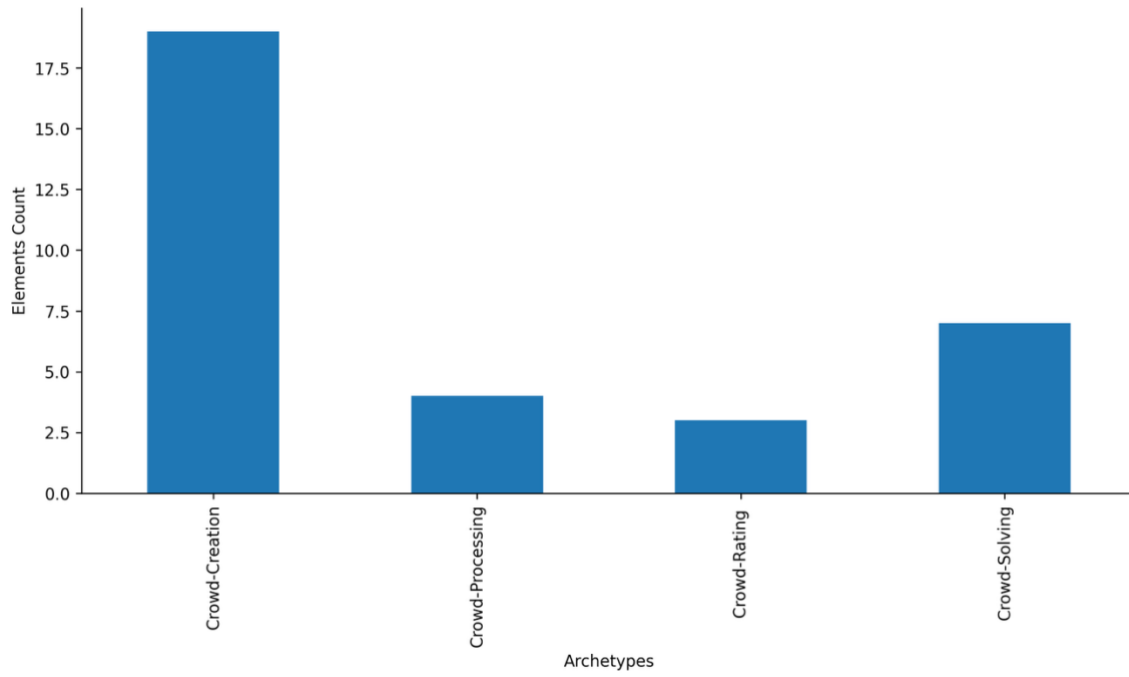


Figure 5.5: Immersion element usage across archetypes

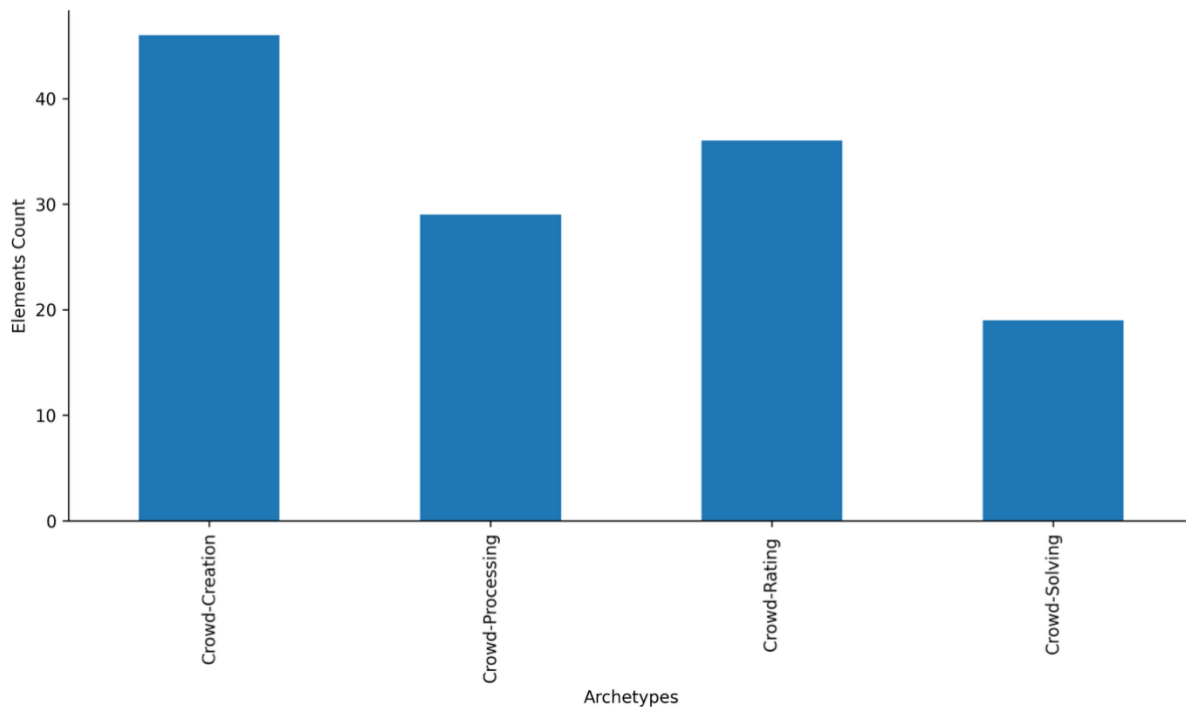


Figure 5.6: Social Gamification elements across all archetypes

5.4.6 Trends and Implications at the Archetype Level

The reviewed study papers revealed that the implementation of gamification elements and the designs vary across the four crowdsourcing archetypes. Each type showed different trends that reflects its underlying structure, motivational features that have been used, and social dynamics (Figure 5.7). This archetype-level perspective is significant because it indicates that ‘gamification in crowdsourcing’ is not a monolithic design phenomenon, the same feature (e.g., points, leaderboards, etc.) is repeatedly utilised, but the functional uses of those things change according to whether systems prioritise throughput, judgement aggregation, creativity, or complex problem solving.

Crowd-processing systems were dominated by competitive, achievement-oriented elements like points, levels and leaderboards. For repetitive assignments it was often used as a means to speed up work and so improve efficiency. Studies implemented such mechanisms (Bin Hussein et al., 2019, Hajarian et al., 2024, Lendak, 2019, Tinati et al., 2017, Walter et al., 2022, Lichtenberg et al., 2020, Zhou et al., 2017), reported improvements in the workers outcomes in term of accuracy. However, peer-based or cooperative features were largely absent in crowd-processing archetypes. This indicated a clear limitation of experimental studies in such trend by which most of these study archetype prioritise individual productivity over social engagement or shared goal structure. A future design that incorporates a lightweight cooperative feedback form or peer validation mechanisms may maintain the workers’ performance while fostering high motivation. In this thesis, this pattern supports treating productive gamification orientation as a baseline style of the design and motivates evaluating whether adding structured interaction changes both the engagement and outcome quality.

Crowd-rating studies are remarkably similar to the design employed in crowd-processing, often considering points, scores, leaderboards, and badges as core motivators (Morschheuser et al., 2019, Sadeghianasl et al., 2021). Such trends in achievements oriented in crowd-rating tasks encourage a high volume of contributors and reduced participant attrition, but also, they risk encouraging quantity driven behaviour over reflective judgment. A significant body of literature in this archetype investigates competition (Brovelli et al., 2018, Fort et al., 2020), but there is more than crowd-processing interest in cooperation features, such as the presence of social aspects, including social networking (Aljanaki et al., 2016, Hallifax et al., 2019, Kalokyri et al., 2023, Pinto and Viana, 2019); emphasises the significance of social influence in rating activities (see Figure 5.6). Feedback components (explicit performance feedback and progress indicators) are predominantly employed in crowd-rating studies as (Cordova Mancheno, 2017, Sadeghianasl et al., 2021), compared to other archetypes. This hybrid design method, combining competitive and cooperation incentives suggests that crowdsourcing systems benefits from immediate, transparent feedback and validation that improve both accuracy and contribution consistency. The finding directly applies later to experimental design decisions in this thesis, as this implies that feedback and shared evaluation cues are a persistent mechanism for stabilising contribution reliability.

Crowd-solving archetypes constituted the most complex and high cognition-demand tasks environment, wherein workers engaged in open-ended problem-solving or optimisation tasks. Gamification resources in this type often included challenges, contests and time constraints, which exhibited hackathon-style environments (Matin et al., 2020). Although these designs produced excitements and time-sensitive engagement, they overemphasised competition at the expense of collaboration. In contrast, other studies implemented more balanced or cooperative models such as in Foldit (Cooper et al., 2010), which indicated the value of shared discovery, and collective learning. These examples indicated that the good balance between motivation for competition with collaboration can not only greatly improve the quality of outcomes but also increased participant satisfaction. This contrast indicates a common design tension in the literature; in comparison, time-limited competition can lead to participation while coordination mechanisms are more directly associated with learning, refinement, and more quality gains.

Another prominent characteristic of this archetype involved abundant use of storyline or narrative elements (Seong and Kim, 2020, Sakamoto and Nakajima, 2014), which facilitated emotional involvement and a greater understanding of the context for the problem. These features, which immerse participants in the problem as a whole, contribute to increased emotional engagement and show that narrative framing can be an effective design strategy, rather than just an add-on, in intense cognitive crowdsourcing scenes. A discernible trend towards tangible incentives in crowd-solving studies is evident, as numerous research studies include monetary rewards (Matin et al., 2020, Choi et al., 2014), which are less frequently emphasised in crowd-processing and crowd-rating, perhaps owing to the substantial cognitive work and competence demanded by these tasks.

Crowd-creation systems have a significant propagation of competitive attributes such as points, leaderboards, ranking techniques as in many studies (Martelaro et al., 2021, Gamage et al., 2020, Michail and Gavalas, 2019, Prandi et al., 2018, Tahmasbi and Sternberg, 2019, Hantke et al., 2018, Kapenekakis and Chorianopoulos, 2017, L'Heureux et al., 2017, Hao et al., 2019, Albarqouni et al., 2016, Cáceres et al., 2019, Packham and Suleman, 2015). While these resources help enhance productivity, they reduced the opportunities for collaboration and sharing the knowledge, which are necessary for co-creation and creativity tasks. A smaller numbers of studies applied the cooperative features such as shared creative space, communication tools or forums for peer discussion, which demonstrated benefits both workers performance and engagement (Ogawa et al., 2020, Spitz et al., 2018, Xu and Zheng, 2016, Pinto and Viana, 2019, Yang et al., 2024, Kalokyri et al., 2023, Kacorri et al., 2015). While that the cooperations in the content creation have been success in production of positive attitude of the crowd's engagement and performance, unfortunately, much research remains to be individuals' efforts instead of promoting a group driven experience, leaving a space for future research into examining how collective quality can be enhanced through social gamification elements. Furthermore, the high emphasis of customisation gamification resource in crowd-creation archetype such as user-customisation avatars personalised interfaces (Xu and Zheng, 2016, Hao et al., 2019), indicated a

focus on player autonomy and user-driven interaction. The features mentioned here seem to work in favour of intrinsic motivation and can help to bridge the gap between competitive and cooperative engagement. Moreover, the integration of social resources such as comments or team conversations discussion may enhance the engagement and collaboration, which are minimal in crowd- processing, rating and creating. This pattern supports the needs for examination of collaborative labelling as a missing configuration in the literature; co-creation tasks often have some competitive motivators, but fewer systems closely integrate real-time coordination and shared progress cues that might facilitate collective sensemaking. Throughout all archetypes, specific design features, such points and badges, appeared universally, though their psychological function varied. In crowd-processing, they enhance productivity, in crowd-rating they indicate reliability, in crowd-creating they embody originality, and in crowd-solving they signify mastery of challenges. Similarly, leaderboards were implemented variably functioning as ranking instruments in processing, reputation systems in crowd-rating, and visibility boosters in crowd-solving and creating. Immersion features, such as avatars, narratives or aesthetics, appeared less overall, but when used, they contribute well to enhance satisfaction and engagement. This indicated a missed opportunity for combined the emotional and dimensions of aesthetic into otherwise performance-driven designs.

Overall, based on the observed distribution of features, as they have been showed in Figure 5.7, suggest that the current crowdsourcing research has predominantly focused on achievement-based and extrinsic motivation mechanisms, while collaborative, social and intrinsically oriented motivational approaches were comparatively less presented. This patterns further underlines the need for integrated and balanced motivational designs which combined competition, collaboration, and immersions to optimise and enhancing performance and engagement across different archetypes. Taken together, these archetype-level patterns led to an interpretive insight generated through the synthesis conducted in this review: the effectiveness and motivational function of gamification cannot be understood independently of the task context and the characteristics of the people performing it. This emerged from comparing how similar gamification elements were implemented and functioned differently across the reviewed crowdsourcing archetypes. For example, achievement-oriented elements were predominantly associated with productivity and throughput in crowd-processing, whereas social, cooperative, narrative, and customisation features were more visible in contexts requiring interaction, creativity, or complex problem solving. These variations suggest that gamification should not be treated as a one-size-fits-all intervention, but as a configurable motivational design space whose effectiveness may depend on both task characteristics and individual worker preferences

5.4.7 Measurement of Psychological and Behavioural Outcomes

The reviewed literature examined how the utilisation of gamification resources affects participants' psychological experiences, including engagement, enjoyment, motivation, and behavioural outcomes such as the quality of the output and the task performance. However, these constructs were often assessed

inconsistently among studies, indicated a lack of a unified framework for measurement within gamified crowdsourcing research. Most papers measured the enjoyment and engagement using either basic questionnaires or indirect participant behaviour, rather than standard psychometric scales dimension. We determined that a limited number of studies measured the enjoyments of the players by using validated instruments to assess the constructs of this enjoyment, such as, flow, positive affect, or immersion. For example, studies use a survey-based questions to measure their enjoyment construct such as in (Spatharioti, 2020, Preist et al., 2014, Saito et al., 2014, Bowser et al., 2013, Sheng, 2013), whereas other literatures measure the enjoyment by other factors of the worker behaviour such as workers number of logging times, number of clicks and accumulated tags they gain during the task on the desired elements (Na and Han, 2023).

Other studies have used platform-specific performance metrics such as accounting for the number of participant and the amount of collected scores, badges or leaderboard as indicator of workers engagement (Mason et al., 2012, Eickhoff et al., 2012, Saito et al., 2014, Brovelli et al., 2018, Feyisetan et al., 2015) (Nose and Hishiyama, 2013, Ipeirotis and Gabrilovich, 2014, Itoko et al., 2014). While these indicators offer tangible proof of activity, they may confuse productivity with authentic psychological engagement, so constraining their interpretive value. Performance-based assessments exhibited greater consistency, primarily emphasising accuracy, task completion rates, or concordance with expert-validated outcomes.

Nonetheless, limited research has directly associated performance enhancements with particular gamification mechanisms or motivated states, resulting in mainly unexamined causal pathways. This highlights the importance of subjective experience measures such as behavioural logs, so engagement is not based solely on more work done. Table 5.4 encapsulates the primary components and instruments distinguished across the four archetypes. In summary:

- Engagement was the most commonly analysed concept, assessed via both subjective evaluations and behavioural data.
- Enjoyment and fun were mostly examined in creative or problem-solving tasks, frequently by self-reported emotional responses.
- Motivation was sporadically examined by qualitative feedback, rather than through validated scales.
- Performance outcomes were extensively documented by comparing user's outcome to expert-validated outcome but barely corroborated with psychological assessments.

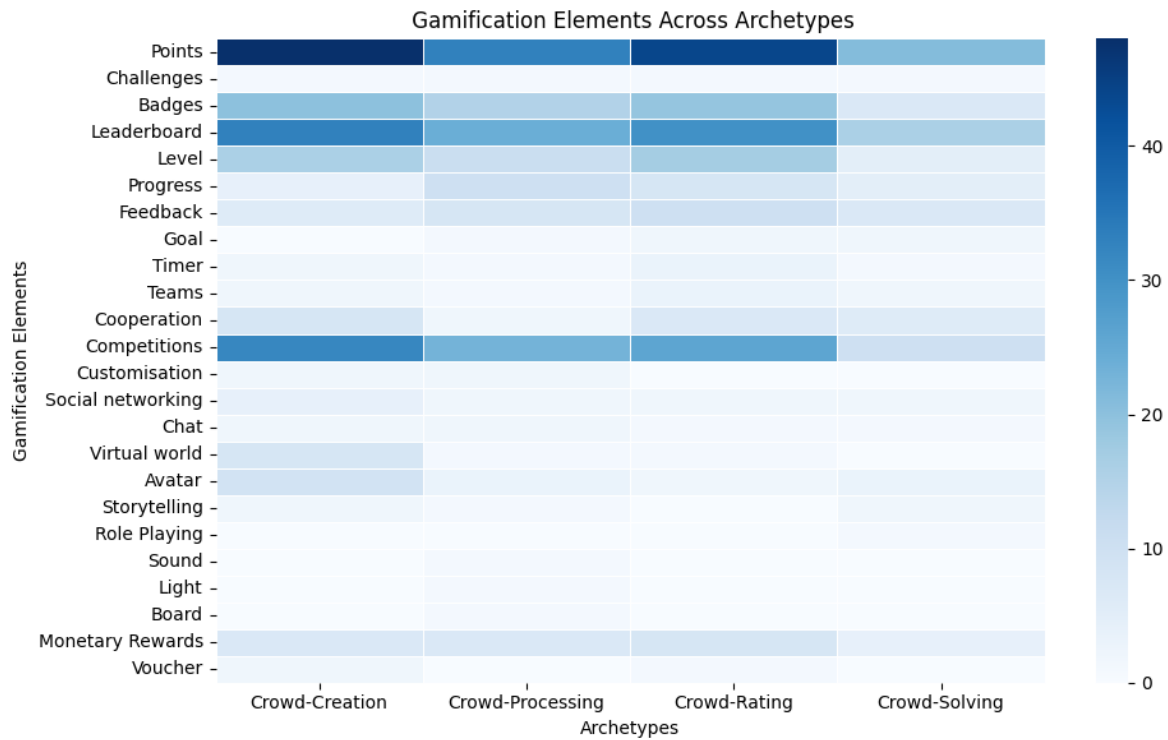


Figure 5.7: Heatmap of gamification elements by archetype

The investigation indicates that whereas gamification significantly improves observable engagement and results in various crowdsourcing scenarios, a comprehensive evaluation of its psychological effects is still lacking. This gap directly supports the measurement strategy used in the thesis, where validated instruments are paired with behavioural and agreement-based quality metrics to test whether a better engagement co-occurs with better data quality. Subsequent research ought to employ validated frameworks, such as SDT-based motivation and engagement inventories, or Flow State Scale (Jackson and Eklund, 2002), to obtain more profound insights into the influence of specific gamification features on user experience and performance. A further insight emerging from this synthesis concerns individual differences in responses to gamified task design. Across the reviewed studies, behavioural and psychological outcomes such as engagement, enjoyment, motivation, participation, and performance were operationalised in different ways and were rarely examined systematically in relation to worker characteristics. Consequently, the literature provides limited direct evidence about whether particular gamification configurations are equally effective for all workers. However, when the patterns identified across the reviewed studies are considered together, they suggest that the motivational effects of gamification may depend not only on the elements implemented and the crowdsourcing context, but also on how individual workers experience and respond to those elements. This interpretation is a synthesis generated from the comparative evidence in this review and also connects with the findings of Study 1. The exploratory study demonstrated heterogeneity in workers' task-selection preferences and engagement

experiences, with task interest ranked second only to monetary reward, while personality-related differences were more evident in subjective engagement and interest than in objective accuracy. Study 1 findings provide empirical support for the broader proposition emerging from this review that motivational task design is unlikely to affect all workers uniformly. Together, the two studies therefore motivate a more configurable and context-sensitive approach to gamification.

Table 5.4: Behavioural and psychological outcomes on gamified crowdsourcing types

Construct	Crowd-processing	Crowd-solving	Crowd-creation	Crowd-rating
Engagement	(Bin Hussein et al., 2019, Purnomo W.P et al., 2019, Ipeirotis and Gabrilovich, 2014, Rokicki et al., 2014, Walter et al., 2022, Yacobson et al., 2021, Lichtenberg et al., 2020, Zhou et al., 2017, Lendak, 2019, Tinati et al., 2017, Packham and Suleman, 2015, Sadeghianasl et al., 2024, Paulino et al., 2024, Rönneberg and Kettunen, 2023, Uzun et al., 2013, Remelhe et al., 2022, Pothineni et al., 2014)	(Puritat, 2019, Cooper et al., 2010, Bowser et al., 2013, Liu et al., 2011, Seong and Kim, 2020, Rojas et al., 2017, Weng et al., 2019, Bounov et al., 2018, Song et al., 2017, Molasy et al., 2023, Lu et al., 2021, Altmeyer et al., 2016, De Franga et al., 2015)	(Cáceres et al., 2019, Eryiğit et al., 2021, Leal, 2023, Packham and Suleman, 2015, Martella et al., 2015, Pothineni et al., 2014, Kashian et al., 2014, Wang and Lee, 2021, Prandi et al., 2016, Wernbacher et al., 2022, Ogawa et al., 2020, Michail and Gavalas, 2019, Tahmasbi and Sternberg, 2019, Hantke et al., 2018, Spitz et al., 2018, Tsai and Lee, 2017, Kapenekakis and Chorianopoulos, 2017, Ling et al., 2021, Puritat, 2019, Bär, 2022, Uhlgren et al., 2024, Mendoza et al., 2024, Tocchetti et al., 2024, Ortiz and Tang, 2024, Chalwe et al.,	(Morschheuser et al., 2019, Brovelli et al., 2018, Sadeghianasl et al., 2021, Bujari et al., 2016, Macakoğlu et al., 2022, Olinsky et al., 2021, Sadeghianasl et al., 2020, Cáceres et al., 2019, Eryiğit et al., 2021, Bin Hussein et al., 2019, Amorim et al., 2024, Leal, 2023, Mason et al., 2012, Massung et al., 2013, Saito et al., 2014)

			2024, Nummenmaa et al., 2025, Weber et al., 2023, Miyata et al., 2022, Snijders et al., 2015, Wei and Anwar, 2017)	
Fun/Enjoyment	(Rönneberg and Kettunen, 2023)	(Lee et al., 2013a, Choi et al., 2014, Altmeyer et al., 2016, Melenhorst et al., 2015)	(Lin and Ding, 2023, Kalokyri et al., 2023, Köse et al., 2019, Prandi et al., 2018, Weber et al., 2023, Sheng, 2013, Roengsamut et al., 2015)	(Morschheuser et al., 2019, Brovelli et al., 2018, Good et al., 2014, Sadeghianasl et al., 2021, Lin and Ding, 2023, Kalokyri et al., 2023, Eickhoff et al., 2012)
Motivations	(Na and Han, 2023, Nose and Hishiyama, 2013, Itoko et al., 2014, Kawajiri et al., 2014, Lee et al., 2013b, Riegler et al., 2015)	(Matin et al., 2020, Seong and Kim, 2020, Lu et al., 2021, He et al., 2014, Nagai et al., 2014, Altmeyer et al., 2016)	(Eryigit et al., 2021, L'Heureux et al., 2017, Hao et al., 2019, Jaworski et al., 2023, Baer and Purves, 2022, Miyata et al., 2022, Roengsamut et al., 2015)	(Gutiérrez Páez et al., 2021, Hallifax et al., 2019, Eryigit et al., 2021, Bastanfard et al., 2024, Preist et al., 2014)
Performance	(Malatini et al., 2017, Carlier et al., 2015, Prestopnik et al., 2017, Spicher et al., 2024, Jacob et al., 2022, Brenner et al., 2014, Dziedzic, 2016, Deng et al., 2016)	(Haworth et al., 2021, Matin et al., 2020, Allahbakhsh et al., 2018, Yakushin and Lee, 2014, Choi et al., 2014, Deng et al., 2016, Melenhorst et al., 2015)	(Duggan et al., 2024, Lin and Ding, 2023, Kalokyri et al., 2023, Stamatiadou et al., 2023, Gamage et al., 2020, L'Heureux et al., 2017, Albarqouni et al., 2016)	(Hajarian and Hemmati, 2021, Harris and Srinivasan, 2019, Aljanaki et al., 2016, Gutiérrez Páez et al., 2021, Linares et al., 2019, Guy et al., 2011, Good et al., 2014, Fort et al., 2020, Harris, 2017, Bastanfard et al., 2024, Duggan et al., 2024, Lin and Ding, 2023, Kalokyri et al.,

				2023, Duhaime et al., 2023, Stamatiadou et al., 2023, Harris, 2014, Chamberlain, 2014, Dumitrache et al., 2013, Eickhoff et al., 2012, Goncalves et al., 2014)
Engagement + Performance	(Cordova Mancheno , 2017, Feyisetan et al., 2015, Spatharioti, 2020, Hajarian et al., 2024, Maddalena et al., 2023, Dergousoff and Mandryk, 2015)	(Mairitha et al., 2021, Dergousoff and Mandryk, 2015)	(Pinto and Viana, 2019, Yang et al., 2024, Apostolopoulos et al., 2016, Gorjup et al., 2021, Martelaro et al., 2021, Xu and Zheng, 2016, Laso Bayas et al., 2016, Kacorri et al., 2015, Biegel et al., 2014)	(Hong et al., 2013, Rinner et al., 2020, Kirkham et al., 2015, Chen et al., 2018, Lumsden et al., 2016, Pinto and Viana, 2019, Cordova Mancheno , 2017, Kratchanov and Luchev, 2024, Yang et al., 2024)

5.4.8 Theories of Enjoyment

Only a limited number of studies on crowdsourcing gamified tasks explicitly anchored their GDE or evaluation within the established psychological theories of motivation or enjoyment in their work. However, those few studies provided valuable theoretical foundations, illustrating how gamification can affect participant outcome quality and their engagement. Based on Flow Theory (Csikszentmihalyi, 1990), several studies drew their conclusion to explain how gamification elicits immersion and an intrinsically rewarding state. For example, (Dumitrache et al., 2013) and (Eickhoff et al., 2012) linked competition and worker performance feedback to workers' immersion, interpreting the experience as an indicator of the workers' involvement and flow in the gamified task, this a state characterised by focus and balance between challenge and skill. Other researchers, such as (Hong et al., 2013), incorporated three motivational theory aspects, identifying achievement, social and immersion as key design motivational features to enhance worker involvement and flows in the gamified task. Cognitive Dissonance Theory (Aronson, 1969), was deployed in (Choi et al., 2014) to examine how the diversity and anticipation of the rewards influence perception of the level of fun and task satisfaction. Additionally, the study assumed that reward diversity creates a stimulating sense of challenge, and sense of fun which led to higher engagement and improved the quality of the contribution. Moreover, other crowdsourcing studies such as Morschheuser et al. (2019), combined two other theories in order to analyse the

workers behaviour by using their gamification elements, the Goal-setting theory (Locke and Latham, 1990), and Social Interdependence theory (Johnson, 2003), to analyse how goal structures and team dynamics shape motivation in gamified environments. Their results demonstrated that clear, attainable goals enhance commitment, while cooperative interdependence among team members fosters persistence and mutual reinforcement of effort. Table 5.5 summarises the theoretical frameworks adopted across different crowdsourcing archetypes, showing that SDT and Flow Theory appeared most often where theory is explicitly referenced, while other motivational frameworks appear only sporadically. Despite these contributions, the review reveals that most studies treat gamification as a design strategy rather than a theoretically grounded motivational framework. The lack of explicit theory use limits the interpretive depth of many findings and hinders the development of cumulative knowledge across studies. Moreover, the study of Humayun et al. (2024) reported that across 91 studies on crowdsourcing, only a limited number of combined aspects of two motivational theories simultaneously, and that the usage of theories is mainly selective rather than comprehensive. This indicates that there is still a need to apply the full motivational capabilities of SDT, and the relatedness dimension is particularly applicable under collaborative configurations, when designing crowdsourcing in the field. Thus, to advance the field, future research should systematically align gamification design elements with established multiple motivational theories, enabling the measurement of psychological constructs, such as autonomy competence satisfaction,

Table 5.5: Implemented theories based on different crowdsourcing types.

Game Theory	Crowd-processing	Crowd-solving	Crowd-creation	Crowd-rating
Self-Determination Theory	(Zhou et al., 2017, Lichtenberg et al., 2020, Sadeghianasl et al., 2024, Na and Han, 2023, Jacob et al., 2022, Ipeirotis and Gabrilovich, 2014, Feyisetan et al., 2015, Prestopnik et al., 2017)	(Lu et al., 2021)	(Martella et al., 2015, Hao et al., 2019, Weber et al., 2023)	(Morschheuser et al., 2019)
Flow Theory				(Dumitrache et al., 2013, Eickhoff et al., 2012)
Motivational Theory				(Hong et al., 2013)
Theory of Norm				(Preist et al., 2014)
Cognitive Dissonance Theory		(Choi et al., 2014)		
Goal Setting Theory				(Morschheuser et al., 2019)

Social Interdependences Theory				(Morschheuser et al., 2019)
--------------------------------------	--	--	--	--------------------------------

flow, and goal clarity, and their causal links to behavioural outcomes like task accuracy, persistence, and engagement.

5.5 Discussions

5.5.1 Objectives of Review

This scoping study presented a thorough analysis and evaluation of gamification application in crowdsourcing, synthesising evidence from 146 empirical studies. Through the integration of the categorisation framework of the (Geiger et al., 2012, Koivisto and Hamari, 2019) archetype model, this scoping review systematically examined how different gamification resources are implemented across diverse crowdsourcing archetypes and how these designs affect the participants' engagement, performance, and motivation. By combining the gamification taxonomy with the crowdsourcing archetype framework, the review tells us what elements have been used and also why they are used, taking into account the task structure and value-creation logic of each archetype.

This scoping review mapped the gamification elements onto four categories (crowd-processing approach, crowd-rating, crowd-creating, and crowd-solving), revealing that most dominant design trends, showing the prevalence of achievement mechanisms, and showing the relative dearth of action or socially driven strategies.

5.5.2 Gaps, Problems, and Recommendations

1. Problem: Dominance of competition

Across all archetypes, competitive gamification mechanisms, such as leaderboard points and badges considered dominate the current gamified crowdsourcing systems. For example, research by Malatini et al. (2017) and Rokicki et al. (2014)] into crowd-processing, as well as Morschheuser et al. (2019) and Sadeghianasl et al. (2021) in crowd-rating, illustrates the utilisation of these competitive factors to enhance productivity and efficiency. This has also been used in other archetypes, for example, in crowd-creation, researchers such as Gamage et al. (2020, Prandi et al. (2018) which employed competition elements to stimulate innovation, whereas in crowd-solving, the studies by Cooper et al. (2010)) and Rojas et al. (2017) highlighted the widespread dependency of those competitive mechanisms to encourage participation to address the complexity of problem-solving issues. While these features motivate productivity, they often promote individual performance over collaborative effort, potentially discouraging cooperation and creating stress or unhealthy comparison among the participants. This is in line with the archetype-level patterns presented

previously in section 5.4.6, in which achievement-oriented elements emerge as the default design option in crowd-processing, crowd-rating, and crowd-creating. From the perspective of a SDT, the reliance on competitive extrinsic rewards may affect intrinsic motivation by constraining autonomy and limit the relatedness among the participants. Flow Theory similarly suggested that excessive competitiveness may disrupt the balance between the challenge and the skill, which leads to anxiety rather than enhancing engagement.

Recommendation: Collaboration-Oriented Functionalities

- Future studies should build hybrid models that balance the motivational competition advantages with the benefits of collaboration. Examples include team-oriented leaderboard, cooperative missions, or dual-reward systems that acknowledge both the individual effort and collaborative achievement. Such hybrid design models better supported SDT needs, specifically relatedness and competence, while also fostering flow by creating better balancing and socially a more engaging task environment.
- Designers or requesters should also contextualise competitive elements: adapting its intensity to the nature of the task and ensuring competitive features complement intrinsic motivation and cooperation.

2. Problem: Limited Utilisation of Social and Collaborative Mechanisms

The review showed revealed a limited and consistent under-implementation of social GDE, such as chat, peer feedback, or shared goal, across most gamified systems. For example, Table 5.3 shows that only a few crowd-processing research projects, including study of Carlier et al. (2015), expressly employed team or cooperative frameworks. In the realms of crowd-solving Cooper et al. (2010) and crowd-creation Ogawa et al. (2020), where collaboration is frequently vital, the majority of research centred on individual competition. Although these features were valuable for problem-solving tasks, they were rarely integrated even where teamwork was crucial. This presented a significant gap that limits the ability to leverage collective intelligence, exchange of the knowledge, and social relatedness, which are the most important components of motivation and engagement. In particular, SDT defined relatedness as an essential psychological need, and the absence of such needs may reduce the engagement level directly, while Flow Theory indicates that social interaction between participants could enhance immersion and attention throughout the task. Moreover, this gap indicated that many systems lack structured coordination support (e.g., shared progress visibility, feedback loops that help teams converge, and mechanisms that make collaboration performance-relevant rather than optional).

Recommendation

- Gamified crowdsourcing should integrate communication and cooperating elements that enable real-time interaction communication and peer support. For example, Martelaro et al. (2021) asserts that chat functions enhance user experience and facilitate effective collaboration.

- Future research should also empirically evaluate purely competitive systems with hybrid cooperative design, to evaluate how the systematic structure of collaboration affects outcome quality and engagement.

These design orientations align with SDT by establishing strong social connectedness and intrinsic motivation, while also supporting flow through interactive and responsive engagement structures.

3. Problem: Collective action problem

The recent issue in team-based studies is the challenge of collaborative endeavours, which arise when certain individuals did not know their equitable contribution. Most systems focus on group-level performance metrics without considering of individual accountability. For example, studies that include a teams' mechanism, like Hong et al. (2013), focus on how well the team performs overall or how teams compete against each other, but they don't explain how to ensure everyone in the team contributes fairly. This constraint is especially pertinent to our research objectives, which aim to develop a collaborative image-labelling system that integrates team-oriented goals, individual progress metrics, and real-time feedback scores to reduce such phenomena by enhancing the visibility and reward of individual contributions. This issue is of special concern in the interpretation of team-based performance effects, as increases in team scores may obscure uneven participation unless individual contribution signals are visible and rewarded.

Recommendation:

- The designers should incorporate mechanisms that ensure individual accountability within collaborative work. For instance, including individual progress or feedback attributions, peer-feedback, proportional scoring that can reward effort equitably while preserving cooperation.
- Hybrid reward structures should be evaluated to identify the optimal balance between individual and team-based rewards as these will be most effective at encouraging cooperation and individual sense of their contribution to the team. This approach aligns the design principles of collaborative gamified labelling systems develop in this thesis, which explicitly combine individual and team-based scoring feedback in order to maintain balancing motivation and improving quality.

These recommendations directly align with this thesis design choice to combine team-level outcomes with individual progress cues so that collaboration remains motivating without removing accountability.

4. Problem: Data Quantity vs Quality

A major problem embedded in the design of many reviewed systems is the inherent tension between incentivising a high quantity of contributions rather than accuracy or reliability. The most common GDEs, such as points, badges, and leaderboards, often lead to surface-level of participation at the expense of quality (e.g., number of images labelled, tasks completed) by default. This may prompt the participants to complete the task quickly, resulting in poor data quality. This issue is pronounced in high-stakes fields such as medical image analysis (Duggan et al., 2024, Linares et al., 2019), or intricate data validation (Brovelli et al., 2018),

where precision is paramount above rapidity. The literature shows a gap in explicitly designing gamification mechanics that prioritise and reward quality rather than sheer quantity. This risk is most visible throughout the studies reviewed in that output metrics tend to be predominantly volume-based, reinforcing the need for quality-linked feedback and validation mechanisms rather than point-per-task scoring. This reflects a misalignment with the SDT, where more emphasis on external rewards may weaken the intrinsic motivation for accuracy, and with the Flow Theory, where the repetitive, low challenge could decrease the engagement and the absorption in the task.

Recommendations:

- The designer should integrate the implementation of quality-centric feedback loops, rather than rewarding only submission. For instance, a feedback loops system that gives bonus points for contributions that reward accuracy.
- Design a robust scoring mechanism: instead of simple "one point per task" models; future designs could implement scoring algorithms that weigh contributions based on difficulty, consensus with other players as seen in study of Pinto and Viana (2019), or agreement with a known set of gold standard data. For example, controlled progression systems that employ levels or challenges as quality gates so the participants should only be allowed to access higher-value tasks after achieving a defined accuracy threshold. This creates a natural incentive for high-quality work as a means of progression.
- Integrate expert validation or consensus bonuses also encourage workers to focus on precision rather than quantity.

These approaches support the development of the competence SDT and help to maintain the challenge and skill balance that is required for the flow, thus encouraging deeper and more meaningful participation.

5. Ethical Considerations in Gamified Crowdsourcing

Few studies addressed the ethical dimension consideration of the gamified system, including the issues of fair compensation, data privacy and possibility exploitation. The reviewed studies on gamification in crowdsourcing have revealed several ethical concerns and did not explicitly address them. One primary consideration is the possibility of exploitation. The use of Monetary Rewards such as in Matin et al. (2020) and Melenhorst et al. (2015), with highly competitive mechanics can create environments where crowd workers overperform for minimal financial compensation, blurring the line between playful engagement and digital piecework. The psychological impact of constant performance measurement and social comparison via individual leaderboards can also lead to stress, burnout, and a focus on quantity over quality can led to worker fatigue, stress or reduced autonomy. Furthermore, systems that collect user data for tasks like labelling personal photos or location tracking raise significant concerns about informed consent and data privacy. The review suggests that ethical considerations are often secondary to design choices (e.g., persistent comparison,

accelerated pacing, and competitive pressure) that can have direct implications for worker wellbeing and perceived fairness.

Recommendations:

- The platforms designer and researchers must adopt and promote fair compensation models, by providing clear information about the usage of the data and allowing one to withdraw without penalty. Ethical consideration design should focus on well-being, autonomy, and consent informed, includes investigating and establishing fair pay standards for gamified tasks and being transparent with users about expected earnings and effort. This ensures that engagement mechanisms enhance rather than exploit worker participation.

In conclusion, the results indicate the urgent need to develop modern gamified crowdsourcing systems with collaborative, ethical, and quality-oriented features. In addition, more studies should be conducted to identify the impact of various sets of gamification elements on individual and group outcomes. The designers also need to base their decisions on motivation principles such as SDT and the Flow model to design systems that will encourage people to play fair, transparently share their activity on projects, and protect the data from becoming corrupted. Overall, the gaps identified from this review indicate a consistent misalignment between recent gamified crowdsourcing designs and established motivational theories. In particular, many studies overemphasise extrinsic competition at the expense of intrinsic motivation (SDT) and also fail to keep the balanced challenge-skill for the flow, indicting the need for a more theoretically grounded design of the human-centred approach in future systems.

5.6 Conclusion

This chapter advances our understanding of how gamification has been applied within the field of crowdsourcing by systematically mapping and synthesising the finding from 146 empirical studies. Based on this scoping review, it was possible to identify trends in gamification, and the theoretical and methodological paradigms describing the gaps that exist in design practice across different crowdsourcing archetypes. Gamification reviews showed that crowdsourcing systems are diverse where design and implementation are concerned. Most studies, integrated achievement-based style mechanisms such as seeing points, getting badges and being listed on the leaderboard, that increased participation rates and task throughput. By contrast, collaboration and social design elements remained underrepresented, despite the existence of evidence that these features enhance both data quality and levels of engagement. The same applies to immersion-related features (e.g., storytelling, avatars, aesthetics) implemented infrequently, even though they interact in meaningfully ways with intrinsic motivation and user involvement. Taken together, this evidence suggests that the predominant design in the literature is achievement-focused and individually oriented, while theoretically useful supports for relatedness, coordination, and intrinsic engagement are less frequently designed. Our scoping review also showed that engagement, enjoyment, and motivation were often measured with ad hoc or

inconsistent instruments that limited the ability to compare studies. Few articles grounded the findings in research traditions like SDT and Flow Theory, indicating the need for integrating a broad theoretical foundation within future research on gamified crowdsourcing. This measurement gap supports a design-and-evaluation approach of using validated subjective instruments in conjunction with behavioural logs and quality-oriented performance measures instead of activity volume as the central indicator of engagement.

The review also shed light on how most designs are centred on competitive mechanisms and it also highlighted several key problems, including the perniciousness of collective action, the disjunction between quantity and quality in data, and ethical concerns about compensation and transparency. To address these challenges, the chapter outlined clear design principle recommendations, including using hybrid incentive structures which balance individual reward and team reward, quality driven feedback, and integrate social communicative means to facilitate cooperation and collective intelligence.

All the result of the scoping review with all insights addresses the RQ2 and provide the conceptual basis for the next phase of this thesis. This review synthesis also connects back to the empirical findings of Study 1 in Chapter 4. Study 1 showed that workers differed in their task-selection preferences and engagement experiences, with task interest ranked second only to monetary reward, while personality traits were associated more clearly with engagement and interest than with objective accuracy. These findings provided an empirical indication that workers may respond differently to motivational properties of a task. When considered together with the patterns identified in this scoping review, they strengthen the rationale for examining gamification as a configurable rather than uniform intervention. Consequently, this thesis approaches gamification as a configuration problem, emphasising how multiple elements can be combined and how their motivational value may vary according to task context and worker characteristics, rather than evaluating only the presence or absence of individual elements. This perspective subsequently informed the collaborative gamified image-labelling design presented in Chapters 6.

Thus, Chapter 6 incorporates these results into the actual design and analysis of an interactive image-labelling system, featuring a blend of cooperation and competition bonded by gamification to examine their combined effects on labelling accuracy, worker engagement, and agreement strength. By grounding the designing of the system in terms of the gaps identified in this scoping review using SDT and Flow theory, Chapter 6 operationalises these theoretical insights gained into experimental investigation.

Chapter 6 :Collaborative Gamified Image-labelling Task Design

6.1 Introduction

This chapter builds on the empirical findings and theoretical insights established in earlier chapters to present the design and the implementation of the collaborative gamified crowdsourcing image-labelling system developed for this thesis. Based on the motivational principles introduced in Chapter 2 and the systematic gaps highlighted in Chapter 5, this chapter provides a concrete, testable system design. In doing so, it advances transitions from conceptual and empirical foundations to the operationalisation of those insights within completely functional experimental platforms. Particularly, the chapter defines the protocol for Study 3, by detailing how the theoretical constructs and empirical gaps are translated into experimentally controlled system design.

Chapter 5 demonstrated that most previous studies have concentrated primarily on competitive and achievement-oriented mechanics with a very limited focus on collaboration, social interaction, or quality-based incentive structures. This imbalance exposed a notable deficiency in the existing literature: whereas competition may yield short-term productivity gains, it fails to reliably enhance engagement, boost relatedness, or improve accuracy in repeated or complex crowdsourcing tasks. The review also identified that collaborative resources, such as shared objectives, communication, and team-based feedback, remain unexplored, despite robust support from SDT and Flow Theory. This gap is relevant for task contexts where ambiguity resolution, social sense-making, and motivational are essential for ensuring high-quality annotations. These observations encourage the need to design a system that allows for the combined and isolation effects of collaboration and gamification to be empirically tested under controlled conditions. Based on these gaps, the present study proposes a collaborative gamified system that has been explicitly designed to evaluate the mediating effect of integrated competitive, cooperative, and quality design elements on worker performance and engagement. The framework design integrates shared goal structures, synchronous communication, level task progression, real-time accuracy feedback as motivational based on autonomy, competence, relatedness, and challenge–skill alignment. Collectively, these characteristics operationalise the theoretical mechanisms revealed in previous chapters and provide rigorous empirical assessment for how collaboration and gamification jointly affect crowd performance. The system functions as a between-subjects experimental platform, facilitating direct comparison of various configurations of collaboration and gamification. The chapter starts by explaining our design rationale on the basis of the gaps presented in Chapter 5 and the motivational frameworks discussed in Chapter 2 and presents the full-featured system comprising all collaborative and gamified components. The comparative experimental conditions, (Chat-Only, No-Elements Teams, and Individual), were constructed by systematically removing selected design features from the Full-Featured version to isolate their effects. This structure enables the study to test not only the value of each component but also how the combinations of

competition and cooperation affect the performance, and the engagement in a real-time collaborative task environment. Such controlled feature subtraction ensures that variations in performance across conditions can be causally attributed to the presence or absence of specific motivational and collaborative mechanisms. This chapter delineates the experimental circumstances, measuring tools, participant recruitment approach, and data processing plan employed to address the research questions established for Study 3. Figure 6.1 illustrates the conceptual framework for the system's design and execution, which directs the exposition of feature alignment, system architecture, and workflow logic in Section 6.3. This figure serves as the integrative roadmap linking theoretical constructs, design choices, and experimental conditions. The components presented in this chapter form the methodological basis for the empirical findings detailed in Chapter 7.

6.2 Study Objective

This section states the research objectives and research questions for Study 3 and describes how these are operationalised through the experimental design and the measurement framework described in this chapter. The aim of this chapter study is to translate the evidence from Chapter 5 of the scoping review into an actionable design and evaluation plan grounded in the integration of the SDT with Flow Theory to present the developed system in a format appropriate for empirical assessment in the subsequent chapter. Specifically, this chapter illustrates how theoretical constructs become operationalised in practice into real world design decisions that setup the collaborative image-labelling environment. This chapter operationalises theoretical and empirical insights into particular design rules that govern the collaborative image-labelling process and rationalises the presence of each feature. The chapter thereby demonstrates its traceability from theoretical motivation, design intention and then the behavioural outcomes by systematically translating each conceptual requirement into a functional interface component. The study transforms the review findings into actionable design requirements and demonstrates how each requirement is implemented through features that facilitate collaborative interaction (synchronous team chat), progression and feedback (five levels with immediate level-end feedback), and quality-aligned incentives (agreement, and honeypots points). These mechanisms directly address the gaps identified in Chapter 5 by embedding cooperation, structured progression, and accuracy-oriented incentives within the system logic.

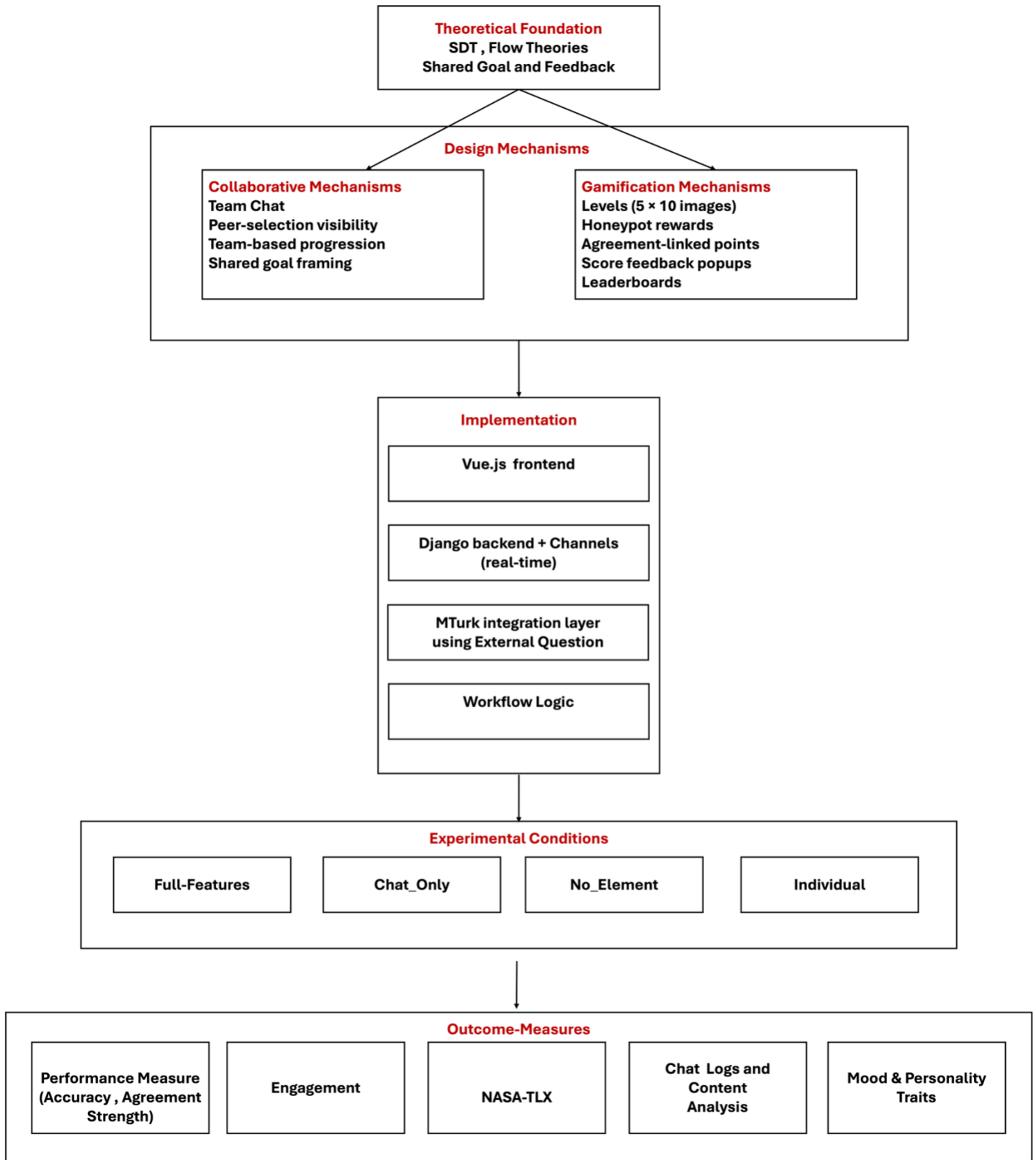


Figure 6.1: Conceptual framework for the design and implementation of the collaborative gamified image-labeling task

The study addresses the following research questions by using observable, measurable performance, engagement measures:

RQ3: How can such features be systematically designed and integrated to improve workers' performance and engagement in crowdsourcing image-labelling tasks?

RQ4: How do such features affect the reliability of the crowdsourcing outcomes, particularly in terms of worker agreement?

Based on the motivational features gaps that we identified, these wider questions are solved by answering the subsequent sub-research questions for Study 3:

SRQ3.1: ➔ (Performance effect)

What impacts individual accuracy, majority accuracy, and agreement strength are there with different team-based gamification arrangements compared with individual non-gamified work?

SRQ3.2: ➔ (Engagement effects)

In which direction do conditions of collaboration and gamification affect psychological engagement among workers (autonomy, competence, relatedness, enjoyment, and perceived usability)?

SRQ3.3: ➔ (Perceived Workload)

How do collaborative gamified conditions influence perceived cognitive workload as per NASA-TLX dimensions?

SRQ3.4: ➔ (Collaboration Mechanisms)

Can synchronous chat enhance labelling performance and what are the conditions where communication volume has direct associations with accuracy?

SRQ3.5: ➔ (Workers Differences)

Do personality traits or mood significantly moderate labelling performance across experimental conditions? In this chapter, we outline the system architecture, encompassing the front end, back end, real-time messaging, storage, and MTurk integration, to illustrate the requisite reliability and auditability for teams, messaging, and data integrity. The exposition effectively differentiates between the fundamental task-design logic and the particular technical implementation employed for its realisation, ensuring that the generalisable design principles are distinct from platform-specific engineering decisions. Moreover, we describe how the architecture supports synchronous interaction, mitigates coordination delays, and preserves data quality across conditions. Additionally, it delineates an ethical compensation strategy that incentivises quality. This includes transparent point-to-bonus conversion rules, minimum wage payment thresholds, and safeguards aligned with ethical crowdsourcing practice. We operationalise each sub-research question through specific metrics, validated self-report instruments, and behavioural interaction measures. Experimental conditions described in following Section 6.3 are explicitly designed to allow a comparison between such SRQs. The empirical findings for each SRQ are presented in Chapter 7.

6.3 Design Rationale

This section elaborates on the rationale underlying the experimental design in Study 3 and describes the creation of the experimental conditions employed to investigate the sub-research questions addressed in Section 6.2. A between-subjects design is employed for the purpose of experimental comparison of different configurations of collaboration and gamification mechanisms. The rationale of the design is structured around three components, which are: the research gap identification, the theoretical grounding and translation of these insights into the experimental conditions.

Building on the theoretical framework of Chapter 2 and the scoping review in Chapter 5, we provide the design rationale for the collaborative gamified image-labelling system employed in this research. The aim is to make a direct link between each design decision and the empirical limitations exposed in the literature, and to justify the structural design of the four experimental conditions, including the primary multiple features interface. This alignment guarantees that each implemented feature meets the target motivational-, collaborative-, or quality-related gap as clearly stated in previous chapters. The findings from the previous chapter identified several research gaps in the current crowdsourcing systems: over-dependence on competitive-based gamification (points and leaderboard features) (De et al., 2013, Riar et al., 2022), lack of application of cooperative or social approach elements (chat, shared goals, or peer collaboration) (Ogawa et al., 2020, Pinto and Viana, 2019), poor alignment between gamification design and human psychology motivation theory, SDT and Flow Theory (Csikszentmihalyi, 1997, Ryan and Deci, 2000), and general lack of quality-oriented feedback strategies that emphasise accuracy over speed (Feyisetan and Simperl, 2019, Hattie and Timperley, 2007). These gaps create a disconnect between worker motivations and system designs, which in turn impacts performance as well as engagement across most crowdsourcing archetypes. To address these gaps, the design is guided by three integrated considerations: empirical evidence from Chapter 5, the theoretical principles from Chapter 2, and quality-oriented design requirements that have been identified across the chapters. First, from Chapter 5, the Scopus review showed that gamification in crowdsourcing is dominated by achievement-oriented features, such as points, badges, leaderboards, and competitive scoring loops (Koivisto and Hamari, 2019), while collaborative elements, such as chat, shared goals, team-based feedback, or cooperative mechanics, are underexplored, even though their benefits have been shown in some archetypes (Ogawa et al., 2020, Tang et al., 2019). The majority of the studies lack balanced or hybrid designs and therefore limits our understanding of how the combining of competitive and cooperative features contribute to improved data quality. The Full-Featured system addresses this gap in empirical evidence by incorporating chat, shared progress visibility, and team-level decision into one coordinated workflow. Second, from Chapter 2, according to Ryan and Deci (2000) intrinsic motivation is enhanced when systems encourage: autonomy (choice, control), competence (specific objectives, feedback), and relatedness (social connection, shared experience). Flow Theory of Csikszentmihalyi (1997) emphasises also that deep engagement is promoted when

challenge aligns with defined goal, and feedback is immediate. However, most crowdsourcing systems rarely align interface features with these psychological states. Therefore, the Full-Featured system explicitly integrates elements that satisfy all three SDT needs and Flow Theory principles. Chat and shared visibility lead to relatedness. Points, levels, and diagnostic feedback encourage competence, avatar selection supports autonomy. Level progression and honeypot weighting produced challenge–skill alignment and immediate feedback (Locke and Latham, 1990, Sweetser and Wyeth, 2005). Third, the current designs often prioritise the quantity over the quality, resulting in fast but inconsistent labelling. Previous research indicates that rewarding alone may increase error rates, particularly in unclear image tasks (Brovelli et al., 2018, Kazai et al., 2012). Hence, this research includes the following quality-oriented mechanics, including weighted honeypots with increasing scores, agreement-linked points, level-end feedback summaries and peer-visible selections to encourage accuracy, consensus-building, and reflection. These design choices directly combat the “quantity-over-quality” bias experienced in current platforms.

Within this design rationale, the system is positioned as a hybrid crowdsourcing model. We position the current study as a crowdsourcing hybrid type of system with a crowd-processing orientation, supplemented by crowd-rating mechanisms and selective collaborative elements that partially overlap with those typically associated with crowd-creating. The task analysed, image labelling, consists of large volumes of homogeneous tasks that are undertaken by multiple workers, each contributor generating data that can be transformed into a usable output. At the level of performing a task, workers complete labelling tasks and the value of the system is measured from these collective contributions.

In addition, the study also includes important dimensions of crowd-rating with the use of its measures of quality and evaluation. In particular, the relevance of labels is determined by majority accuracy, and agreement strength measurements. Thus, while the task execution itself follows a crowd-processing model, the quality of the data depends on combining many judgments, a characteristic of crowd-rating systems. This means that the final value of the constructed dataset is not only built on individual submitting but also on group evaluations. Furthermore, with collaborative experimental conditions, we have added certain limited but meaningful features to crowd-creating. Workers work in teams with real-time data exchange and a common goal, and synchronous chat allows interaction and coordination to be key contributors to work and outcomes. In contrast, in canonical crowd-creating systems, participants are collectively building a shared and evolving artifact; however, contributions here remain limited by tasks, with no iterative co-construction of the artifact.

This hybrid positioning enables the study to investigate how competitive, cooperative, and quality mechanisms interact within the unified system. Accordingly, the Full-Featured system integrates competition (scores, leaderboards) and cooperation (chat, group-based progress, shared-visibility) on one hand and explicit structures aligned to SDT and Flow on the other. We adopt such a hybrid approach to fill the specific empirical gaps identified in Chapter 5, by integrating competitive, cooperative, and quality-driven components under

single evaluable environmental framework. In contrast, the rest of the conditions were designed by systematically eliminating one or more of these mechanisms in order to evaluate their separate and joint impacts. This comparative design supports causal inference by isolating features such as chat, visibility, feedback, and scoring mechanisms and extracts their individual contributions. The above insights guide the development of four experimental conditions, which are as follows:

The first condition is Full-Featured System (all gamification features + collaboration). Includes chat, shared workspace awareness, shared-visibility, levels, honeypots, score feedback, and leaderboards. This condition is optimised to satisfy SDT needs, enable Flow through structured progression and enhance social presence. The Full-Featured design puts the theoretical insights and the practical issues identified in Chapters 2 and 5 into practice, providing a robust and justifiable basis for the empirical analysis in the later sections. The design explicitly aligns with previous evidence relating to positive interdependence (Johnson and Johnson, 2005), goal-setting (Locke and Latham, 1990), consensus-building (Tang et al., 2019), and the support of motivational needs, (Deterding, 2015, Ryan and Deci, 2000).

The second condition of participation is chat-only collaboration. No gamification features, only collaboration (chat, shared visibility) tools. This method can isolate the social interaction effects, peer discussion and negotiated consensus (Tang et al., 2019), enabling evaluation of communication features alone without competitive reinforcement.

The third condition is no-gamification elements (collaboration without chat). Participants see only each other's selections (shared visibility). This tests whether minimal coordination improves agreement simply by being passively aware. This design mirrors similar "view-only" crowd workflows described in Geiger et al. (2011). This also serves as a minimal-collaboration baseline.

The fourth condition is the individual non-gamified condition. No collaboration, no gamification, pure baseline. This represents typical labelling tasks in previous labelling research (Snow et al., 2008). This condition provides the reference point against all collaborative and gamification improvement compared.

Collectively, these conditions enable a quantitative and qualitative assessment of cooperation, motivating design, and competitive resources on worker labelling accuracy, engagement, and agreement strength of the rater. This systematic feature subtraction approach allows us to isolate the effects from different design components and their combinations and facilitates direct assessment of SRQ3.1–SRQ3.4. By holding task content, interface structure, and compensation constant across conditions, observed differences in performance and engagement can be causally attributed to the presence or absence of certain collaborative and gamified features.

Collectively, these experimental conditions provide a foundation to operationalise the research questions in Study 3 through the control of the comparison of activities in individual, collaborative with partially gamified, and fully gamified tasks. The system implementation described in the following sections.

The experimental conditions and presence of collaborative and gamification in every configuration are summarised in Table 6.1. This table describes the controlled manipulating of the design elements across conditions and forms the structural basis for all comparative results covered in Chapter 7. This design rationale establishes a theoretically informed and experimentally controlled foundation for understanding and interpreting the results presented in Chapter 7. Specifically, the variations noted in performance, engagement, and consensus strength among different conditions are analysed in relation to the presence or absence of the design mechanisms that have been operationalised in this section.

6.4 Prototype Architectural Framework Components and Implementation

In this section, we describe the technical architecture and implementation of the collaborative gamified crowdsourcing system developed for this study. The description is given to enable reproducibility and transparency of the experimental platform.

This prototype system allows us to operationalise the motivational principles that were identified in Chapter 2 and the design gaps identified in Chapter 5, specifically the need for hybrid cooperative–competitive features, quality-oriented feedback loops, and theory (SDT, and Flow)-aligned task structures. Accordingly, the platform encompasses a set of modular gamification and collaboration resources that can be enabled or disabled to construct the experiment conditions.

Table 6.1: Features Design mapping across theoretical constructs, design rationale

Gamification Feature	Design Problem Addressed	SDT Need Satisfied	Flow Principle
Avatar & Nickname	Autonomy deficit in crowdsourcing; social anonymity reduces accountability and cooperation	Autonomy, Relatedness	Personalisation, social presence
Team Chat	Absence of peer communication prevents negotiated consensus and quality improvement. collaborative for contribution quality	Relatedness	Immediate feedback, social flow
Team answer visibility and shared workspace	Workers lack structural awareness of collective progress; passive social information reduces inter-worker variance	Relatedness	Social presence
Level Progression	Flat task structure induces monotony and quantity-over-quality behaviour; goal scaffolding needed	Between-teams competence	Challenge–skill balance, goal clarity
Honeypot Feedback	Inattentive workers obscured by group scores; data quality assurance required	Competence	Challenge–skill balance

Level-end Feedback	Speed-incentivising scoring increases error rates; consensus bonuses redirect incentives to quality	Between- teams Competence	Immediate feedback, mastery signal
Team- Leaderboard	Individual leaderboards discourage lower-ranked participants and undermine cooperation; team-level ranking preserves cooperation	Between- teams competence	Competitive feedback loop

This modular design allows accurate feature manipulation and controlled comparison across different conditions and aligns with the recommended practice for adaptive and user-oriented crowdsourcing systems (Law and Ahn, 2011, Morschheuser et al., 2016). At its core, the system is a real-time collaborative labelling environment, deployed through Amazon Mechanical Turk and supports synchronous interaction between the workers, controlled progression through levels, and continuous measuring of accuracy and agreement. Real-time coordination is highly recognised as a crucial element in collaborative human computation systems to achieve consensus-building and accuracy enhancement (Tang et al., 2019, Ho and Yin, 2018). Hence, the design of the platform was consciously created to reduce latency and maintain a uniform state across teams and support shared decision-making. The platform basically includes three main components: the MTurk integration layer, the front-end web-based labelling interface, and the Django-based back-end framework that support real-time collaboration. The MTurk integration layer within the back end is to oversee worker redirection, session binding, condition assignment, team building, scoring, and bonus disbursement, hence providing experimental control and data integrity. In addition, all participant events such as label submissions, timestamps, chat messages, and progression events were kept at the server level to maintain data completeness and avoid client-side manipulation. This logging approach allows accurate reconstruction of task sequences and interaction patterns during the analysis process. Also, this architecture was deliberately modular to accommodate specific collaborative and gamification components to have controlled activation and deactivation across experimental conditions, while maintaining a consistent task interface and data pipeline. Having established the architecture and implementation of the system, the following sections illustrate experimental procedure and participant flow, and measurement instruments employed to assess performance, engagement levels, workload, and collaboration outcomes across this platform. This distinction between primary system logic and feature toggles serves to guard against performance or engagement divergence as a function of motivational mechanisms, rather than differences in the presentation of tasks or workflow.

To promote research openness, reproducibility, and future enhancement, the collaborative gamified crowdsourcing system created in this thesis has been made publicly accessible via GitHub². The repository

² https://github.com/Banuqitah/Collaborative_Gamified_Task.git

includes the front-end and back-end implementations, experimental configuration files, and deployment documentation related to the research presented in Chapters 6 and 7.

6.5 Prototype frameworks

The construction of the prototype system required the merging of various technologies and contours to ensure real-time collaboration, work management, and data processing. The selection of the needed technologies has been determined by their scalability, efficiency, and adaptation to the norms in the study, following the best practice for scalable real-time HITL systems (Holovaty and Kaplan-Moss, 2009, Marani et al., 2019). Figure 6.2 offers an overview of the three-layer system architecture, showing how front-end interaction, back-end coordination, and hosting infrastructure help one another execute tasks in real-time by collaborating. The next paragraphs describe each layer in detail.

A. Front-end Layer: this layer was developed using Vue.js, due to its lightweight nature and its capabilities of reactive databinding, which allowed for real-time WebSocket updates to be displayed dynamically in the user interface (Hanchett and Listwon, 2018) such as avatar visibility, peer selections, chat messages, countdown timers, level transitions, and score and agreement notifications. Vue.js facilitates an interactive deployment. These are the attributes that make Vue.js fit particularly well for synchronous multi-worker tasks. This responsiveness also promotes SDT-aligned resources by providing competence feedback (through immediate scoring and level updates) and relatedness support (via visible peer activity) (Ryan and Deci, 2000).

B. Back-end Layer: this built using Django-based, due to its strong security features, swift development capabilities, and structured Object-Relational Mapping (ORM). Holovaty and Kaplan-Moss (2009), which facilitates integrated support for user authentication and database management, worker tracking, room assignment, image retrieval, scoring and honeypot evaluation.

This back-end structure ensures that task logic, scoring rules, and team coordination are consistently enforced across all experimental conditions, thereby preventing procedural variation from confounding performance or engagement outcomes. The system also enables synchronous collaboration via Django Channels, which extends Django to support WebSocket, enabling real-time chat, instant broadcast of selections, majority-vote triggers, collective countdown timers, and synchronised level progression. Real-time communication and synchronisation were critical to facilitate synchronous team-based interaction, enabling immediate coordination and feedback while ensuring that all participants in a team operated on the same task state.

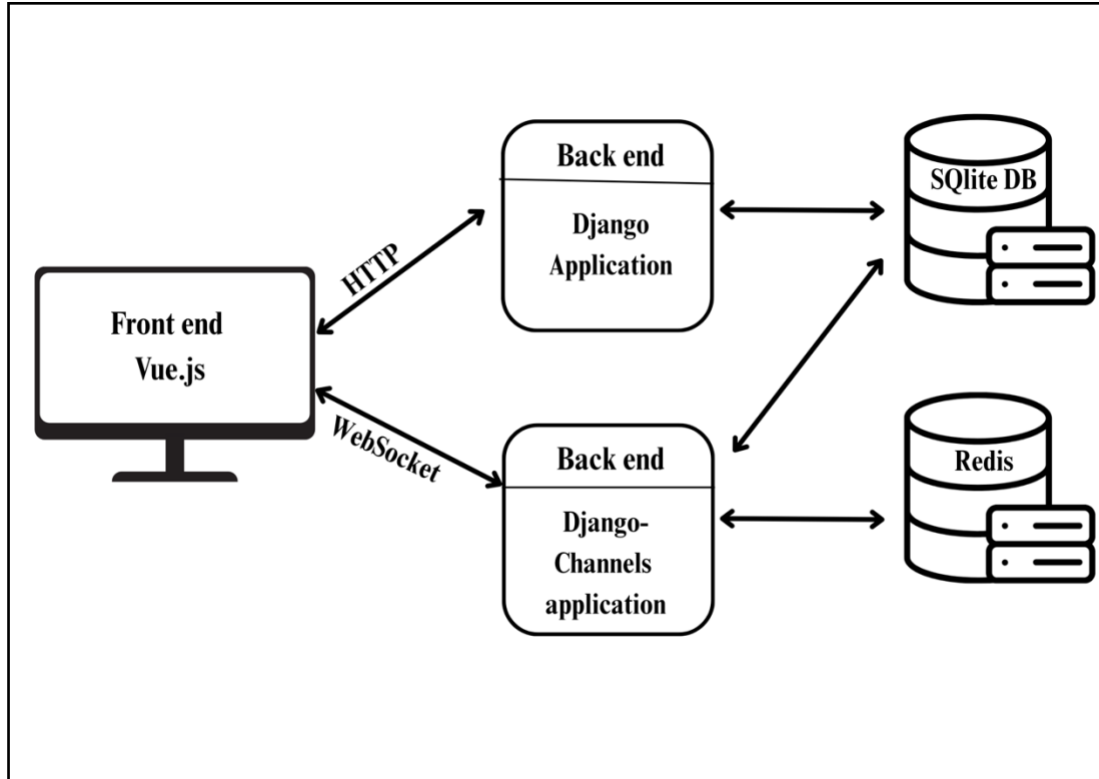


Figure 6.2: Prototype system architectural framework

This design decision directly supports the empirical evaluation of research questions related to collaboration in Study 3 (Tang et al., 2019, Ho and Yin, 2018). Django Channels provides long-running connections required for maintaining state consistency among team members, ensuring that all workers receive updates instantly as actions occur across the team (Marani et al., 2019). This functionality underpins the shared-goal, consensus-driven framework which are central to the study.

C. Hosting and Infrastructure: The system is hosted on Amazon Web Services (AWS) to ensure high availability, low-latency communication, secure handling of session data, and scalability for large-scale MTurk experiments. Cloud hosting also ensures stability under the variable arrival rates of the worker and supports the synchronous room-matching process required for real-time team formation.

6.6 MTurk Integration

Our prototype system employs MTurk External Question infrastructure for guiding qualified workers into the experimental platform. Every HIT generates a secure worker-specific session ID, which is essential for verifying consent, linking the survey task and bonus data, prevent multiple entries and maintaining the worker-to-team allocation integrity. Due to its dependability, assignment control, and integration flexibility, MTurk is still a widespread infrastructure for controlled crowdsourcing experiments (Buhrmester et al., 2016, Difallah et al., 2018). The Back end employs Boto3 SDK Python automation script (Appendix B), that enables programmatic creation, monitoring, and closing of the HITs. This workflow is guided by existing literature

that promotes the use of secure session-binding and in-depth back-end validation to implement the integration of custom experimental systems with MTurk (Chandler et al., 2014). This automation script handles HIT initialisation, qualification checks, worker arrival logging, session assignment, and database synchronisation, ensuring seamless transition from recruitment to participation and guaranteeing consistent execution across all experimental conditions. The MTurk integration layer guarantees that experimental manipulations for assignment control, session management, and condition allocation are centralised within the back end, avoiding any selection bias caused by worker self-selection or platform-level biases.

6.7 Workflow Logic in Real Time

This section presents the workflow of the Full-Featured experimental condition, which serves as the reference configuration for illustrating participant interaction with the system.

This walkthrough illustrates the participant’s complete experience. Once we run the script, the task is initiated in the Amazon Turk dashboard with a title, description, and initial reward (Figure 6.3) and appears to only qualified workers.

6.7.1 Consent and Instructions Screen

The workers select the tasks from the Amazon Turk dashboard as shown in the Figure, where they can read the description and click accept if they decide to participate. The consent form and brief instructions are the first page that participants arriving from MTurk will encounter (Figure 6.4). This page explains:

- The aim of the study
- The collaborative labelling task
- The role of levels, points, and honeypots
- The requirement to complete at least 20 images for base pay, with bonuses proportional to performance

All the instruction screenshots for all experimental conditions are included in the Appendix B. After the worker agrees on the consent form, worker is forwarded to the system website login page to fill out the registration form.

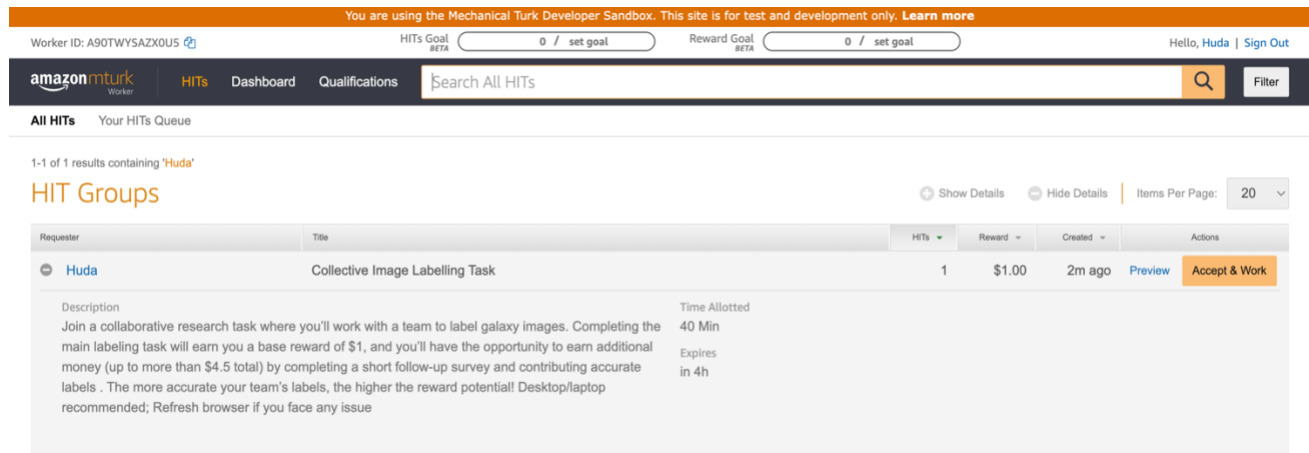


Figure 6.4: HITs Dashboard in Amazon Mechanical Turk

Welcome To Collaborative Gamified Image Labeling

Task Tutorial
User consent

This research is being conducted by Huda Banuqitah (huda.banuqitah@strath.ac.uk) under the supervision of Dr Mark Dunlop (mark.dunlop@strath.ac.uk) and Dr Sotirios Terzis (sotirios.terzis@strath.ac.uk) – all of Computer and Information Sciences, University of Strathclyde.

I am a Ph.D. student at the University of Strathclyde (Scotland, UK), My research aims to improve the overall quality of crowdsourcing platforms by developing techniques that help to retain and improve workers engagement and performance in image labeling crowdsourcing. Your collaboration is highly appreciated, and your input is very important to us. Thank you for considering participating in our study. We have designed a collaborative game that aims to encourage worker engagement and performance, where you will work collaboratively anonymous (by Nickname) with a team of workers to label images of galaxies based on the number of arms in each image.

Your participation will contribute significantly to our research efforts in improving task design in crowdsourcing platforms. The task will involve labeling up to a total of 50 images. For your participation in this study, you will be compensated in accordance with common US minimum wage levels \$8.50/h.

To be eligible for payment, you are required to complete a minimum of 20 labels and fill in a follow-up survey which will take around 20 minutes. Your payment of the minimum tasks will be based on the minimum wage levels \$8.50/h for both the labeling and post survey tasks. However, we encourage you to label more images, as increasing the number of labels completed will result in higher compensation, especially if you collaborate live with your team to improve the accuracy of the labels. The more correct image labels, the higher your reward will be.

The research was granted ethical approval by the University of Strathclyde Computer and Information Sciences Ethics Committee. The collected data may be used in academic publications, but your identity will be protected. No personal, or identifiable information will be collected except for your worker's platform ID. An anonymous data of the tasks and post survey data set will be created once the study is complete by replacing worker IDs with randomly generated ones. Your contributions are invaluable to the success of our study, and we appreciate your time and effort in advance. If you have any questions or concerns, please feel free to contact us. If you have any questions/concerns, during or after the research, or wish to contact an independent person to whom any questions may be directed or further information may be sought from, please contact the Ethics Committee of the Department of Computer and Information Sciences, Livingstone Tower Richmond Street Glasgow G1 1XH email: ethics@cis.strath.ac.uk

I confirm that I have read, understand and agree with the above policy and procedure for enrollment in this task.

I have read the tutorial and agree to rules
Reject (quit without reward)

Figure 6.3: Consent form

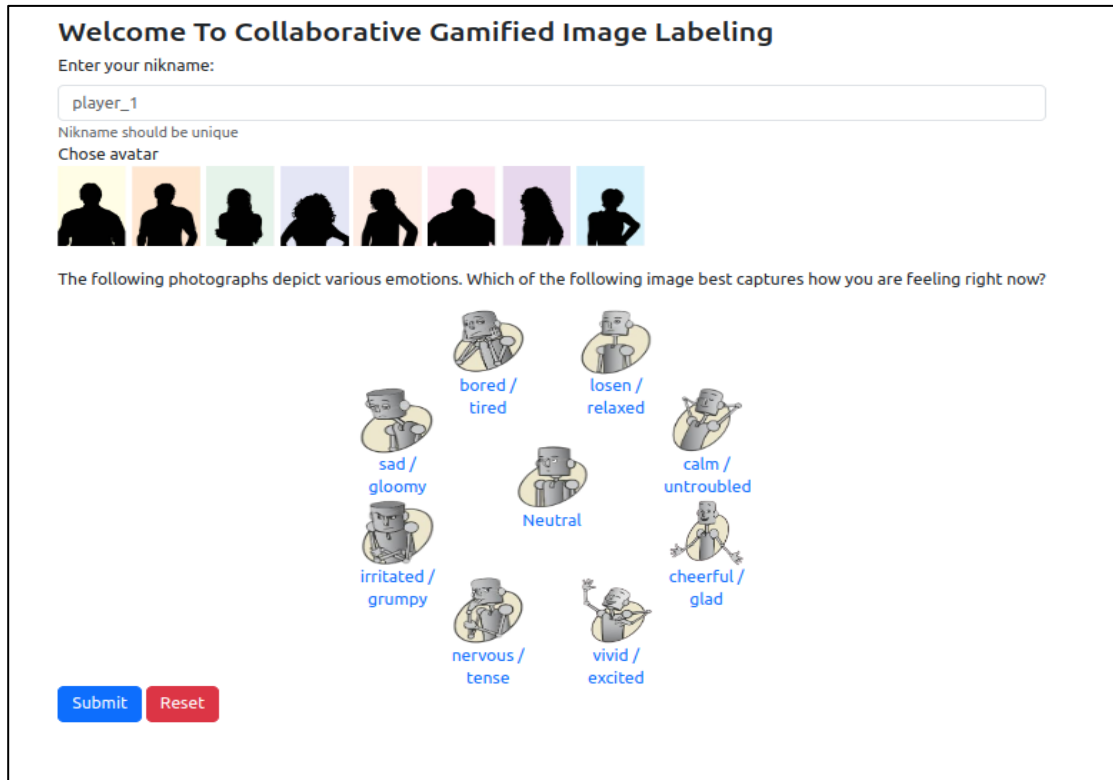
6.7.2 Login, Avatar, and Mood Selection Screen

On the login page, workers must register with the nickname, choose avatar, and the current mood (Figure 6.5). The selection of the avatar is consistent with SDT principles of autonomy and relatedness, that could create engagement before the main task interaction.

6.7.3 Waiting Room (Team Formation)

Each collaborating team is housed in a synchronous session container called a “room” in real-time-system parlance that provides a shared workspace to its participants for seeing collective progress and communication.

Following login, workers move into a synchronised waiting room. On that page the team formation starts and the team size ranges between two and five workers. Using a between-subjects design, participants were assigned into the experimental conditions, where each worker participated in only one condition to prevent learning and carryover effects. In the collaborative conditions, workers entering the system were temporarily placed into a pending queue until adequate participants were available for room allocation purposes. A Python-based script executed via a time-based job scheduler (Cron) runs every 15 seconds to assign workers to teams dynamically in order to preserve synchronization between team members, only those who joined rooms that had not progressed beyond the first image of the first level were added, or a new room would be automatically created. This was carried out in such a manner to reduce disturbances to current collaborative interactions and to ensure that workers in the same room experienced similar task progression and communication opportunities.



The registration page is titled "Welcome To Collaborative Gamified Image Labeling". It features a form for entering a nickname, with a text input field containing "player_1" and a message "Nickname should be unique". Below this is a section for choosing an avatar, displaying eight different silhouette icons. The main part of the page asks the user to select an emotion from a grid of ten options, each represented by a cartoon character and a label: "bored / tired", "loosen / relaxed", "sad / gloomy", "Neutral", "calm / untroubled", "irritated / grumpy", "nervous / tense", "cheerful / glad", and "vivid / excited". At the bottom left, there are two buttons: "Submit" (blue) and "Reset" (red).

Figure 6.5: Registration page

6.7.4 Main Full-Featured Task Page

The main interface (Figure 6.6), this condition implements the full gamified–collaborative design based on the identified gaps, theoretical background, and insights summarised in Chapters 2 and 5. This condition comprises multiple primary components which are: one Galaxy Zoo image at a time, predefined labels, all team members’ avatars next to their chosen labels, synchronous chat box, level progress indicators, plus a feedback popup message that will appear after each level completion. A one-minute countdown is triggered when $\geq 50\%$ of the team selects a label. Such a $\geq 50\%$ threshold is in line with consensus-based task workflows in allowing timely progress while preventing strategic delays (Snow et al., 2008, Tang et al., 2019).

The gamification resources are accumulation of points, agreement bonus honeypot bonus, constant visibility of team status, and increasing challenge across levels with chat communication. These elements combine for an all-around SDT and flow experience: clear goals, timely feedback, optimal challenge progression, relatedness via teamwork, and autonomy in decision making.

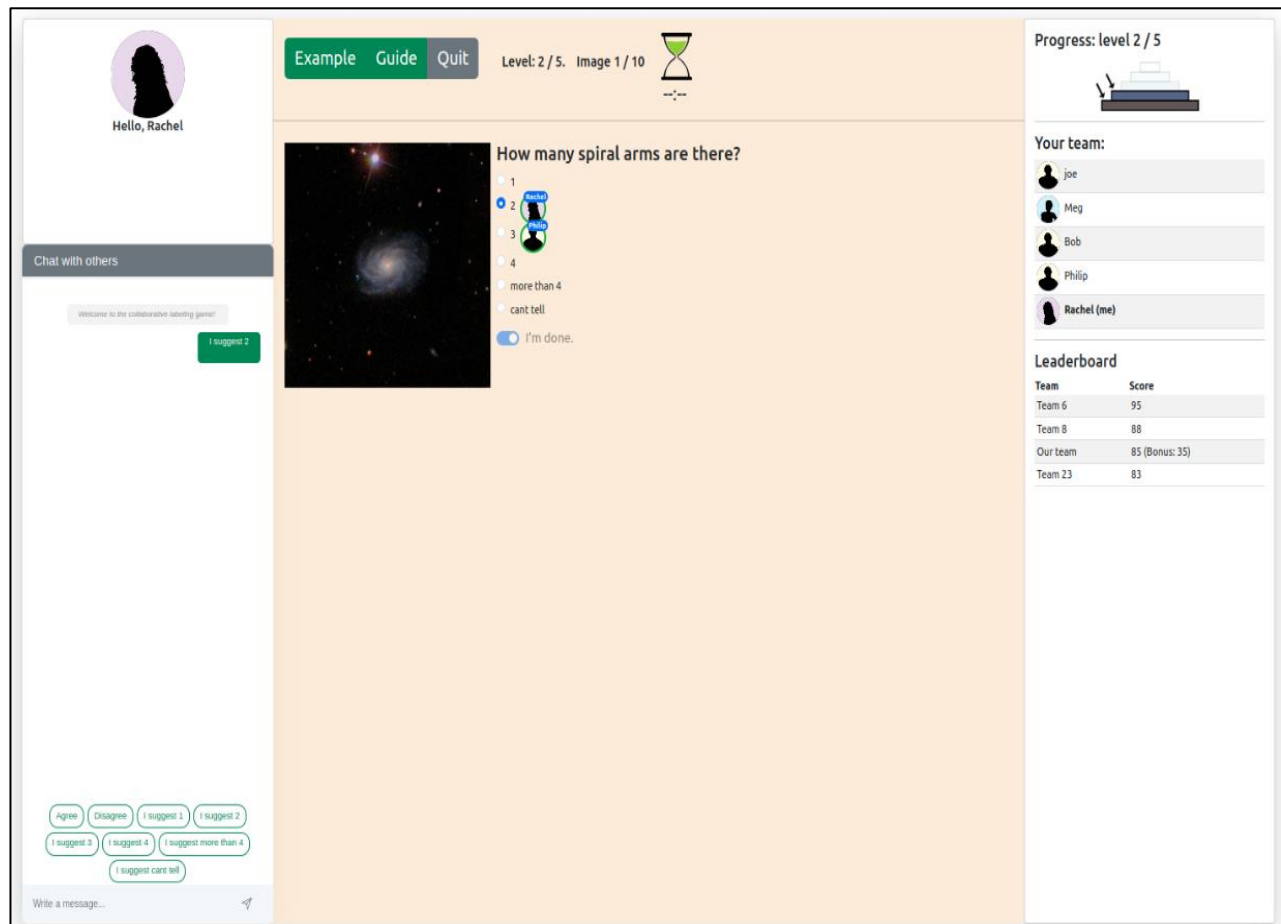


Figure 6.6: Full-Feature interface

6.7.5 Level-End Feedback:

Workers in the Full-Features condition get a 'popup message' whenever they finish each level, summarising their performance and contribution to the team score (Figure 6.7). The message contains the team total score; the individual contribution includes agreement points, honeypot points. This level-end diagnostic feedback is one of the key innovations of the system. It directly addresses the “quantity over quality” problem highlighted in the scoping review and reinforces competence and mastery.

6.7.6 Leaderboard (Between-Team)

The leaderboard shows the aggregate rankings of the teams. It promotes performance through moderate social comparison, avoids within-team competition to preserve cooperation, and aligns with task-level shared goals rather than individual rivalry. This well-designed interplay mitigates the harmful effects of competitive overload yet benefits from a well-known motivational advantage.

Participant compensation was established using a point-based reward system closely linked to task correctness and consensus, rather than direct monetary feedback during task performance. This solution guaranteed equitable compensation in accordance with US minimum wage standards while mitigating prominent financial signals that can influence worker behaviour. The comprehensive scoring and reward framework, encompassing agreement bonuses and level-dependent honeypot weighting, is explicitly delineated in Appendix B through simplified pseudocode to enhance transparency and reproducibility. This compensation design ensures that performance incentives are designed to support accuracy and consensus instead of speed alone to get paid, aligning with the quality-oriented evaluation criteria of this study.

6.7.7 Final Completion

Once Level 5 has been completed, participants can view final team score, their individual contribution, confirmation of bonus qualification, and link to the post-task survey with a completion code for MTurk (Figure 6.8). This confirms the end of the Full-Featured condition and moves workers toward post-study measurement.

6.8 Reduced-Feature Experimental Conditions

The next three sections illustrate how the reduced-feature conditions systematically eliminate or reformulate the features of the Full-Featured system to create experimental isolation for specific mechanisms. Each reduced-feature condition corresponds directly to sub-research questions by selectively disabling mechanisms present in the Full-Featured configuration (cooperation, communication, competition, and feedback).

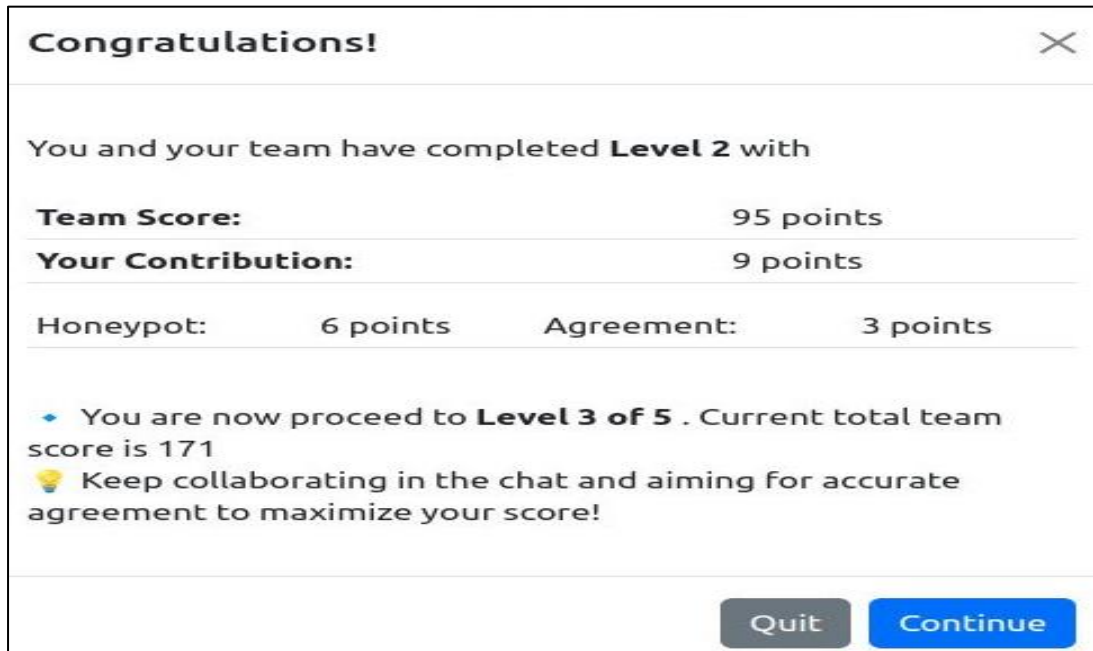


Figure 6.7: Level-End feedback message

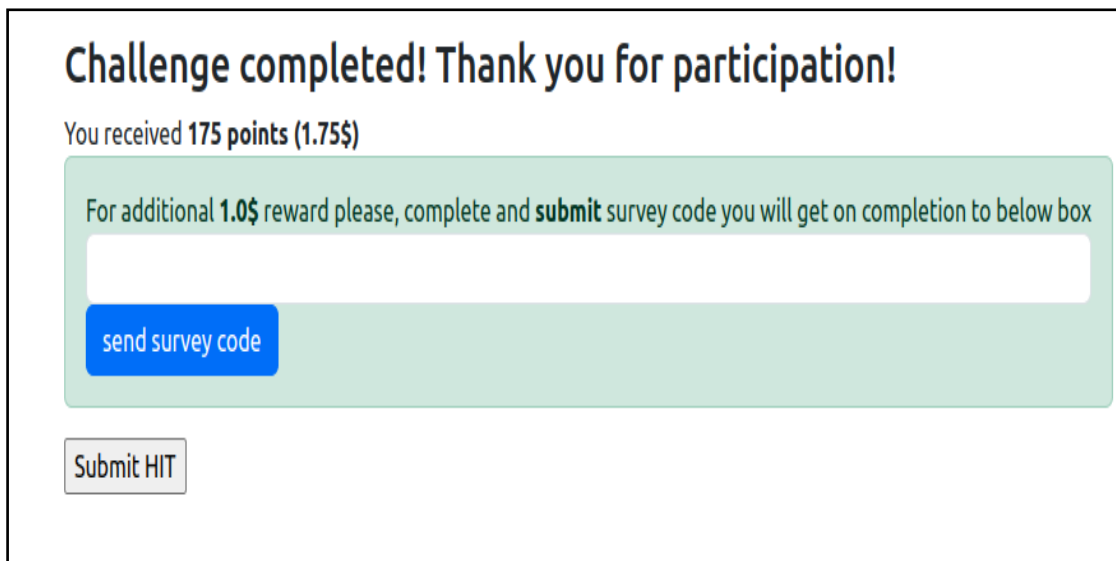


Figure 6.8: End of task message

6.8.1 Chat-Only Condition

In this experimental condition we include teamwork that offer visible peer selections, synchronous chat with no levels, no leaderboard and no level-feedback messages, Figure 6.9 show this condition interface. The only message in this condition is the final messages by the end on the task that show the rewards and survey link. The purpose of this condition is to isolate social interaction as a mechanism. It tests whether discussion and visible peer influence alone, without any gamification features.

6.8.2 No-Elements Condition (Collaboration Without Chat)

The feature included in this condition is the visibility of the peer selections only (Figure 6.10). The purpose of the condition is to test whether minimal team coordination (visibility of the other choices) is good enough to raise quality. This weak cooperative baseline isolates the value of explicit communication and feedback.

6.8.3 Individual Condition (Control)

This is an experiment control condition by which the workers accomplish the task by themselves as the fundamental labelling task with no gamification or collaboration. This condition serves as baseline to measure the effect of collaboration using chat, gamification, feedback, competition and quality-oriented incentives. The improvements in later conditions are counted in comparison to this control. Figure 6.11 shows the interface of individual control condition, which serves as a non-collaborative fundamental for subsequent experimental comparisons.

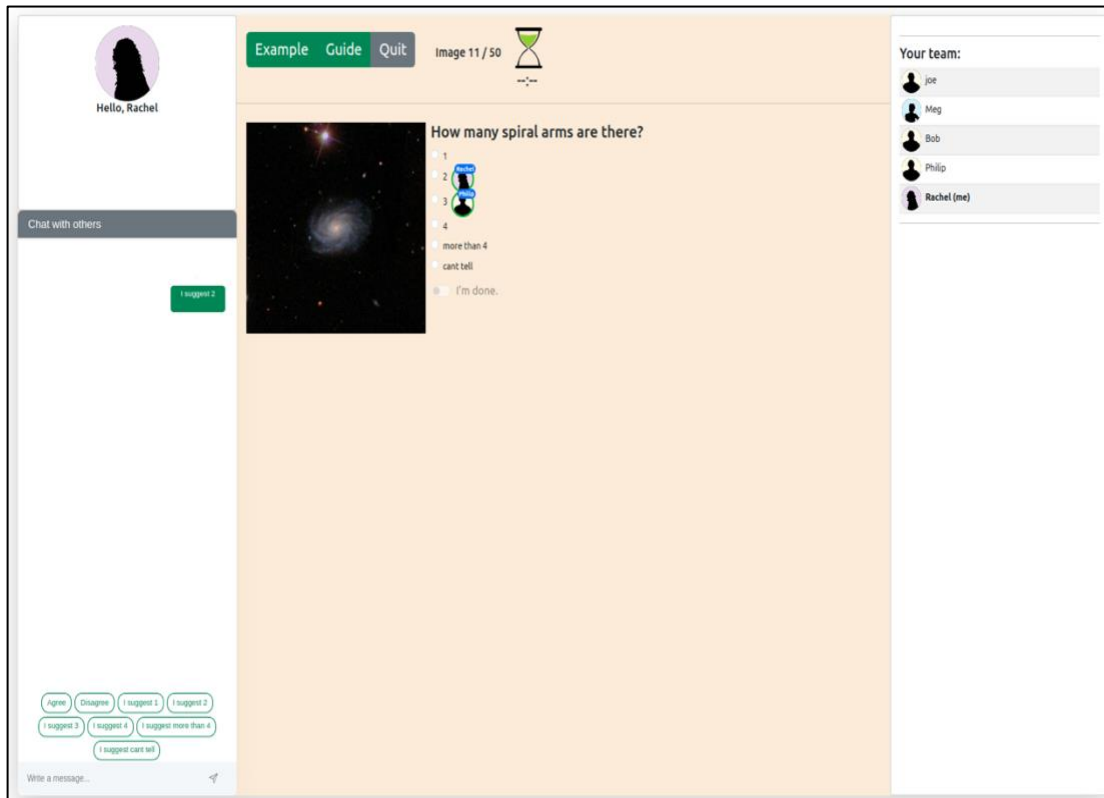


Figure 6.9: Chat-only condition interface

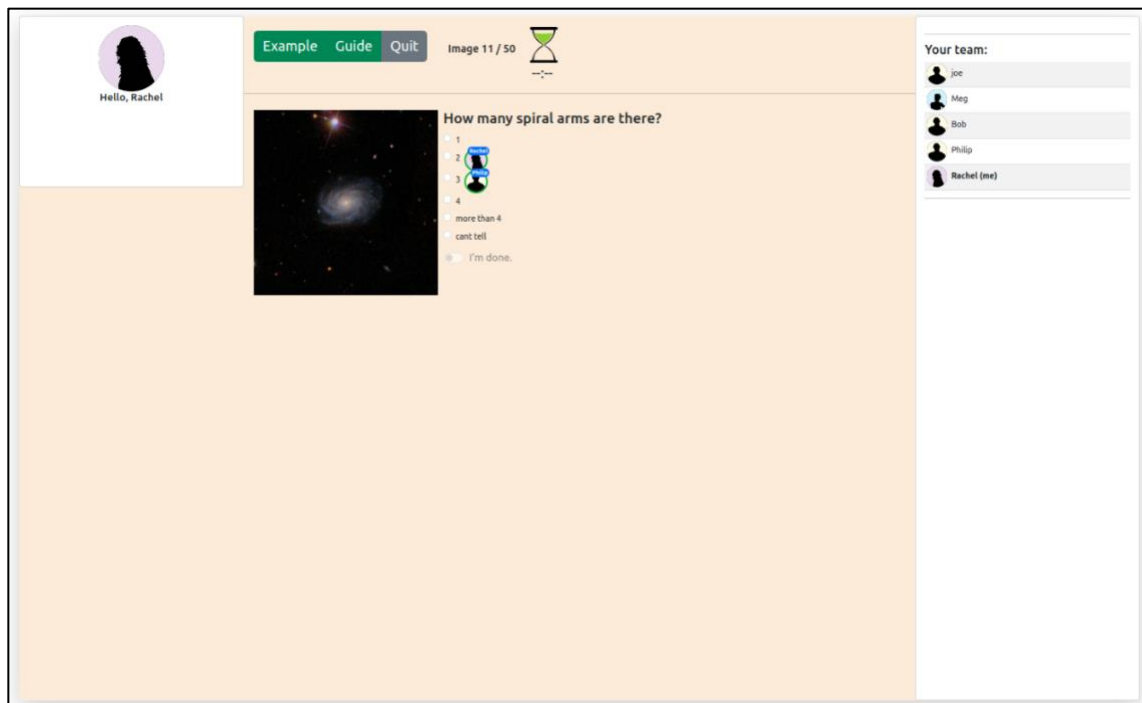


Figure 6.10: No-Elements condition interface

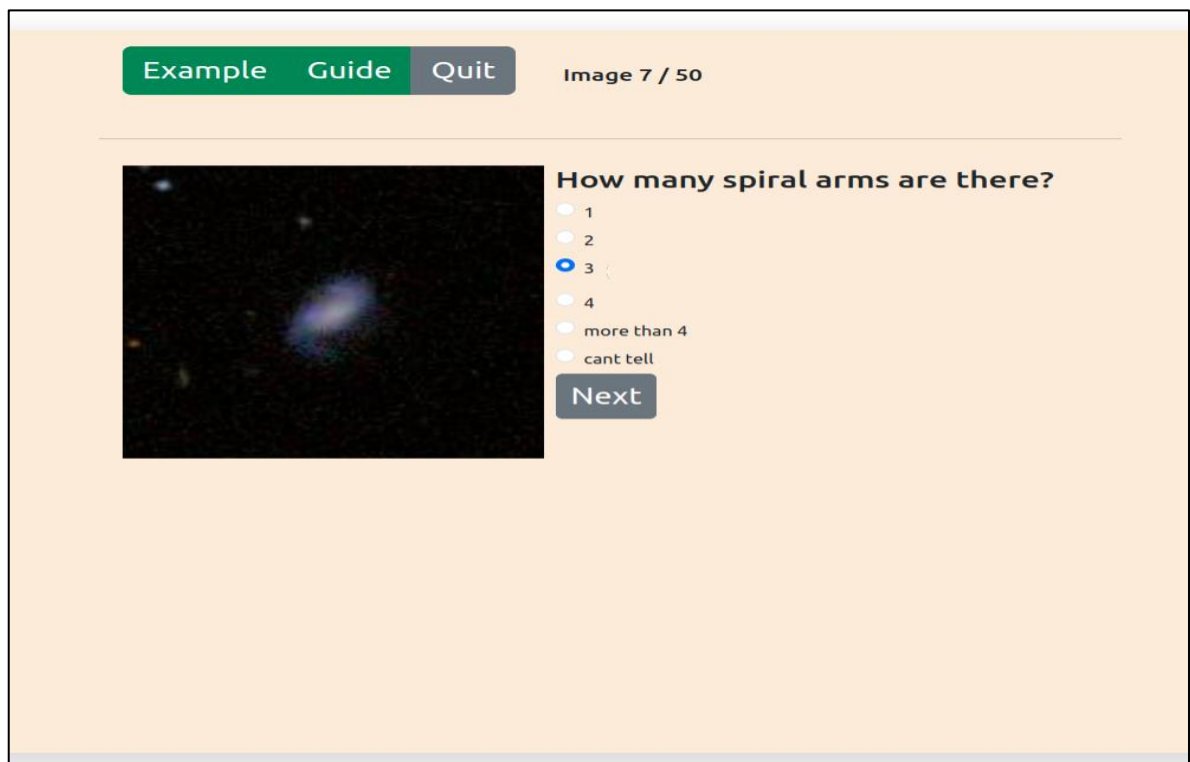


Figure 6.11: Individual control-condition interface

6.9 Dataset

This study used a selected sample of images from the Galaxy Zoo 2 project (Willett et al., 2013), which has been a popular large-scale citizen science data set for galaxy morphology classification (Lintott et al., 2008). Galaxy Zoo was chosen for the following reasons:

- It represents a canonical image-labelling task with well-validated labels
- It provides expert consensus ground truth, essential for accurate honeypot construction
- It contains images of consistent clarity and clear arm-structure visibility
- It is widely used in HCI, crowdsourcing, human-in-the-loop (HITL), and computer vision research.

For every image, participants answered a closed-ended question designed to guarantee simplicity and consistency: "How many spiral arms are there?" Among the accessible response choices were 1, 2, 3, 4, more than 4, and can't tell. For this experiment, 200 images were selected following a structured screening procedure. Crucially, the selection of images was intended to provide consistent task difficulty over all experimental degrees. The images were sampled from the "very clear, prototypical morphology" subset identified in prior work of Willett et al. (2013). By constraining the difficulty and visual clarity of the images in all conditions, the dataset design is sufficient to guarantee that differences in performance or engagement are accounted for by experimental manipulations as opposed to divergent task complexity. This ensures the stable difficulty across conditions, clear visibility of spiral structures, minimal variance due to noise, distortions, or ambiguity, and ensures consistent cognitive load irrespective of condition. Each worker labelled 5 levels $\times$ 10 images per level = 50 images, including one honeypot per level. The honeypots were made using classification frequencies found in the Galaxy Zoo 2 public repository to verify the reliability of expert-validated answers. The choice of this dataset also ensured strict control of image difficulty, and thus, the impact of collaboration, gamification, feedback, and social interaction on performance, devoid of any confounding by image complexity or dataset imbalance. This setup increases internal validity of the comparisons among conditions.

6.10 Pilot studies

In the first stage of the pilot work, we assessed the technical robustness, usability, and procedural feasibility of the four experimental settings before we began the full-scale study. Across two iterative pilots (each with ten MTurk workers per condition, restricted to US participants with $\geq 95\%$ approval rate and ≥ 1000 approved HITs), we assessed the stability of the room-assignment logic, the clarity of onboarding instructions, timing and synchronisation of chat and visibility-based interactions, and the functionality of the scoring and level-progression mechanisms. These pilots revealed several practical issues, notably in delays of team formation, occasional registration timeouts, and confusion regarding the transition between consent, login, and the waiting room. In order to avoid this result, we optimised the room-matching script to automatically assign late workers to existing teams when still at Level 1, shortened the instruction text, modified the waiting room countdown,

and upgraded the back-end logging to track session integrity. In open-ended pilot feedback, satisfaction showed good engagement, easy teamwork in chat-based contexts, and high interface usability. In general, this pilot phase verified that the system functioned properly across all four conditions (Full-Featured, Chat-Only, No-Elements, and Individual), confirmed that the reward structure was applicable, and that the experimental workflow was sufficiently stable and comprehensible to be deployed in the main study. The pilot study thereby proved not only a means of technical validation but also of procedural calibration, to ensure that the experimental protocol minimised friction, ambiguity, and unintended sources of variance before full-scale deployment.

6.11 Conclusion

In conclusion, this chapter has detailed the design, implementation of the collaborative gamified image-labelling task used in this study, demonstrating how each system's component operationalises the theoretical and empirical foundation established in Chapters 2 and 5. The chapter has concentrated on the translation of established motivational principles into a working, tangible, testable experimental system. The platform was purposefully designed to fill long-standing gaps in the literature, including the dominance of competitive gamification, the underuse of cooperative resources, the lack of SDT or flow aligned design, and the focus on accuracy-oriented feedback. As we design a shared-goal, positive interdependence, individually accountable system, we combine mechanisms supporting relatedness (visibility as part of the team, concurrent chat), competence (levels and agreement-based scores and honeypot testing), autonomy (avatar choice), while ensuring flow conditions via clear goals, immediate feedback and a measured challenge–skill progression. The Full-Featured condition balances competition and cooperation via levels, honeypots, team-level scores, and leaderboards, while the reduced-feature condition systematically eliminate collaboration or gamification features in the system to isolate their individual effects. So, this modular construction enables controlled comparison across collaborative, semi-collaborative, or individual baselines. This real-time workflow comprising team formation, synchronised decision stages, level-end feedback, and performance-based rewards, was intended to facilitate reflection, decrease repetitive errors, and strengthens consensus in relation to the majority accuracy without a loss of efficiency. The back-end instrumentation, MTurk integration, real-time communication guaranteed that the system performed reliably across all conditions.

Collectively, this chapter provides a methodologically rigorous implementation which makes clear how the causal links between gamification and collaboration are expected to affect performance and engagement. This establishes the platform, the experimental manipulations, and the technical precautions required for empirically valid research. Based on this methodological basis, Chapter 7 presents the experimental findings of the full-scale experiments and examines the impact of various combinations of collaboration and gamification configurations on accuracy, agreement strength, worker engagement, and perceive

Chapter 7 :Experimental Evaluation of the Collaborative Gamified Image-labelling Task

7.1 Introduction

This chapter discusses the empirical results of the collaborative gamified labelling study that we presented in Chapter 6. As mentioned, we employed four between-subjects experimental scenarios on Amazon Mechanical Turk, by which the participants were assigned to one of the four conditions that we discussed in Chapter 6: Full-Features (FF), Chat-Only (CO), No-Elements (NE), and Individual (IND). These conditions were built to systematically differentiate the presence of collaborative gamification mechanisms to investigate the effects of these mechanisms on worker performance, agreement, engagement, and perceptions of workload. Considering the methodological role of previous chapter, this chapter reports the empirical findings and deliberately refrains from any theoretical interpretation or design speculation. All analyses discussed here directly operationalise the experimental manipulations and outcome measures described in Chapter 6.

7.2 Overview of Analysis and Result Organisation

The results in this chapter are organised to directly address the sub-research questions defined for Study 3 in Chapter 6. Analyses proceed from objective performance outcomes to subjective and behavioural measures. Specifically, following a brief description of the participant characteristics and data-filtering procedures, the chapter first reports individual and team-level performance, including individual accuracy, majority accuracy, and agreement strength, followed by engagement and perceived workload. Communication behaviour and workers' feedback are then examined as complementary evidence concerning collaboration and worker experience.

To provide greater transparency and traceability between the sub-research questions of Study 3 and the empirical findings detailed in this chapter, Table 7.1 provides a structured overview mapping each SRQ to the corresponding outcome measures and results sections. This table serves as an informative resource, so that all information returned can be directly linked to the underlying research questions provided by Chapter 6. The chapter reports principal sub-research questions (performance, engagement, workload, and collaboration) as its focus of analysis, and analyses of individual differences (personality traits and mood) are explicitly considered as exploratory and supplementary to the core research objectives.

Table 7.1: Mapping of Sub-Research Questions to Measures and Result Sections

Sub-Research Question	Measures Used	Corresponding Results Sections
SRQ3.1 What impacts individual accuracy, majority accuracy, and agreement strength are there with different team-based gamification arrangements compared with individual non-gamified work?	Individual accuracy, majority accuracy, agreement strength	7.8.2
SRQ3.2: In which direction do conditions of collaboration and gamification affect psychological engagement among workers (autonomy, competence, relatedness, enjoyment, and perceived usability)?	Engagement based on (Autonomy, Competence, Relatedness, Enjoyment, Perceived Usability)	7.8.3
SRQ3.3: How do collaborative gamified conditions influence perceived cognitive workload as per NASA-TLX dimensions?	NASA-TLX Measures	7.8.3
SRQ3.4 Can synchronous chat enhance labelling performance and what are the conditions where communication volume has direct associations with accuracy?	Chat message and Feedback analysis	7.8.4 7.8.5
SRQ3.5: Do personality traits or mood significantly moderate labelling performance across experimental conditions?	Big Five personality traits, Mood	7.8.6

7.3 Experimental Conditions

The experimental conditions discussed in this chapter correspond exactly to those defined and justified in Chapter 6. Subjects were assigned to one of four conditions that varied systematically in the presence of collaboration and gamification features. For clarity during result interpretation, key features are listed in Table 7.2. Detailed descriptions of conditions design, rationale for feature specification, and deployment of features are presented in Chapter 6 and are not reported here.

Table 7.2: Experimental conditions and Gamification Feature

Condition	Teams	Chat	Levels	Score Feedback	Visibility of member selection	Leaderboard	Honeypots	Avatar
Full-Features	☐	☐	☐	☐	☐	☐	☐	☐
Chat-Only	☐	☐	☐	☐	☐	☐	☐	☐
No-Elements	☐	☐	☐	☐	☐	☐	☐	☐
Individuals	☐	☐	☐	☐	☐	☐	☐	☐

7.4 Sample Size Determination and Power Analysis

We conducted a prior power analysis utilising G*Power 3.1 (Faul et al., 2007). Given a four between-subjects experimental design with four separate conditions, we planned for an omnibus comparison across the condition (one-way ANOVA model). One-way ANOVA is appropriate for comparing means among many independent groups, rendering it the optimal method for assessing the impact of gamification components in collaborative crowdsourcing projects (Field, 2013). A priori power analysis was performed to guarantee the study possessed adequate power to identify a significant effect. This analysis has used an effect size (f) of 0.25, indicative of a medium effect size as delineated by Cohen (1992). The effect size was chosen based on the previous empirical studies on gamified crowdsourcing tasks, which have shown that medium to large effect sizes are found in assessing engagement and performance. For example, Mekler et al. (2017) discovered that points and leaderboards greatly enhanced performance metrics in online annotation tasks with medium-range effect sizes (Cohen's $d \approx 0.4$ – 0.6). Similarly, Eickhoff et al. (2012) showed a moderate effect of gamification on label quality and user motivation in the crowdsourced relevance assessment (Cohen's $d \approx 0.5$). These results support employing a medium effect size ($f = 0.25$) used in G*Power analysis and justify its usage in the power calculation of this study. The significance level $\alpha = .05$, and the statistical power $\beta = .80 - 1$, a standard criterion that guarantees an 80% likelihood of identifying real effects while reducing Type II error (Cohen, 2016). Figure 7.1 visualises the achieved power curve for $f = 0.25$ and $\alpha = .05$ across feasible sample size in this study.

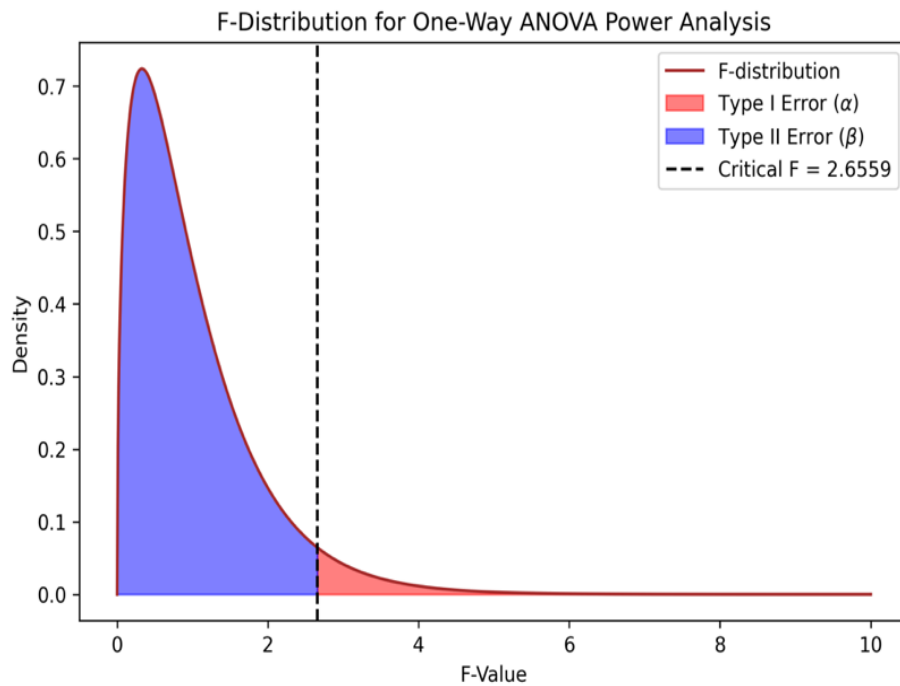


Figure 7.1: F-Distribution Power Analysis for One-Way ANOVA (Cohen, 1992)

Under these parameters, G*Power indicated an approximately $N \approx 180$ per condition (total ≈ 720) for 80% power to detect $f = 0.25$ in the omnibus test. To accommodate the real-time constraints in team formation, the final recruitment strategy was to recruit 180 participants in each team-based condition and 90 participants for the individual condition, resulting in a total sample of 630 workers with adequate statistical power. This ensured equitable team assignment while statistically maintaining the integrity of studies. The final sample size selection adheres to accepted power analysis principles in behavioural research as showed by Lakens, (2022) and guarantees adequate statistical sensitivity to identify moderate impacts of gamification features in crowdsourcing tasks. The meticulous calculation and justification of sample size bolster the credibility of the study's findings, aiding in the comprehensive comprehension of engagement and performance enhancements in collaborative gamified settings.

7.5 Participant Recruitment and Qualification Criteria

The MTurk platform has been chosen as the crowdsourcing platform to recruit the workers due to its widespread use in academic and industrial research concerning data labelling and task performance (Paolacci et al., 2010). MTurk provides a dependable and scalable framework for administering extensive studies with human participants (Difallah et al., 2018). The choice to utilise this platform corresponds with previous studies demonstrating MTurk efficacy in executing controlled experiments for assessing crowdsourcing task designs (Rand, 2012). Utilising MTurk, we intend to guarantee that the results of our study are relevant to a diverse array of scholars and practitioners within the domain of gamified crowdsourcing (Mason and Suri, 2012).

To ensure data quality and participant reliability across experiment conditions, we required only workers with a minimum approval rating of $\geq 95\%$ and ≥ 1000 approved HITs for participation. This approach conforms with advice from studies stressing the need of worker reputation measurements for reducing validity concerns related to MTurk samples (Aguinis et al., 2021, Litman et al., 2017). While the minimum number of approved HITs guarantees that participants have sufficient experience with the platform and task completion, so reducing possible errors due to unfamiliarity, this qualifications mechanism helps to filter out workers with a history of low-quality submissions (Arechar et al., 2018, Litman et al., 2017). A total of 180 workers were hired for each of the three team-based experimental conditions (Full-Features, Chat-Only, and No-Elements) and 90 workers were recruited for the individual condition to provide enough statistical power and flexibility in analysis. Team-based participation required flexible grouping mechanisms, so the decision to recruit 180 participants in the team-based condition was necessary due of the variability in team composition and group sizes and enabled us to create dynamic groupings of different sizes while preserving enough data for comparative study. 90 workers were hired under the individual condition to enable powered exploratory comparisons and balanced analysis against portions of the team-based data. We applied a pre-specified filtering: (1) early quit (who dropped out before completed 20 labelling tasks (2 levels); and (2) performance (≤ 0.26 average accuracy), to remove outlier, non-engaged or random selection responders, thus

reducing noise and ensuring data reliability, guaranteeing strong datasets, and control for potential attrition. The random worker selection filtering mechanism has been predicated on the presumption that workers regularly performing below chance level might not be completely engaged or might be arbitrarily selecting labels, hence adding noise into the outcome. This is also supported by the recent study of Saravanos et al. (2021), that found that even among the best MTurk workers, those with an approval rate of 98% or higher and more than 1,000 approved HITs—about 22.3%—failed at least one question check. This highlights the presence of negligence even among the top approved workers. The retained samples of workers after filtering were as follows: FF:121; CO:138; NE:112, and IND:68. These recruitment procedures ensured that all the analyses were applied on data from workers who completed a meaningful portion of the task and demonstrate baseline engagement, therefore supporting the validity of the statistical comparisons between experimental conditions.

7.6 Evaluation Measures

7.6.1 Average Accuracy Measurement

Individual accuracy was computed against gold-standard using the confusion matrices formulation. Workers predicted labels were compared to the labels in the gold standard; this measure was utilised to calculate their score as Equation 7.1. Accuracy is alternatively expressed as the ratio of correct predictions to errors, as in $\text{accuracy} = 1.0 - \text{error rate}$.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP + TN + FP + FN)} \quad \text{Equation 7.1}$$

where True Positive (TP) is the number of correctly labelled items as positive classes by workers, and True Negative (TN) is the number of correctly marked items as negative classes by workers. False Positive (FP) is the number of items incorrectly labelled as positive by workers, whereas False Negative (FN) is the number of items incorrectly labelled as negative by workers. Table 7.3 present the 2*2 confusion matrix used to compute the accuracy.

Table 7.3: Confusion matrix for positive and negative labels

		Predicted Labels of Workers	
		Positive	Negative
Gold Standard (original class)	Positive	TP	FN
	Negative	FP	TN

7.6.2 Majority Accuracy:

For each team, majority accuracy assesses the proportion of labels where the team's absolute-majority label matches the ground truth. It demonstrates the effectiveness of collective decision making. The following steps define the calculation process used in this study to compute team majority accuracy.

Step 1: Determine Majority Answer for Each Question.

For each team t and question q , let

- $A_{t,q} = \{a_1, a_2, \dots, a_n\}$, the list of all answers submitted by the team members.
- $N_t = |A_{t,q}|$, the number of workers in the team t who answered question q .

We define the majority label $M_{t,q}$ as the answer provided by a clear majority of the team members.

$$M_{t,q} = \text{mode}(A_{t,q}) \quad \text{Where:} \quad \frac{\text{count}(M_{t,q} \in A_{t,q})}{N_t} > \frac{1}{2}$$

Thus, the absolute majority for team t for question q is defined when the modal label exceeds 50% of the answers within the team on that item.

Step 2: Compare Majority Answer with Correct Answer

Let C_q be the ground-truth label for question q ; then the **correctness indicator** $C_{t,q}$ is defined as

$$C_{t,q} = \begin{cases} 1 & \text{if } M_{t,q} = C_q \text{ and } \frac{\text{count}(M_{t,q})}{N_t} > \frac{1}{2} \\ 0 & \text{Otherwise} \end{cases}$$

Step 3: Compute Majority Accuracy per Team

The majority accuracy of a team t is the average correctness across all questions:

Let Q_t be the set of all questions answered by team t with a valid absolute majority.

$$A_t = \frac{1}{|Q_t|} \sum_{q \in Q_t} C_{t,q}$$

$|Q_t|$ is the total number of questions where an absolute majority was achieved.

Step 4: Aggregate the Majority Accuracy By experimental condition.

The condition-level majority accuracy is the mean across all teams in condition c using Equation 7.2:

Let T_c be the set of all teams assigned to condition $c \in \{FF, CO, NE\}$.

$$A_c = \frac{1}{|T_c|} \sum_{t \in T_c} A_t \quad \text{Equation 7.2}$$

7.6.3 Worker's Agreement Strength

Agreement strength represents the average within-team proportion of workers who selected the correct label aggregated across all items. The following steps define the calculation process used in this study to compute team agreement strength.

Step 1: Compute Agreement Strength for Each Team

Let T_c be the set of teams in a condition C . Each team $t \in T_c$ includes a set of workers W_t and labels L_{wq}

$w \in W_t$: a worker in the team t

$q \in Q_t$: the set of questions this team labelled

L_{wq} : denotes the label that worker w assigned to question q .

C_q : correct label for question q

We compute the agreement strength per team level:

$$Agreement_t = \frac{1}{|Q_t|} \sum_{q \in Q_t} \left(\frac{\sum_{w \in W_t} 1(L_{wq} = C_q)}{|W_t|} \right)$$

Where:

- $1(\cdot)$ is an indicator function: 1 if $L_{wq} = C_q$ is correct
- The inner fraction measures the proportion workers labelled question q correctly.
- The outer average aggregates this across all questions answered by the team.

Step 2: Agreement Strength for all teams of the condition

The condition-level agreement is the mean of the team scores as shown in Equation 7.3:

$$Agreement_c = \frac{1}{|T_c|} \sum_{t \in T_c} Agreement_t \quad \text{Equation 7.3}$$

As we mentioned, the agreement strength reflects the strength of consensus with ground truth as opposed to inter-worker consistency, independent of ground truth. This differs from standard inter-annotator agreement metrics like Fleiss' κ , which measure worker-to-worker consistency (i.e. consistency irrespective of whether the label is correct). The two constructs are complementary: our measure captures collective accuracy and Fleiss' κ measures collective consistency. Because expert-validated ground truth labels were available for the dataset, accuracy-referenced agreement was prioritised as the more informative metric for the research aims.

7.6.4 Engagement Measurement

Engagement is defined as a multidimensional construct in human-computer interaction and motivational psychology, which incorporates affective, cognitive, and behavioural components of user experience (O'Brien and Toms, 2008). The engagement instrument comprised ten items measured using seven-point Likert scale and represented five dimensions: autonomy, competence, enjoyment, relatedness and perceived usability. We derived the sub-scales from widely used questionnaires (PENS, IMI, UES-SF) and adapted them to our labelling context in order to provide higher granularity. The subscales have been derived from validated instruments; each addresses a fundamental psychological need or usability principle that influences participation in digital contexts. Table 7.4 presents the complete engagement construct and the items that have been used and administrated to participants. The notions of autonomy, competence, and relatedness were assessed utilising subscales from the Player Experience of Need Satisfaction (PENS) questionnaire (Ryan et al., 2006), which is based on SDT. These scores evaluate the extent to which participants perceived themselves as autonomous, competent, and socially engaged during the task. The Enjoyment subscale was extracted from the Intrinsic Motivation Inventory (IMI), a validated tool utilised to evaluate interest and intrinsic value in digital environments (McAuley et al., 1989). An example of a reverse-coded item stated, “This activity failed to engage my attention”. Perceived usability was evaluated using items derived from the User Experience Questionnaire Short Form (UES-SF) (O’Brien et al., 2018), which measures subjective assessments of system efficiency and user-friendliness.

Table 7.4: Engagement Measures

Constructs	Questions		Source
Autonomy	Q1	I did things in the task because they interested me.	(Ryan et al., 2006)
	Q2	I felt controlled and pressured to be a certain way.	
Competence	Q3	I am satisfied with my performance in the task.	
	Q4	After doing the task for a while, I felt reasonably competent.	
Relatedness	Q5	When I visit the online community, I feel supported by other users.	
	Q6	When I visit the online community, I feel like other people care what I have to say and what I do.	
Enjoyment	Q7	I enjoyed this task very much.	(McAuley et al., 1989)
	Q8	This task didn't retain my attention.	
Perceived Usability	Q9	I felt frustrated while visiting this task site.	(O'Brien et al., 2018)
	Q10	I found this task site confusing to use.	

7.6.5 NASA Task-Load

A multi-dimensional scale is intended to help establish the estimation of the workload from one or more operators during task execution or immediately thereafter. NASA-TLX comprises six subscales that denote somewhat independent clusters of variables: Mental, Physical, and Temporal Demands, Frustration, Effort, and Performance (Hart, 2006). The premise is that a certain mix of these dimensions is likely to signify the "workload" encountered by the majority of individuals undertaking the tasks. Table 7.5 shows the six subscales of the NASA-TLX that assessed using a 7-point Likert scale.

7.6.6 Perceived Usefulness of Gamification Elements

In addition to the validated engagement, workload measures and team-based conditions, participants completed a small set of feature-specific questions that revealed their perceptions of the gamification and collaboration features within their assigned interface. These items were not collected in the Individual condition since the corresponding features (e.g., chat, levels, leaderboards, peer visibility) were not present. Because all participants in the Full-Features, Chat-Only, and No-Elements condition were intentionally exposed to features in the system, each condition contained only the corresponding items relevant to features it contained. All items were measured on a 7-point Likert scale and analysed descriptively (means and visual comparison only), since they were designed to add contextual rather than psychometric information and were not intended to support inferential testing. The complete questions of all feature-specific items, along with their condition-level availability, is provided in Appendix B.

7.6.7 Big Five Personality Traits and Mood

The Five-Factor Model of personality traits comprises Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. These five overarching categories contain a diverse array of human personality traits (Rammstedt and John, 2007). These measures often employ a Likert scale (e.g., 1-5, ranging from strongly disagree to strongly agree) to assess an individual's position on each feature. Mood is a pictorial tool for conveying and articulating emotions. It assesses eight unique mood states efficiently and intuitively, suitable for both qualitative and quantitative research (Desmet et al., 2016).

7.7 Evaluation Metrics and Analysis Plan

Our independent variable was the between-subjects condition, four nominal levels: Individual (IND), No-Elements (NE), Chat-only (CO), and Full-features (FF). Dependent variables comprised (i) accuracy measures (individual average accuracy, majority accuracy, and strength of Agreement; (ii) Engagement scale and NASA-TLX workload (Likert/ordinal); and (iii) chat messages.

Table 7.5: NASA-TLX Measures

NASA-Measure	Question
Mental Demanding	How mentally demanding was the task?
Physically Demanding	How physically demanding was the task?
Temporal Demand	How hurried or rushed was the pace of the task?
Performance	How successful were you in accomplishing what you were asked to do
Effort	How hard did you have to work to achieve your level of performance?
Frustration	How insecure, discouraged, stressed and annoyed were you?

Because these outcomes include proportions bounded in 0 and 1, ordinal scales, and skewed counts, we first assessed normality distributional assumptions with the Shapiro–Wilk test (Shapiro and Wilk, 1965). We used parametric tests (e.g., ANOVA to test the effect of gender on accuracy within condition. Otherwise, we applied non-parametric tests. For the main condition effect, including accuracy, majority accuracy, strength of agreement, engagement, workload, and chat, we employed the Kruskal–Wallis H test for omnibus comparisons and Mann-Whitney U for planned pairwise contrasts. Effect sizes were expressed as rank-biserial r (Kerby, 2014), which indicate the direction and the strength of the difference between two pairs conditions by evaluating the proportion of the pairs that favour the hypothesis. The positive correlation value ($r > 0$) means the value of the first condition tend to be higher than the second condition, negative value ($r < 0$) indicate that value of the second condition tend to higher than the first, while zero mean no effect or dominance.

The family-wise error rates were controlled solely for the RQ-relevant comparisons (Individual vs each team-based condition and specified within-team contrasts) via the Bonferroni correction. All statistical tests were chosen prior to implementation based on a data scale and distributional properties, and no arbitrary post hoc decisions were made beyond those described in Chapter 6.

7.8 Results

7.8.1 Participant Characteristics and Analytical Sample

Detailed demographic distributions are provided in Appendix B. As an additional check, gender was examined in relation to individual accuracy because gender composition differed across conditions. Although gender distribution differed significantly between conditions, no significant gender-related differences in accuracy were observed within conditions or across the overall sample.

7.8.2 Performance Measures

To analyse the effectiveness of the collaborative and gamified conditions, the majority of accuracy metrics were analysed across team-based experimental conditions. The analysis aimed to evaluate the extent to which gamification elements influenced participant performance in image-labelling tasks. Majority accuracy has been computed for team-based conditions (FF, CO, and NE) to investigate the degree of consensus in image-

labelling tasks. In the Full-Features condition, participants again demonstrated superior performance, achieving the highest majority accuracy ($M = 0.66$), followed by the Chat-Only ($M = 0.55$), and No-Elements conditions ($M = 0.50$). A Shapiro-Wilk test revealed that the majority of accuracy was normally distributed for all conditions. The distribution of majority accuracy is presented in Figure 7.2. An ANOVA test indicated statistically significant differences in majority accuracy differences between conditions ($F = 8.91, p < .001$). Furthermore, Tukey's post-hoc comparisons highlighted statistically significant differences between the FF condition and both Chat-Only ($p = .007$) and No-Elements conditions ($p < .001$), by small-to-moderate effect sizes (FF vs NE: $r_{rb} = -.45$; FF vs CO: $r_{rb} = -.36$). These findings indicated that the comprehensive gamification and collaboration mechanisms in the FF feature condition facilitated more reliable collective decision-making than the reduced feature conditions. The majority accuracy differences are directly related to SRQ3.1 because the experimental condition was quantified to determine to what extent it affected the consensus accuracy of the team level (collaborative settings).

Moreover, to further quantify agreement levels and the strength of the workers' consensus across the team-based conditions, the strength agreement measure identified in Section 7.6.3, was used. The analysis indicated that Full-Features workers exhibited the highest strength of agreement ($M = 0.68$), while the Chat-Only and No-Elements participants obtained lower scores by ($M = 0.59$) and ($M = 0.57$) respectively, this showed in Figure 7.3. Kruskal analysis confirmed these differences as statistically significant ($H = 14.37, p < .001$).

Pairwise Mann-Whitney comparisons reinforced these findings, showing higher significant differences for Full-Features vs Chat-Only with small-to-moderate effect ($U = 908.50, p = .006, r_{rb} = -.29$), and vs No-Elements ($U = 860.00, p < .001, r_{rb} = -.39$). No significant difference was detected between the Chat-Only and No-Elements conditions ($p = .084$). This observation underscores that real-time feedback and comprehensive gamification enhance rater consensus strength among collaborative teams. To conclude, the results of accuracy and agreement analyses support the hypothesis that comprehensive gamification (FF) robustly enhances outcome quality compared to minimal gamification or no gamification conditions.

To further examine the nature of collective decision-making across conditions, we analysed the proportion of questions which resulting in different majority outcomes. Particularly, five categories were examined: first: correct absolute majority (AM), second: incorrect absolute majority, third, evaluating the percentage of questions where there is a correct relative majority (RM) decision when no absolute majority; fourth, measure the rate of questions with incorrect relative majority, and finally the rate of no majority at all.

Table 7.6: Tukey's Post-Hoc and Rank Biserial Correlation Analysis for Majority Accuracy

Conditions	Statistic	p-value	Lower CI	Upper CI	Significant	r_rb
FF vs CO	0.115	.007	0.027	0.203	Yes	-.36
FF vs NE	0.156	.001	0.065	0.247	Yes	-.45
CO vs NE	0.041	.496	-0.045	0.127	No	-.11

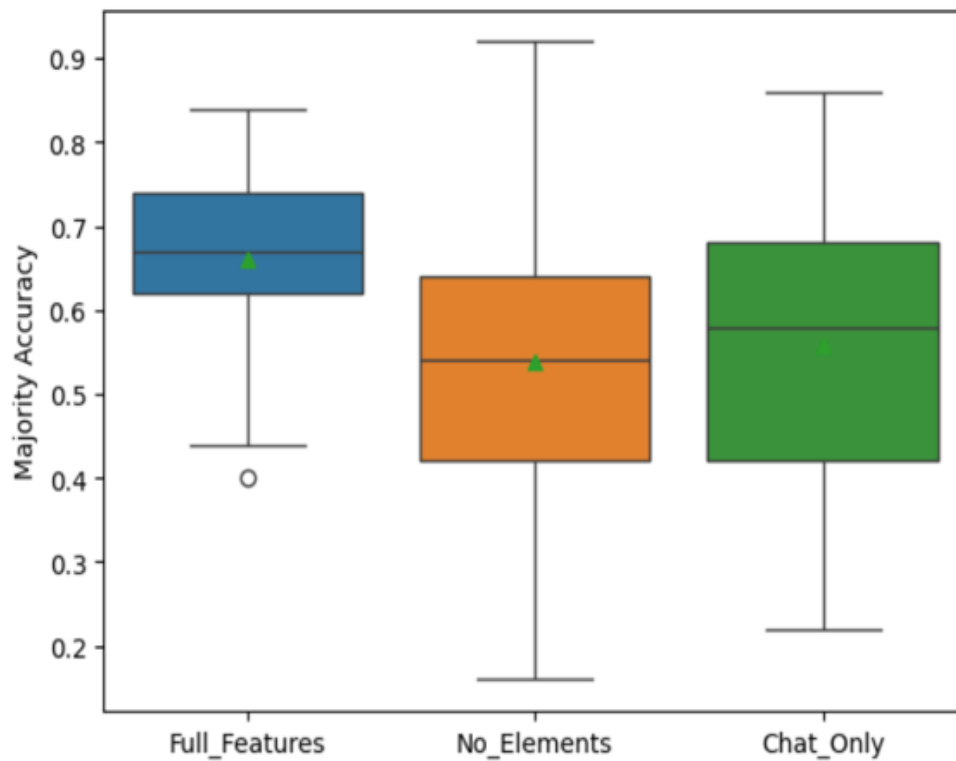


Figure 7.2: Boxplot of Majority Accuracy Across Conditions

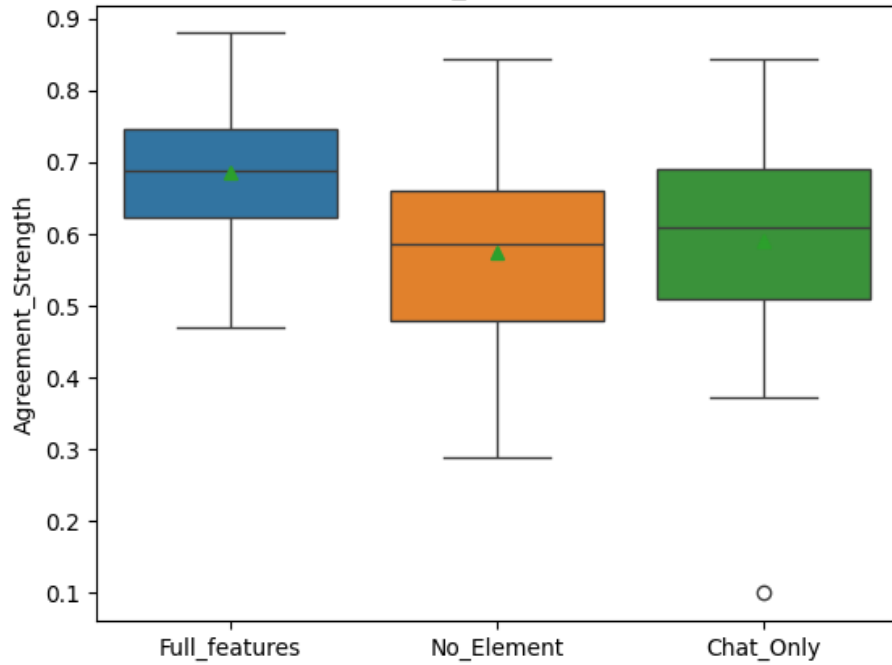


Figure 7.3: Agreement strength

Table 7.7 summarises the distribution of these outcomes across the experimental conditions. The FF condition has the highest proportion of correct absolute majority (0.66), the lowest ratio of questions with no majority (0.08) and incorrect absolute majority (0.23) compared to the CO and NE conditions. A Kruskal–Wallis test across the conditions showed that there is a significant difference for the correct absolute majority and no majority (Correct *AM*: $H = 17.00$, $p < .001$; No Majority: $H = 11.94$, $p = .003$). The pairwise test showed a significant difference of No Majority questions between the FF and the NE ($p < .001$ and $r_{rb} = .44$). No significant differences were observed for correct or incorrect relative majority outcomes (all $p > .44$) (see Table 7.8). This pattern suggests that the Full-Features condition not only increased correct consensus but also substantially reduced indecision at the team level. These majorities decisions also address SRQ3.1, by evaluating how each condition design change affects the strength of team consensus beyond accuracy alone. To compare Individual condition (control) to team-based conditions in term of performance, the average of individual accuracy within each condition has been calculated, as shown in Figure 7.4, revealed that participants in the Full-Features (FF) attained the highest mean accuracy ($M = 0.69$), succeeding over both the CO ($M = 0.60$) and Individual (IND) conditions ($M = 0.61$), but the (NE) condition recorded the lowest mean

Table 7.7: Ratio of different majorities analysis:

Condition	Correct AM	Incorrect AM	Correct RM	Incorrect RM	No Majority
CO	0.55	0.27	0.01	0.03	0.15
FF	0.66	0.23	0.01	0.02	0.08
NE	0.50	0.25	0.01	0.02	0.21

Table 7.8: Statistical Analysis of Majority-Decision Differences Across Conditions

Measure	H-statistic:	p-value
Correct AM	17.00	< .001
Correct RM	0.33	.846
No majority	11.94	.003
Incorrect AM	1.60	.450
Incorrect RM	0.27	.872

accuracy ($M = 0.59$). Histograms of accuracy distributions in Figures 7.5 -7.8 showed that the accuracy data for each team-based condition appeared to be approximately normal, supported by formal tests of normality (p-values: FF = .127, CO = .060, NE = .501). However, accuracy scores for the individual conditions deviated significantly from normality ($p < .001$). Thus, a Kruskal-Wallis test was used; the test confirmed a significant difference in individual accuracy among conditions ($H = 17.49$, $p < .001$). Subsequent pairwise Mann-Whitney U tests revealed significant differences specifically between the Full-Features and individual conditions ($U = 5,027.00$, $p = .006$), with a small-to-moderate negative effect size ($r_{rb} = -.25$), indicating that the Full-Features condition significantly improved accuracy compared to individual task conditions. No significant differences were found between Individual and team-based conditions (CO and NE). This confirmed that the improvements on the average of individuals accuracy were mainly associated with the full integrated collaboration richer gamified design rather than collaboration alone. This individual-accuracy comparison meets SRQ3.1, because it examines whether condition differences span from team consensus measures to individual performance outcomes across all four experimental conditions.

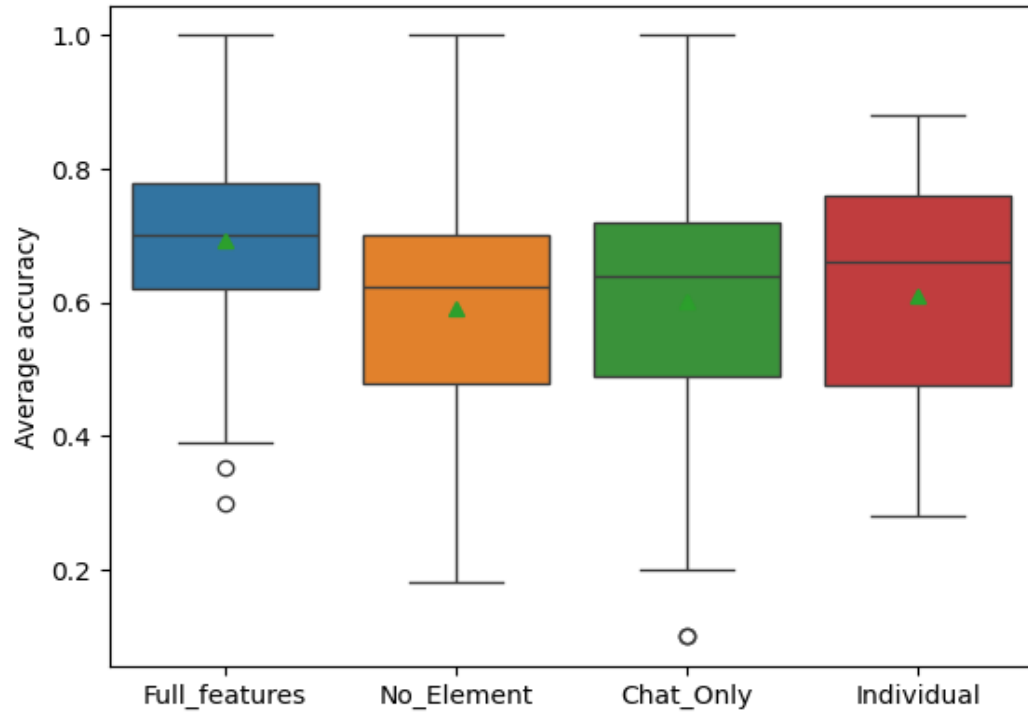


Figure 7.4: Average accuracy of workers across conditions

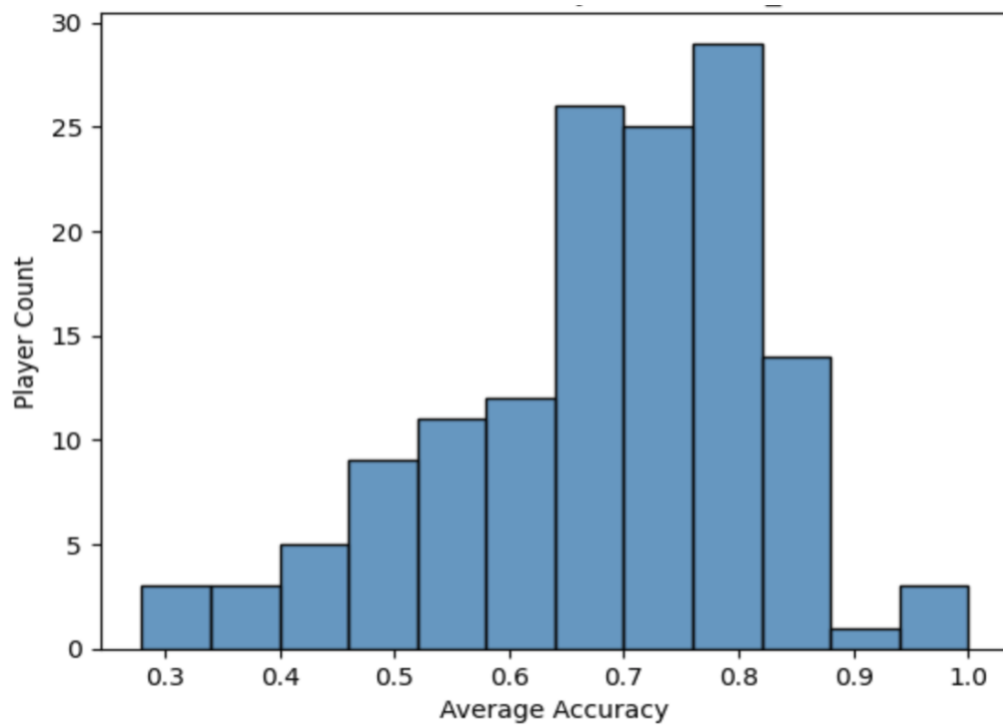


Figure 7.5: Histograms of average accuracy in Full-Features experimental conditions

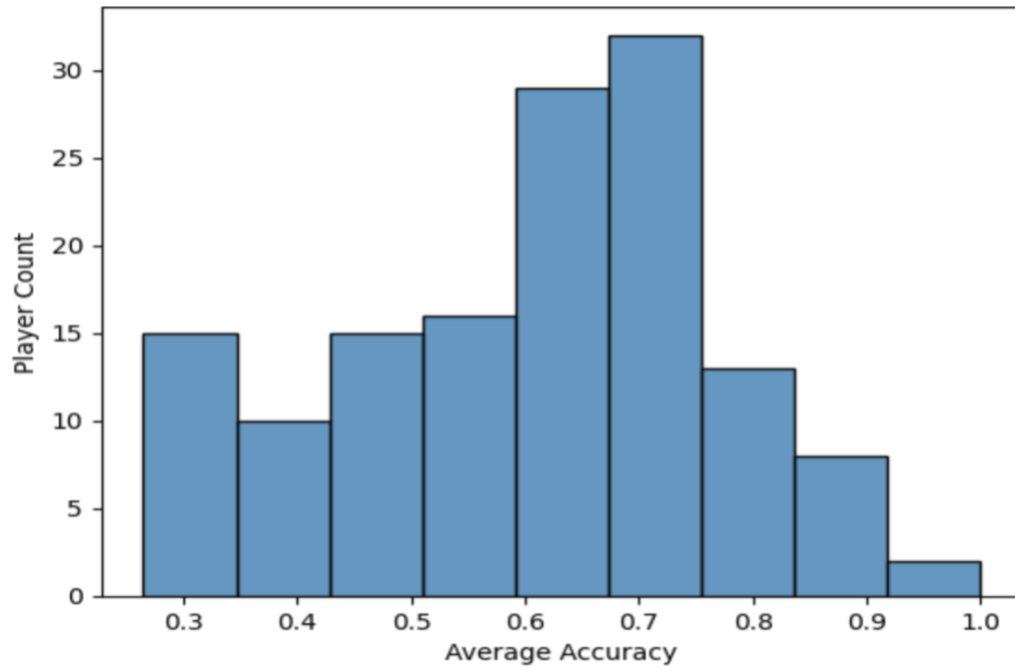


Figure 7.6: Histograms of average accuracy in Chat-Only experimental conditions

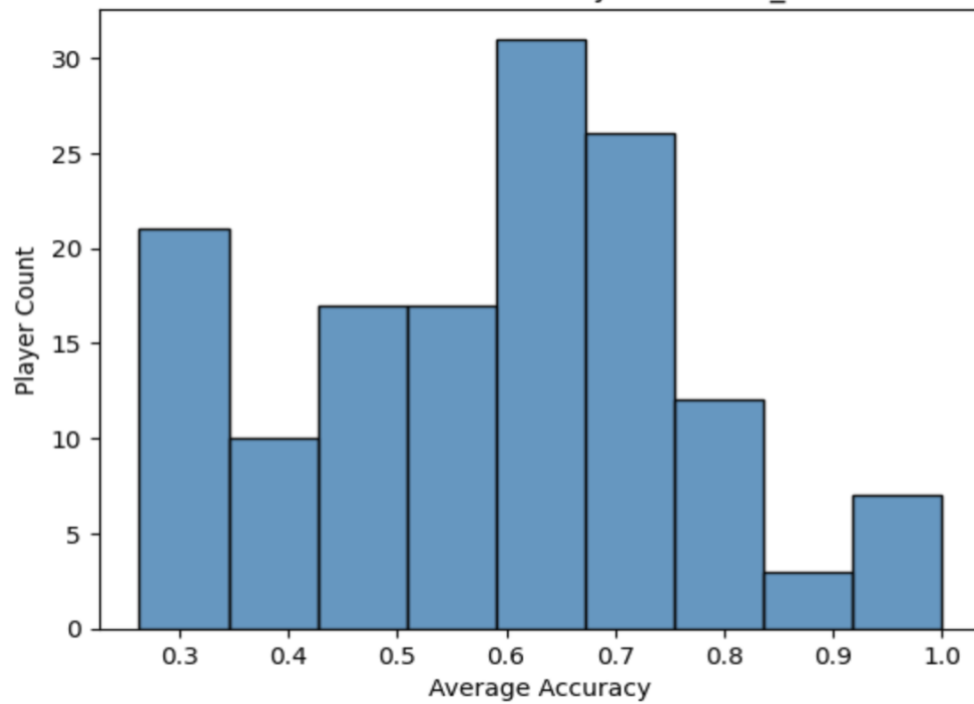


Figure 7.7: Histograms of average accuracy in No-Elements experimental conditions

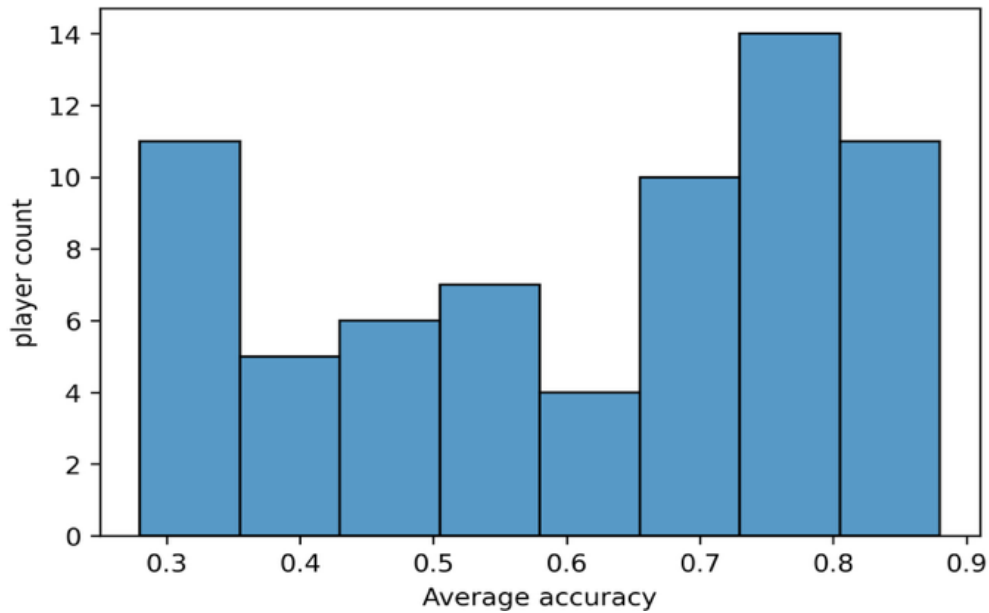


Figure 7.8: Histograms of average accuracy in Individual experimental conditions

7.8.3 Engagement and Workload Measures

Workers' engagement among collaborative image-labelling tasks was initially measured using the ten-item instrument (see Table 7.4), which comprised Autonomy, Competence, Relatedness, Enjoyment, and Perceived Usability. However, the experimental conditions were intentionally different in the extent of social interaction and collaborative affordances available to the workers, not all items were conceptually accurate for the direct comparisons among the conditions. Particularly, Relatedness items (Q5 and Q6) measured the perception of the social support and interpersonal connection. Such construct would be meaningful for both Full-Features and Chat-Only conditions in that workers could interact with teammates but were not equivalently represented in the Individual condition and partially represented in the No-Element condition. Inclusion of these items in a cross-condition engagement measure, then, would confound engagement with the experimental manipulation itself. Therefore, the Relatedness items were removed from the comparative engagement composite on conceptual and measurement-equivalence grounds. Similarly, a deeper examination of the remaining items showed that Q1 (“I did things in the task because they interested me”) and Q4 (“After doing the task for a while, I felt reasonably competent.”) were not consistently applicable to all experimental conditions. These items were also limited to context-specific motivational experiences (i.e., not a common engagement construct that could be construed essentially the same way across the four conditions). As such, they were also omitted from the comparative engagement measure. The final engagement composite consisted of six items (Q2, Q3, Q7, Q8, Q9 and Q10) representing dimensions that were conceptually comparable across all conditions. Reliability analysis confirmed that the derived scale demonstrated sufficient internal consistency across the dataset (Cronbach’s $\alpha > 0.70$) for a comparative analysis of the scale. The results shown in Table 7.9 indicated

a hierarchical pattern of engagement among the four experimental conditions. Participants allocated to the Full-Features condition exhibited the highest average score of engagement ($M = 4.51$), closely followed by participants in the Chat-only condition ($M = 4.49$). In contrast, lower engagement score has been observed in the NE ($M = 3.98$) and Individual ($M = 3.97$) conditions. These descriptive discrepancies indicate that conditions incorporating collaboration and system feedback were associated with higher engagement score than the minimal or individual conditions. The descriptive similarity of Full-Features as compared to Chat-Only condition also led to a further inferential test to determine whether these differences were statistically meaningful and reliable. The Shapiro-Wilk test revealing data violations of assumptions of normal distribution across all conditions and measures. Accordingly, a Kruskal-Wallis H test applied to assess a reliable difference in engagement among the four interface conditions ($H = 18.16, p < .001$). Post-hoc pairwise analyses employing the Mann-Whitney U test with Bonferroni correction that helped to explain these group-level effects. Rank-biserial correlation (r_{rb}) was computed for every significant difference found between any pair to evaluate effect size and guide interpretation of the direction of the difference. Pair-wise Mann-Whitney U tests showed that the Full-Features condition outperformed the Individual condition ($U = 4,003.00, p = .006, r_{rb} = -.23$). The Full-Features condition also demonstrated higher engagement than No-Element design ($U = 4,718.00, p = .004, r_{rb} = -.24$), indicating a small-to-moderate effect in favour of the Full-Features condition. Similarly, CO condition significantly outperformed NE approached ($U = 5,682.00, p = .001, r_{rb} = -.24$) and outperformed IND ($U = 4,792.50, p = .002, r_{rb} = -.23$). There was no significant difference between Full-Features and Chat-Only conditions ($p = .920$) suggesting that there are similar engagement differences in these two collaborative environments. Together, these engagement results show that collaborative task designs, specifically those with synchronous communication, are considerably more likely to generate engagement than either independent or minimally collaborative conditions. Although the Full-Features condition consistently showed the highest rates of engagement, the fact that Full-Features and Chat-Only had no significant difference indicates that communication alone contributes to a significant percentage of the engagement benefits which have been observed.

Table 7.9: Overall Engagements Across all Conditions

Condition	Mean	N
Full Features	4.51	107
Chat-only	4.49	126
Individual	3.98	61
No Element	3.97	72

These engagement findings address directly SRQ3.2 by showing how different combinations of collaboration and gamification affect psychological engagement dimensions including competence, enjoyment, perceived usability.

Perceived workload was measured on the basis of the six subscales of the NASA-TLX system (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration), which described how participants in each experimental condition perceived their cognitive and physical loads during labelling. As shown in Table 7.10, Full-Features participants also showed the lowest workload ratings per NASA-TLX dimension, Chat-Only were next in values and No-Elements were followed by the Individuals conditions. The Full-Features condition workers reported Mental demand of ($M = 4.65$), Physical demand of ($M = 4.29$), Temporal Demand ($M = 4.42$), Effort ($M = 4.58$), and Frustration ($M = 3.87$). These descriptive findings indicated that the participants experienced lower workload perception when collaboration and feedback features were available. In contrast, as shown in Table 7.10 and Figure 7.9, No-Elements presented the highest workload levels in several dimensions: Physical demand ($M = 5.11$), Temporal demand ($M = 5.30$), Effort ($M = 5.65$) and Frustration ($M = 5.07$), suggesting that collaborating with no supportive elements led to a heightened subjective effort for the tool use. The Individual condition similarly revealed elevated Mental Demand ($M = 5.18$) and Performance ratings ($M = 5.73$), indicating increased perceived cognitive and self-evaluated exertion in the non-collaborative control settings. In order to determine whether the differences in NASA-TLX were statistically significant, Kruskal-Wallis tests were conducted for each of the NASA-TLX dimensions. The resulting analyses showed that there were significant main effects on conditions for Physical Demand with ($H = 7.87, p = .049$), Temporal Demand ($H = 12.61, p = .006$), Effort ($H = 18.56, p < .001$), and Frustration ($H = 22.68, p < .001$). No statistical significance was found for Mental Demand or Performance ($p > .05$). Pairwise Mann-Whitney comparisons showed that participants in the No-Elements condition perceived significantly higher Physical Demand than those in Full-Features ($U = 2,919.50, p = .010, r_{rb} = .22$), while no other Physical demand differences were statistically significant ($p > .05$). Temporal Demand scores were significantly different between No-Elements comparing to Full-Features and ($U = 2,651.00, p < .001, r_{rb} = .30$). No-Elements comparing to Chat-Only show higher Temporal demand ($U = 3,522.00, p = .014, r_{rb} = .21$). These results indicated that the absence of collaborative aids or supportive features were associated with increased perceived time pressure during the task execution. Though mean Performance ratings showed a slight edge for Individual ($M = 5.73$) and No-Elements ($M = 5.70$) conditions, compared with Full-Features ($M = 5.54$) and Chat-Only ($M = 5.44$), these differences did not reach statistical significance ($p > .05$). This suggests that the perceived task success remained comparable across the conditions despite the differences experienced workload

Table 7.10: Average NASA-TLX Constructs Across Conditions

Condition	Mental Demand	Physical Demand	Temporal Demand	Performance	Effort	Frustration
Chat-Only	4.93	4.53	4.74	5.44	4.71	4.23
Full Features	4.65	4.29	4.43	5.54	4.58	3.87
No-Elements	5.00	5.11	5.29	5.70	5.65	5.07
Individual	5.18	4.90	4.92	5.73	5.07	4.88

However, Effort demand ratings showed that the condition effect was strong. Participants in the Full-Features group reported significantly lower average subjective Effort than those with No-Elements ($U = 2,414.00$, $p < .001$, $r_{rb} = .36$). The Chat-Only group also showed lower Effort than the NE condition ($U = 3,174.50$, $p < .001$, $r_{rb} = .28$), while the remaining pairwise comparisons were not statistically significant. For Frustration scores, No-Elements showed higher Frustration than Full-Features ($U = 2,356.00$, $p < .001$, $r_{rb} = .37$) as well as the Individual condition worker reported higher Frustration compared to Full-Features ($U = 2,153.00$, $p < .001$, $r_{rb} = .32$). Chat-Only also scored lower than both No-Elements ($U = 3,428.50$, $p = .007$, $r_{rb} = .23$) and Individual ($U = 3,035.00$, $p = .032$, $r_{rb} = .19$). Across all experimental conditions, Full-Features resulted consistently in lowest perceived workload: Individual and No-Elements produced the

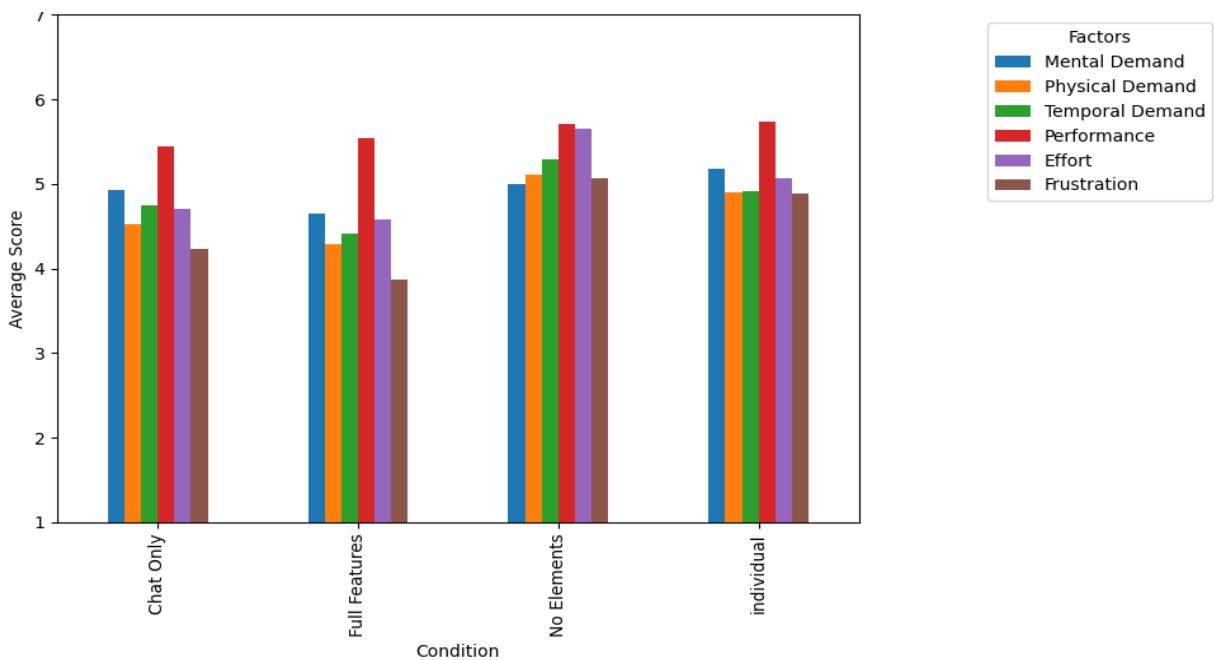


Figure 7.9: NASA-TLX distribution across conditions

highest strain across Physical, Temporal, Effort and Frustration dimensions. These workload results are a direct address to SRQ3.3, revealing that collaborative gamified conditions (more specifically, the richer features design) significantly reduce key aspects of perceived and emotional workload compared to those that are minimally supported or non-collaborative. To support these assessed engagement and workload measures, we extracted participants' self-reported individual perceptions of gamification properties in the three team-based environments (Full-Features, Chat-Only, No-Elements). The aim of this analysis was to offer more contextual insight into how specific the design elements were encountered rather than to present an alternative form of engagement. Average ratings of each item are visualised in Figure 7.10. Overall, participants in the Full-Features presented a high interest in the items related to progress visibility and group feedback (mean score awareness, team level, and leaderboard related). There is also show a general tendency for participants to place higher importance on those features as meaningful contributors to their experience of the task. In contrast, the Chat-Only condition found superior ratings of communication-related items (chatting with other people, sharing ideas), but significantly lower ratings of the importance of moving forward-oriented aspects, which were not observed in this condition. This is consistent with the idea that chat facilitates interpersonal social interaction but cannot replace more-advanced feedback and progression. The No-Elements condition, while also collaborative, presented a significantly weaker rank on key feature scores that pertain to motivation and monitoring of progress, with respect to most features related to these features. This pattern is helping to explain why collaboration without gamification support failed to produce similar gains in engagement or in accuracy. This analysis is descriptive in nature and is not meant to substitute for valid measuring engagement strategies. Instead, it contextualises. The perceptions support SRQ3.2 and SRQ3.3 because the participants demonstrated subjective valuing of those motivational resources from each condition, providing additional empirical support to the quantitative engagement and perceived workload findings.

7.8.4 Chat Logs Analysis

To analyse the role of communication resource in collaborative performance, chat logs were analysed for the two experimental conditions, those enabled communication via chat, Full-Features and Chat-Only. The analysis integrates descriptive interaction metrics, and inferential comparisons to characterise how the communication patterns that related to the accuracy and the behaviour of the team. These analyses directly address SRQ3.4. A content analysis was applied to classify the chat messages into types of communication based on predefined keyword patterns. Messages were coded into the following categories

- Suggested keywords: Explicit labelling proposals using structured keywords (e.g., "I suggest 2").
- Agreement keywords: Consensus words such as "agree," "yes," or "sure".

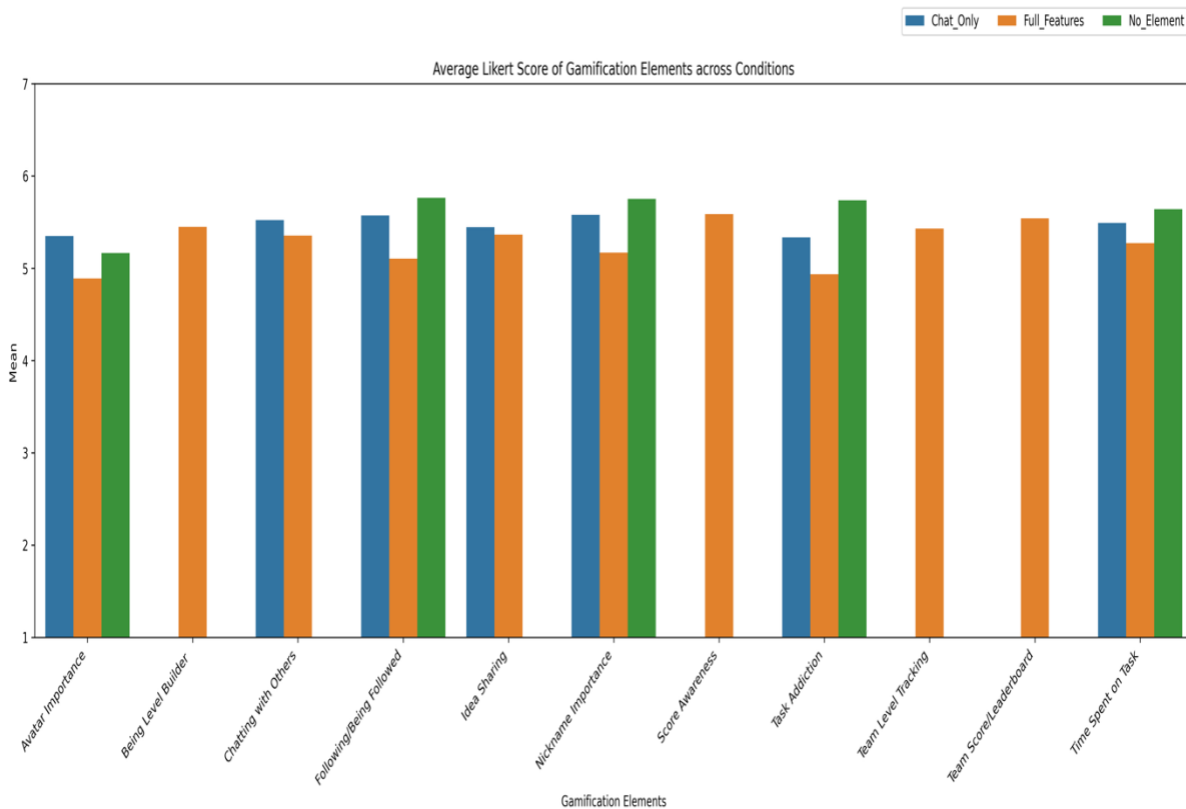


Figure 7.10: Average Likert ratings for gamification elements across team-based conditions.

- Disagreements: expressions of dissent, such as “disagree” or “no”.
- Social Exchanges: messages that express some sociality, like the word “Hi”.
- Game-Element Encouragement: expressions of competence such as “our score is low”.
- Free-Text: Messages not fitting in any of the above classifications.

The Full-Features condition produced a substantially higher number of suggestion and agreement messages, specifically those predefined from structured interface prompts. This indicated that interface scaffolding influenced message composition and density, supporting more task-focused collaborative exchanges. For the full-feature condition, 24 of 34 teams used the chat textbox and 28 of 44 teams from Chat-Only did. Then, the comparative breakdown revealed a total of 1,050 messages in Chat-Only and 1,612 messages in Full Features.

The average messages per team were:

- Full-Features: 47.41
- Chat-Only: 24.42

By calculating the average accuracy for the teams that used the chat messages vs. the teams that did not use chat in Full-Features, the results showed (Table 7.13) that the teams that used the chat feature had an average rate of ($M=0.69$) compared to ($M=0.63$) for teams that did not use the chat-text-box feature. In the CO

condition, the average accuracy for the teams that used chat was ($M=0.58$) and those that did not, was ($M=0.52$), indicating no meaningful performance difference resulting from using the chat feature alone.

To investigate the correspondence between communication intensity and labelling accuracy, teams were divided into two categories: small chat messages teams, and moderate chat messages teams, according to the median number of messages exchanged per team.

For the Full-Features condition:

- Mean chat messages = 67.17, median = 12.00
- Small chat teams: Average accuracy = 0.65
- Moderate chat teams: Average accuracy = 0.76

A Mann-Whitney U test indicated a significant difference between these two groups by which ($U = 14.50, p = .011$) with a moderate effect size ($r_{rb} = -.53$). This finding shows that moderate levels of intensity of communication, when implemented in a fully gamified environment, was associated with higher collaborative accuracy.

For Chat-Only:

- Mean chat messages = 37.50, median = 14.50.
- Small chat teams: Average accuracy = 0.54
- Moderate chat teams: Average accuracy = 0.53

No significant difference was found on the average accuracy between small and moderate chat under Chat-Only ($U = 70.50, p = .907$), indicating that an increased in communication volume alone did not enhance worker performance when not combined with other mechanisms such as feedback progress or scoring. Overall, SRQ3.4 has been answered by indicating that chat can be associated with better accuracy, within and as part of a broader gamified platform with shared performance feedback and cooperation support, whereas chat alone did not yield quantifiable accuracy improvements.

Table 7.11: Chat messages analysis

Experimental condition	# Teams _{chat}	Avg acc	#Teams _{no_chat}	Avg acc	#Teams _{small}	Avg acc	#Teams _{moderate}	Avg acc
Full-Features	24 teams	0.69	10 teams	0.63	14	0.65	7	0.76
Chat-Only	28 teams	0.58	15 teams	0.52	17	0.54	8	0.53

7.8.5 Big Five Personality Traits and Mood

To examine whether the workers' difference influenced labelling accuracy, the effect of mood and Big Five personality traits were investigated across the conditions. This analysis addresses SRQ3.5. Using a two-way ANOVA, there was no significant main effect of mood on labelling accuracy ($F(8, 327) = 0.98, p = .448$), and no significant mood * condition interaction ($F(24, 327) = 1.49, p = .071$). This means that people's self-reports of mood did not systematically affect their performance at labelling across various experimental conditions within the present study. Multiple regression analyses were performed to examine the predictive power of personality properties (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism). The omnibus Wald F-test for all five traits yielded no significant effect ($F(5, 345) = 1.65, p = .145$). In separate within-condition analyses, no significant associations were also found. Collectively, the results show that neither mood nor personality characteristics were a meaningful predictor of labelling accuracy in the present study. Performance levels were more influenced by task design and interaction structure rather than stable personal differences.

7.9 Discussion

This section synthesises the empirical results of Study 3 by summarising the findings in the context of the study's sub-research questions (SRQs). The goal is not to theorise the results but simply to streamline the results shown on the empirical performance, engagement, workload, collaboration, and individual differences prior to a theoretical discussion about them in theory in Chapter 8. Overall, the results provide an appropriate response to RQ3 and RQ4: collaborative gamification enhanced labelling results and worker experience, with the most reliable effect emerging when collaboration was combined with multiple integrated, complementary design elements (progress feedback, scoring, levels, shared performance cues with team interaction), rather than when collaboration of gamification elements was in isolation. For SRQ3.1, the Full-Features condition indicated the strongest and most consistent outcomes on quality-related factors. Compared to the other conditions, in terms of individual accuracy, majority accuracy and agreement strength, and a reduced percentage of images without a accuracy majority. In comparison, the Chat-Only and No-Elements conditions provided intermediate and low-performance results, respectively, and the Individual condition exhibited the lowest performance. These results, together, indicate that the improvements in performance were not only attributed to collaboration (in-team performance), but to the inclusion of collaboration with structured feedback, progress cues, and incentive mechanisms that helped the teams converge on reliable, shared decisions.

For SRQ3.2, the Level of engagement results were distributed according to the richness of the task environment and followed a clear graded pattern across the conditions. Overall, the Full-Features condition gave the highest overall engagement scores, closely followed by Chat-Only, with the No-Elements and Individual conditions having the lower engagement. This trend suggests that synchronous collaboration,

specifically the ability to communicate in real time, was a key determinant of perceived engagement, while the inclusion of broader gamification features primarily reinforced and stabilised the overall level of engagement rather than generating a large gap between the two chat-enabled conditions.

For SRQ3.3, the NASA-TLX-based workload ratings presented comparable distributions across conditions. Workers in the Full-Features condition achieved the smallest perceived effort, frustration, temporal demands, and physical demands, while the No-Elements and Individual conditions rated workload highest on various dimensions. Crucially, perceived performance did not differ significantly between conditions, suggesting that reductions in participants' workload under the Full-Features reflect a more supported task experience rather than a perceived trade-off in task success.

For SRQ3.4, chat log analysis clarified that the message volume alone did not consistently predict performance gains. In the Chat-Only condition, additional message exchange was not associated with accuracy or agreement while in the Full-Features mode, moderate and task-focused communication was associated with increased team accuracy level. This suggests that communication is performance-relevant when it is anchored in shared performance cues (e.g., progress feedback and scoring visibility) that allow teams to coordinate attention and interpret evidence against a common frame, rather than simply increasing message frequency.

For SRQ3.5, personality trait and self-reported mood analyses showed no main effects or interactions with the experimental condition, indicating that stable individual differences were not meaningful in predicting the labelling accuracy in the study. This strengthens the interpretation that the observed differences in performance, engagement, and workload were driven primarily by the design of the task environment and interaction structure, rather than by systematic differences in participant traits across conditions.

Overall, the findings indicate a cumulative and interactional basis for collaborative gamification effects across all sub-research questions. Performance quality, engagement, workload efficiency, and collaborative outcomes were consistently stronger in the Full-Features condition compared to reduced-feature and individual baseline cases. Partial implementations, e.g., chat alone, provided limited benefits and did not reach the benefits obtained by the richer integrated design. Together, these findings establish a strong empirical basis for explaining how and why the collaborative gamification mechanisms explain shape the workers engagement and their outcome quality. Chapter 8 therefore, builds on this evidence to develop the theoretical perspectives and explanation across-study synthesis of the thesis.

7.10 Threats to Validity

The study methodology addressed various validity issues to guarantee the dependability of the conclusions. In this section we discuss major threats to internal, external, construct and conclusion validity, as well as what steps were taken to alleviate them before interpreting the reported findings.

The internal validity is related to whether the gamified collaborative mechanisms rather than influences from outside produce the observed changes in worker engagement and labelling accuracy. Worker

diversity in knowledge and their past experience presents one possible hazard that can affect labelling accuracy independent of gamification. Strict worker qualifications were thus enforced, requiring an HIT approval rate of at least 95% and a minimum of 1,000 authorised HITs to guarantee that only experienced and trustworthy workers participated, hence mitigating this. To ensure and guarantee uniform interpretation of labelling criteria among all participants, task instructions and labelled example images were also given. In addition, Workers' random assignment to experimental settings further controlled for confounding elements, therefore assuring that observed variations in engagement and accuracy were the result of gamification rather than personal variation. Included to weed out outlier employees and raise data quality is a performance filter that has been applied. Together, these controls enhance confidence that differences in performance and engagement reflect experimental manipulations rather than uncontrolled participant heterogeneity. However, along with most online experiments, uncontrolled contextual factors (e.g., interruptions, device/environmental differences) will not be completely removed and are considered residual sources of variance. A further residual factor is uneven gender distribution by conditions driven by random assignment (gender was not a stratification variable); although gender showed no significant association with accuracy either within or across conditions, this imbalance is noted as a boundary on the engagement comparisons and a candidate stratification variable for future replications.

The external validity concerns the generalisability of results outside the particular experimental settings. The results might not fully reflect crowdsourcing populations in other areas, where variables including language proficiency, cultural attitudes toward cooperation, and gamification familiarity could influence participation and performance. Participation was limited to workers situated in the United States. Furthermore, as the study was carried out on Amazon Mechanical Turk, results might not exactly apply to other crowdsourcing sites with different payment systems, worker profiles, or interaction systems. These constraints set the boundaries within which reported results should be interpreted rather than challenging the internal coherence of the findings. Nonetheless, the study provides insights relevant to task-based crowdsourcing contexts by leveraging a varied pool of seasoned crowd workers and real-time team-based cooperation. Generalisability beyond similar task-based, online crowdsourcing environments is therefore bounded and addressed further in the discussion chapter. Further work could enhance the external validity of the study by replicating the study with different worker populations, platforms, and task domains.

Construct validity evaluates if the research fairly gauges worker performance, involvement, and teamwork. The dependence on behavioural engagement metrics such as chat activity, agreement rates, and leaderboard interactions, which might not entirely reflect workers' underlying motivation or subjective involvement levels poses a possible hazard. NASA-TLX workload measurements and self-reported engagement scales such as the User Engagement Scale (UES) were incorporated to augment objective behavioural data in order to solve this issue. Moreover, task instructions and examples guaranteed consistent

task knowledge, therefore avoiding variances resulting from worker interpretation instead of engagement differences. By taking an objective performance indicator and a validated subjective instrument into account and using each to compare conditions, it reduces the chance that the effects observed are due to measurement artefacts rather than differences between conditions. Because the UES refinement encompassed item removal, all reliability checks and exclusions were performed prior to inferential testing to eliminate the risk of post-hoc construct redefinition.

Conclusion validity is distributional checks (e.g., normality testing) allowed for appropriate inferential procedures and effect sizes were reported to support any interpretation beyond p-values. Agreement measurements were then utilised to check labelling accuracy, therefore guaranteeing that results were not distorted by noise or errors. To further support the robustness of inferential conclusions, suitable non-parametric statistical tests, effect size reporting, and correction for multiple comparisons were performed. Since the constraints placed upon team formation meant that the Individual condition needed to be smaller than the original target per-condition, we tested the attained statistical power of the main pairwise contrasts in the Individual condition after applying the pre-defined filtering. Applying G*Power (post-hoc, two-tailed $\alpha = .05$) and converting the calculated Mann–Whitney effect sizes (rank-biserial $r_{rb} \approx .23-.25$; Cohen's $d \approx 0.47-0.52$), we were left with approximately 0.85 for Full-Features vs Individual ($n_{FF}=121, n_{IND}=68$), 0.92 for Chat-Only vs Individual ($n_{CO}=138, n_{IND}=68$), and 0.84–0.90 for No-Elements vs Individual ($n_{NE}=112, n_{IND}=68$, depending on $d = 0.47$ or $d = 0.52$). These checks show that the important Individual-versus-team contrasts in both cases were adequately powered for the small-to-moderate effects observed (though small true effects would be more difficult to detect). Thus, the study raises confidence in the conclusions by means of careful study design, participant selection criteria, task standardisation, and powerful statistical processes handling validity concerns.

7.11 Conclusion

In this chapter, empirical data from the collaborative gamified image-labelling study were reported and analysed across performance, engagement, perceived workload, collaboration behaviour and individual differences under these experimental conditions. Results were systematically organised around the study's sub-research questions, presenting a structured and transparent account of how varied types of collaboration and gamification affected worker outputs. Throughout the chapter, a strong emphasis was placed on reporting both descriptive and inferential analyses, supported by validated quantitative measures and complementary qualitative analyses, ensuring the robustness and clarity in the findings.

In relation to the research focus of this study and its related sub-research questions of Chapter 7—SRQ3.1 (the impact of different collaborative team-based gamification arrangements on individual accuracy, majority accuracy, and agreement strength), SRQ3.2 (the influence of collaboration and gamification on psychological engagement), SRQ3.3 (the effect of collaborative gamified conditions on perceived workload),

SRQ3.4 (the role of synchronous chat feature in enhancing labelling performance and communication dynamics), and SRQ3.5 (the moderating effects of personality traits and mood on performance)—the chapter concludes that:

The Collaborative environments including both competitive elements alongside cooperative aspects, produced improvements in individual accuracy, majority accuracy, and stronger consensus agreement compared to non-gamified individual work. The engagement levels significantly improved under the conditions that integrated social communication features and gamification elements, which also support higher levels of autonomy, competence, and relatedness. The variability in perceived workload based on differing conditions, with collaborative and gamified environments affecting both cognitive demands as well as the effort required for tasks, underscores the necessity for balancing between task complexity and available support mechanisms. Synchronous communication emerges as vital for facilitating coordination processes while also improving consensus-building efforts among teams, leading to enhanced performance achievements. Individual characteristics like personality traits or mood further elucidate variations observed in engagement levels although their impacts tend to be contingent upon specific contextual factors.

The statistical evaluations detailed in this chapter effectively provide a clear answer to sub-research questions SRQ3.1 - SRQ3.5, ensuring that each analysis is theoretical motivated, empirically grounding and directly linked back to study's research questions. SRQ3.1; addressed through performance and agreement analysis across the conditions. Engagement measures address SRQ3.2. SRQ 3.3 through NASA-TLX workload assessments. SRQ 3.4 via the chat and interactions analysis. And lastly, the analysis of personality traits and mood effects answer SRQ3.5. These collective findings lay down a solid empirical groundwork for the subsequent chapter, where interpretation will extend beyond reporting aiming to interpret the findings in relation to motivation theory and prior crowdsourcing literature, framing and drawing the theoretical, design, and practical implication of the study.

Chapter 8 :Discussion

8.1 Introduction

This chapter draws on these findings and provides a cohesive and critical discussion, consolidating lessons learned from all empirical and review-based research conducted in this thesis. The chapter interprets the findings, discussing how they collectively contribute to the theoretical and design-driven debates with relevance to motivation, gamified systems, and collaborative crowd work. The goal is to situate the findings within broader scholarly debates and to articulate how the evidence across chapters advances understanding in this field. Based on these findings, the chapter offers a more general synthesis and reflection on the three studies that have been carried out in this thesis. We leverage the findings of an exploratory study on task-selection strategies, a scoping review of gamified crowdsourcing systems and a large-scale collaborative gamified labelling experiment. In this process, each study is considered not as standalone contribution, but as a complementary component of a whole research trajectory that develops conceptual, methodological, and empirical insight in incremental steps. Collectively, these studies advance a comprehensive understanding of how the designs of motivational gamification mechanisms, and collaborative functionalities could be configured in combination to enhance engagement and performance on crowdsourced labelling tasks.

For the purpose of conceptual coherence, the chapter highlights explicit interrelationships between these three studies, illustrating how early exploratory observations guided later work based on theory and how the experimental system discussed in Chapter 6 operationalised those theoretical insights. This sequential alignment, from exploratory observation study, through systematic syntheses, to the theory-driven experimental validation, support a coherent cumulative narrative rather than standalone results. Therefore, the chapter focuses on underlying mechanisms, design principles and theoretical implications to move beyond condition-specific comparisons. The cumulative structure increases the validity of the thesis by presenting a clear trajectory from initial behavioural patterns to theory-driven design choices and finally to controlled empirical verification.

8.2 Synthesis Across the Three Studies

The collected and integrated studies in this thesis were for the purpose of comprehending how motivational design strategies can let workers themselves be involved in the evaluation and processing of what they are evaluating and processing. These strategies are gamification and collaboration, as previously described in Chapter 2 and Chapter 5, and are believed to increase both worker engagement and accuracy in very repetitive labelling tasks (Ryan and Deci, 2000). Instead of treating gamification and collaboration as isolated interventions, this thesis conceptualises them as complementary design strategies that jointly shape how the workers perceive, engage, and sustain effort in crowdsourcing tasks. Combining results from an exploratory empirical study, a scoping review, and a collaborative experimental study, the thesis expanded knowledge

about motivational design in crowdsourcing. Collectively, the results present evidence that worker engagement and performance are not only influenced by task characteristics, but also by the social and motivational structures embedded within the system. Rather than the motivator being the feature of an individual's gamification, the findings indicate that successful motivational design is determined by the interaction between task design including feedback mechanisms, social participation opportunities, and workers' psychological needs.

The first study was focused on worker engagement and performance based on three common crowdsourcing task types: image labelling, text classification, and survey completion. This study indicated that the image-labelling task engaged workers more than other tasks, and that the task interest was, after monetary reward, among the strongest task-selection preference. While the study was not inherently built from the perspective of SDT (Ryan and Deci, 2000), certain observed behaviours can be interpreted through SDT, including the need to reinforce workers' perceived competence and autonomy. Simultaneously, the largely individual nature of the tasks offered limited social engagement and relatedness. This observation suggests that features of collaborative design and interactivity could be potential important ways to increase engagement beyond what can be accomplished just by task design. These findings are consistent with other crowdsourcing work, where systems are often optimised for achieving tasks independently and at a throughput that reduces social interaction, coordination, and collective sense-making (Bederson and Quinn, 2011, Kittur et al., 2013). However, they extend this literature by showing that the resulting engagement deficit is not intrinsic to the task but follows from the limited support for relatedness in conventional designs, a gap the later studies were designed to address. Importantly, the contribution of the exploratory study was not limited to specifying the differences between types of tasks. It emphasised that engagement is not solely determined by the nature of the task itself, but also by the extent to which the task environment supports workers' psychological needs. Such an observation serves as an empirical basis for a more systematic examination of motivational design in the subsequent scoping review and led to the adoption of Self-Determination Theory and Flow Theory as the principal theoretical lenses used throughout the remainder of the thesis (Csikszentmihalyi, 1990), (Csikszentmihalyi, 1997, Ryan and Deci, 2000). Building on these findings, the scoping review investigated gamified crowdsourcing platforms following the four archetypes suggested by (Geiger et al., 2012): crowd-processing, crowd-solving, crowd-rating and crowd-creating. The review further identified that, despite having been used to a large extent as an engagement trigger, gamification implementations are often just superficially integrated, and the focus is predominantly on points, badges or leaderboards (PBL) without appealing to fundamental motivational foundations. This is consistent with recent gamification literature that indicates that "PBL"-type designs are prevalent and are prone to competitive/achievement framings even when deeper motivational alignment is intended (Hamari et al., 2014, Koivisto and Hamari, 2019). The key important contribution of the review was the identification of an imbalance in the current evidence base. While

competitive and achievement-oriented mechanisms were broadly addressed, collaborative, socially oriented, and intrinsically motivated methods received comparatively less attention. Features that support relatedness, shared goals, teamwork, communication, and collective participation were relatively rare despite their strong theoretical connections to Self-Determination Theory. This finding suggested that many crowdsourcing systems focus on stimulating participation through external incentives while providing limited support for deeper motivational processes associated with engagement (Renard and Davis, 2019). The review thus revealed a notable gap: the absence of systematically well designed, theory-aligned, and socially relevant gamification in crowdsourcing systems. This gap justified the experimental system developed in this thesis, which combined extrinsic rewards (points, levels) and intrinsic motivational resources (collaboration, communication, and joint goals). Importantly, the review outlined that the motivational mechanisms should not be considered in isolation, but it should be interacting components whose effectiveness depend on how they are configured, sequenced, and integrated within the broader system design. In this way, the review shifted the focus of the thesis from examining gamification components individually toward understanding the gamification as a configuration problem. These insights directly informed the design of the collaborative experimental platform and selection of the motivational mechanisms assessed in the final study.

Lastly, we instantiate these conceptual and empirical characterisations in a collaborative gamified labelling platform and gauge it under four conditions: Full-Features, Chat-Only, No-Elements and Individual. In this last phase, the experimental study functioned as a test of theoretically informed motivational validation mechanism focus rather than simple comparison of isolated feature. The findings indicated that the motivational mechanisms were most effective when implemented as part of integrated system combined feedback social communication, shared goal and performance awareness (Csikszentmihalyi, 1990, Csikszentmihalyi, 1997, Ryan and Deci, 2000). The full-Features condition provides stronger engagement and performance outcome than the conditions containing fewer motivational features, together with lower perceived workload on key NASA-TLX dimensions (effort, frustration, and temporal demand). These findings suggested that the collaborative gamified system reinforce one another when designed around shared goal, social interaction and meaningful feedback. Thus, the results demonstrated that the benefit of motivational design does not arise from the inclusion of specific feature solely, but from the relationships between those features and the extent to which they collectively support the workers psychological needs. Collectively, the three studies have demonstrated the consistent theme that the effectiveness of motivational design in crowdsourcing cannot be explained solely by the presence or absence of individual gamification features. Rather, motivation arises out of the interaction between task characteristics, feedback structures, social participation opportunities, and workers' psychological needs. Thus, the contribution of this thesis is not merely to demonstrate any particular design features. It offers empirical support that motivational mechanisms ought to be configured as coherent systems that align with established motivational theory and the contextual

requirements of the task. Through empirical observation, literature synthesis, and experimental evaluation, this thesis proposes a theoretically grounded and empirically validated framework for designing and developing collaborative crowdsourcing systems to increase worker performance and engagement.

8.3 Contributions to knowledge and Practice

Results from this thesis, taken as a whole, combine to make substantial contributions both to theory and practice in the fields of crowdsourcing, gamification and human motivation. This research integrates behavioural, theoretical and experimental perspectives across the three empirical studies to propose a coherent and theoretically informed analysis of how motivation and cooperation mechanisms work together to influence performance and engagement in crowdsourced image-labelling task. These contributions are developed and interpreted in full below, they are restated in summary form, mapped to the research questions, in Chapter 9.

8.3.1 Theoretical Contributions

This study takes existing concepts further in a number of related respects. By combining SDT and Flow Theory in a collaborative crowdsourcing environment, this study has shown that motivation in digital labour is not explainable by incentive structures of isolation. High level of engagement is achieved when system functions together to satisfy autonomy, competency and relatedness, which are the key concepts of SDT, while maintaining the balance between challenge and skill, as described by Flow Theory (Csikszentmihalyi, 1990, Csikszentmihalyi, 1997, Ryan and Deci, 2000). Theoretical progress is achieved by considering motivation as an emergent property of interacting design elements instead of being the sum of independent motivational levers (Deterding et al., 2011). In this sense, the collaboration fundamentally reshapes how motivation needs were felt by the workers within the task environment, not only amplifying the engagement. Motivation is socially located, reinforced through feedback, and maintained through shared progress instead of through individual reward accumulation (Ryan and Deci, 2000). This builds on previous gamification perspectives which frequently assess mechanics as isolated “features”, by proving that their motivational role depends on how they are associated with coordination resources (e.g., feedback visibility, shared progress cues, and synchronous interaction) rather than on the existence of any one single element (Deterding et al., 2011, Hamari et al., 2014, Koivisto and Hamari, 2019). An important theoretical approach is to show how motivational mechanisms are not independent but supportive and complement each other when well designed. As we see, the interactions between structured challenge (Flow), meaningful social presence (SDT-relatedness), and autonomy-supportive features produce motivational experiences which are qualitatively different from those generated by any single mechanism taken in isolation (Csikszentmihalyi, 1990, Csikszentmihalyi, 1997, Ryan and Deci, 2000). This supports a more integrated model of motivation in crowdsourcing, which acknowledges both the individual’s accountability processes as well as the social dynamics of real-time collaboration (Kittur et al., 2013, Retelny et al., 2014). To this end, the thesis resonates with coordination-focused views of crowd work that focus on structuring collaboration and gamification so that teams can reliably converge under

uncertainty rather than assuming that more participation automatically produces better outcomes (Retelny et al., 2014, Valentine et al., 2017).

The research also revises the (Geiger et al., 2012) archetypal classification by finding a hybrid form that combines the features of both crowd-processing and crowd-creating. The gamified collaborative labelling environment discussed in this thesis represents a transitional archetype where individual tasks contribute to emergent collective understanding through synchronous interaction communication and feedback (Geiger et al., 2012, Geiger and Schader, 2014). This refinement builds on crowdsourcing theory by indicating that modern systems appear to be blurring the line between structured task completion and collective knowledge production (Kittur et al., 2013, Prpić et al., 2015). By demonstrating that workers collaboratively build higher-level understanding by means of collective feedback, team discussion, and agreement dynamics, the results indicate a new category of “collaborative micro-creation”, that exists in between established archetypes. This conceptual contribution widens the existing taxonomies to better reflect modern and interaction-rich crowdsourcing systems (Geiger and Schader, 2014, Prpić et al., 2015).

8.3.2 Methodological Contributions

Methodologically, the thesis provides a multi-phase mixed methods research design and evaluation framework that integrates behavioural performance metrics, as well as a psychometric measures and communicational data within controlled experimental conditions (Creswell and Clark, 2017, Denzin, 2017). At the methodological level, the thesis combines such techniques as accuracy measures, agreement strength, workload indicators and constructs of involvement into a robust method for assessing the psychological and performance effects of design interventions in mass systems (Hart and Staveland, 1988, O’Brien et al., 2018). Its contribution here means that future research on motivation may well follow the example it sets as human–AI or human–collective projects (Kittur et al., 2013). This multi-phase mixed methods approach enhances methodological practice in multiple ways:

- It demonstrated how short-term motivational dynamics can be captured in real-time collaborative environments using behavioural logs and validated psychometric scales (Hart and Staveland, 1988, O’Brien et al., 2018).
- It illustrated the analytical benefit of using agreement-based and consensus-oriented metrics alongside accuracy measurements when investigating collective intelligence (Chafetz, 2005, Kittur et al., 2013).
- It has provided an experimental blueprint for evaluating crowdsourcing systems that combine social communication and gamification, domains which are often descriptively rather than empirically treated (Deterding et al., 2011, Retelny et al., 2014).

These methodological contributions strengthen future research on the motivation of digital labour and human–AI collaboration.

8.3.3 Practical and Design Contribution

At a practical level, this thesis provides insights for the design and management of crowdsourcing systems: guidelines for architecture and tools that take gamified collaboration into account (Deterding et al., 2011). The results show that gamification works best when it is built within collaborative workflows rather than added on as a surface-level motivator layer (Hamari et al., 2014, Nicholson, 2014). Characteristics like levels, communication channels feedback mechanisms, and peers answers visibility are designed to work together and be consistent with intrinsic motivational principles (Ryan and Deci, 2000, Csikszentmihalyi, 1997). From the design perspective, the effective collaboration and gamification required coherence across social, feedback, and challenge structures. Systems promoting shared goal, progressive challenge calibration, and transparent performance feedback are better able to maintain engagement and performance thus the data quality (Retelny et al., 2014, Valentine et al., 2017). Crucially, reliance on purely financial incentives is reduced by fostering internally regulated motivation (Ryan and Deci, 2000). The following design-focused principles are thus proposed in the thesis:

- Real-time communication resources that support shared problem-solving and relatedness (Ryan and Deci, 2000, Retelny et al., 2014).
- Structures of progressive challenge (levels, escalating honeypots) that maintain flow while avoiding disengagement (Csikszentmihalyi, 1997).
- Feedback loops that provide timely competence-supportive cues without overwhelming or distracting workers (Ryan and Deci, 2000, Nicholson, 2014).
- Fair reward mechanisms and transparent team structures that maintain trust and perceived equity (Retelny et al., 2014, Ryan and Deci, 2000).

For a platform developer and practitioners, the thesis research provides evidence-based guidance on how to balance extrinsic and intrinsic rewards. Gamified cooperation can improve engagement and performance without raising expenses, as long as incentives support rather than replace intrinsic motivation (Hamari et al., 2014, Retelny et al., 2014, Ryan and Deci, 2000). By doing this, it helps both academic understanding and the practical development of collaborative human computation systems (Kittur et al., 2013). Finally, this thesis has shown that theoretically grounded behavioural analytics can guide actual design decisions. By incorporating insights from motivational psychology, HCI, and human computation this work outlines a replicable model for translating theory into scalable system design. That, in turn, adds to academic knowledge together with contributions that can be applied in the practical field of collaborative human computation systems.

8.4 Limitation and Methodological Reflections

As with any empirical research study, the studies presented in this thesis, are subject to certain limitations. These limitations basically are not deficiencies in the research design, but its primarily reflect the

methodological constraints associated with conducting multi-phase mixed-method investigation and evaluating real-time collaborative system with online workers. Recognising the conditions presence makes our research more transparent and suggests new directions for future studies advancing collaborative human computation into the realm of gamification. Thus, these limitations are not perceived as shortcomings, but rather reflective boundaries that frame the interpretation and transferability of the findings

8.4.1 Constraints and Environment Design Issues

We conducted these studies in controlled yet simulated environments, striking a balance between the accuracy of an experiment and its ecological veracity (Shadish et al., 2002). In spite of Amazon Mechanical Turk's access to diverse populations, its workforce may not represent the vast array of potential contributors to open projects (Difallah et al., 2018, Hara et al., 2018, Martin et al., 2017). Differences in cultural background, task familiarity and intrinsic motivation could dramatically influence people's perceptions about how gamification and collaboration affect them. Accordingly, the findings should be understood as an outcome of a subset of experienced, platform-aware crowd workers instead of universal crowds. So, while those image-labelling tasks might be good for testing engagement and performance, they are just one type of creative crowd work. Results could well differ for text-based or ideation tasks requiring complex reasoning and creativity (Kittur et al., 2013). Further constraint is in the task domain specificity, Galaxy Zoo images provide a wealth of visual cues and clear category boundaries; However, the same motivational structures may not work for other crowdsourcing domains including dialogue annotation, creative ideation, or subjective judgment tasks. This highlights the importance of task-motivation alignment for generalising collaborative gamification strategies beyond visual tasks. Another contextual limitation that should be mentioned concerns the temporal nature of participation. Each experimental session recorded short-term engagement within discrete task levels, rather than letting people interact for days on end or over multiple projects at a stretch. As a result, the conclusions represent immediate motivational effects instead of a need developed in later contextualisation. As such, long-term motivational sustainability, habituation impact, and the effects of something novel are still open research questions (Hamari et al., 2014, Koivisto and Hamari, 2019). Therefore, while the findings demonstrated a strong short-term effect, future work is needed to determine whether these effects persist across a longitudinal study or if novel effects fade over time.

8.4.2 Measurement and Analytical Considerations

Though many kinds of validated quantification tools (UES, NASA-TLX) and rigorous accuracy metrics were used, outfitting participant self-reports still could not escape the shadow of bias (Hart and Staveland, 1988, O'Brien et al., 2018). Maybe subjects responded under the influence of presupposition or temporary emotion, preferring at any rate to describe a stable state as if it were instead of an unsteady past. These limitations are also inherent in self-reported motivational measures, reiterating the requirement for triangulation (Denzin, 2017). To identify and resolve this issue, we supplemented the subjective feeling data with objective

behavioural information, for instance completion times, interaction logs, and chat messages. Nevertheless, further research in the future might amplify validation by adding physiological or subconscious touchstone indicators.

Moreover, although non-parametric methods were appropriate for the analysis which was a consequence of skewed distributions, future work could involve the application of advanced modelling methods such as:

- Multilevel models that capture team-level variance (e.g., to model varying team size produced by dynamic formation as a grouping factor).
- Agent-based simulations exploring emergent patterns of collaboration.

Such methods would provide further critical insights into the interdependence between collaboration, performance, and engagement. Specifically, hierarchical and network-based models may be more suitable to account for the recursive feedback in collaborative crowd work (Kittur et al., 2013).

8.4.3 Methodological Reflections

There were many difficulties in conducting real-time cooperative experiments which affected methodological decisions. The variation of Internet connections and response latencies caused slight asynchrony for participants in the shared workspace. The study compensated for this by employing server-side time calculations and checking data, but minute differences in time still could have impinged upon feelings of collaboration flowing well. This highlights a challenge in studying synchronous collaboration in distributed online environments. Although the bulk of inconsistencies were reduced as a result of technical mitigation solutions, micro-latencies still appear to be inevitable features of online experimentation in real time (Retelny et al., 2014, Valentine et al., 2017). The process of filtering out unreliable or unfocused workers was a necessary measure to maintain data quality and also sometimes resulted in a shrinking of condition sample sizes. This reflects a typical methodological conflict between internal validity and ecological realism in crowdsourcing research (Kittur et al., 2013, Shadish et al., 2002). By recognising these limitations, this thesis provides both empirical evidence and a methodological perspective on the pragmatic aspects of undertaking robust collaborative crowd experiments at scale.

8.5 Conclusion

This chapter has synthesised the facts and theoretical insights from the numerous research studies contained within this thesis to explain how motivational and collaborative design can be used to improve performance and engagement in crowdsourcing tasks. Its purpose has been interpretive: to position the cumulative evidence in relation to Self-Determination Theory, Flow Theory, and prior crowdsourcing work, and to draw out what the findings mean rather than to restate them. The thesis' original research concluded that participation in crowd work is not stimulated by any one reward but by a family of motivation mechanisms that effectively link individual effort to collective outcomes. Throughout the research, this thesis demonstrates that motivation-based gamification works best when it is based in a context of structured, socially interactive task environments

that support autonomy, competence, and relatedness while maintaining an appropriate challenge–skill balance. These findings extend SDT and Flow Theory by applying them to multi-worker, real-time situations in which intrinsic motivation arises as a collective phenomenon in a real-time collaborative context. The corresponding design principles, the consolidated answers to the research questions, and the forward-looking agenda are presented in Chapter 9. Finally, the methodological reflections provided in this thesis highlight the intricate task of achieving true motivation in a digital context and suggest a practical direction for future work aiming to reconcile experimental control with ecological validity. Together, these interpretation position collaborative gamified crowdsourcing as a viable and theoretically sound means for designing high-quality, human computation systems.

Chapter 9 : Conclusion and Future Work

9.1 Overview of the Research

The aim of this thesis study was to explore how motivational design, gamification and real-time collaboration can contribute to improve workers engagement, performance and experience in crowdsourcing image-labelling tasks. Across the three linked studies, an exploratory examination of the task-selection strategies, a scoping exploration of gamified crowdsourcing, and large scale collective gamified study, this research systematically constructed a multi-phase mixed methods research, understanding of the factors that successfully effective and motivating the workers in digital online community environments. These studies, collectively, were designed as a cumulative research programme rather than stand-alone studies, which would enable behavioural observation, theoretical synthesis and experimental validation to inform the following study. The central argument of the thesis is that the best crowdsourcing systems are those that combine motivational design elements with social collaboration, rather than only extrinsic incentives or isolated gamified features. The findings across the studies indicate that engagement and accuracy are most effective when workers are provided with a sense of autonomy, competence, and social connectedness, consistent with SDT, and when they enjoy a task experience that maintains the balance of challenge and skill as posited by Flow Theory. Instead of treating motivation as an individual or static feature, this dissertation shows that motivation in crowdsourcing appears to be a dynamic and socially situated phenomenon influenced by the design of the system, the structure of the interactions and the feedback mechanisms. The collaborative gamified system described in Chapter 6 operationalised these, with far higher engagement, better agreement and decreased perceived workload than those without collaboration or with no gamification. Moreover, this thesis demonstrates that crowdsourcing is becoming more than conventional task paradigms. The findings indicate that in circumstances when collaborating and gamifying are effectively intertwined, tasks can promote collective sense-making, rather than standalone execution. The gamified collaborative environment we created displays aspects of hybrid crowdsourcing type, crowd-processing orientation, supplemented by crowd-rating mechanisms and selective collaborative elements that partially overlap with those typically associated with crowd-creating. This reveals the emergent category of hybrid crowdsourcing where tasks are nested in socially enriched systems that promote collective solution. In particular this conceptual contribution extends existing crowdsourcing archetypes by illustrating how a real-time collaboration may lead to a transformation on task output without giving up on the scalability of tasks. The thesis also states the necessity for crowdsourcing systems based on human motivation, not only labour cost efficiencies. On contrary, when gamification is properly embedded within collaboration, not just as a superficial reward, or as a structured way to address the psychology of engagement, it can offer high-quality engagement even in repetitive and tedious work such as

image labelling. This thesis posits motivationally aligned collaborative gamification not merely as an optional addition, but as a fundamental design technique for forthcoming crowdsourcing and human–AI systems.

The findings from Chapters 4, 5, and 7 cumulatively demonstrate that collaborative and gamified task designs significantly affect worker engagement, performance, and agreement quality in crowdsourcing contexts, as outlined in the overarching thesis research questions. Worker performance and engagement are affected not only by the task types or rewards but are also shaped by the motivational design and the interaction frameworks. Gamification, when integrated with real-time collaboration instead of being applied in isolation, produces improvement in accuracy, agreement strength, and engagement level. Social contact mechanisms, as well as, especially synchronous communication, serve as essential facilitators of coordination, consensus-building, and intrinsic motivation. Overall, these findings produce a direct and comprehensive answers to the thesis research questions by demonstrating that the effective crowdsourcing system should integrate the motivational resources, as well as the social and structural design features to enhance the workers' engagement, and outcome quality.

9.2 Thesis Research Questions

The structure of this thesis was organized into four research questions addressed by a study each with a distinct contribution.

RQ1 How do workers' performance and engagement vary across different crowdsourcing task types? Addressed by the exploratory study (Chapter 4), which established that performance, engagement, and perceived workload systematically differ between image, text, and survey tasks, and that task interest ranks second only to monetary reward in task selection, a finding that highlights the primary design lever (C1) of intrinsic interest.

RQ2 How have additional task features (e.g., gamification elements) been applied across the main crowdsourcing archetypes? Addressed by the scoping review (Chapter 5), which mapped literature (146 studies) and identified a predominance of competitive, points-based gamification and a comparative lack of investigation of collaborative supporting design (C2).

RQ3: How can such features be systematically designed and integrated to improve workers' performance and engagement in image-labelling tasks? Addressed by the design and development of the collaborative gamified platform (Chapter 6) that instantiated SDT- and Flow-based mechanisms over four configurable conditions (C3).

RQ4: How do such features affect the reliability of crowdsourcing outcomes, particularly worker agreement and data quality? Addressed by the experimental evaluation (Chapter 7), that found that the Full-Features condition had a significantly better labelling accuracy and agreement strength and lower perceived workload than the reduced and individual conditions (C4).

9.3 Summary of Key Findings

This study has made significant empirical and theoretical contributions to the literature on how worker engagement and workload, as well as data quality, are influenced by collaborating gamification in crowdsourcing environments. In a controlled series of experiments with the pairing of participants engaged in real-world image-labelling tasks, we find that people who complete tasks in collaborative environments exhibit both significantly higher engagement and accuracy relative to those not using the gamified or interpersonal task conditions. Full-features participants (i.e., collaboration with scoring, feedback, levels, chat, and leaderboard) achieved the highest engagement scores across all conditions. They managed also to obtain greater levels of agreement and accuracy, indicating that the social and feedback-based components increased attentiveness and decision-making consistency. They also reported lower perceived workload on key dimensions (effort, frustration, and temporal) in comparison to those not collaborating, suggesting that providing real-time collaboration and feedback reduced perceived workload without compromising performance. Teams with low levels of gamification (e.g., Chat-Only) showed average engagement and higher workloads, and individual players had the lowest engagement but higher workloads on average. These differences were statistically significant (full test statistics and effect size that reported in Chapter 7 and interpreted 8), indicating that integrated gamified collaboration is more effective than isolated approaches.

The contributions therefore offer empirical evidence that joint resources applied from motivational and collaborative perspectives, drawing on the SDT and Flow Theory, can foster intrinsic motivation, maintain engagement, and contribute to better-quality outcomes in human-AI data-labelling workflows.

9.4 Thesis Contributions

This research contributes to the theoretical, methodological, and practical aspects of crowdsourcing, human-computer interaction, and motivational design. The contributions are defined here in summary form and mapped to the studies; they are developed, interpreted, and positioned against the literature in Chapter 8. The thesis progresses through three main studies, each one producing a distinct contribution (C1-C4); these together advance our knowledge on understanding how motivational and collaborative mechanisms influence performance and engagement in the field of crowdsourcing. Across studies, worker outcomes are consistently examined across three dimensions, performance, engagement, and perceived workload. Performance is operationalised progressively based on each study's design, baseline task accuracy in the exploratory Study 1 (C1), accuracy-related outcomes synthesised from the literature in the scoping review (C2), and labelling accuracy and worker agreement strength in the experimental Study 3 (C4).

C1 – Empirical insight into how workers perform and engage in baseline crowdsourcing tasks

This thesis provides an empirical insight about workers behaviour, at the performance, engagement, and perceived workload across baseline crowdsourcing tasks on various task types, including 750 crowd workers recruited through Amazon Mechanical Turk. It indicates how differences in task characteristics influence

performance, engagement, perceived workload, with performance captured through baseline accuracy. This study sets a basis for interpreting worker behaviour within the following gamified and collaborative designs. It contributes to enhancing the theoretical understanding of worker diversity, motivation, and behaviour in large-scale crowd environments, as well as providing empirical evidence in the context of task selection and participation within digital labour markets.

C2 – A theoretical synthesis on motivation-oriented factors in crowdsourcing

This thesis offers an in-depth overview of the way motivational aspects, and particularly gamification design, have been employed throughout various crowdsourcing archetypes. It broadens the scope of crowdsourcing studies with an integration of SDT and Flow Theory, which reflects a holistic account of motivation and engagement. It adopts a systematic mapping methodology to provide an overall picture that answers the question of why gamification elements can motivate workers, examining how these elements relate to worker performance (including accuracy), engagement and workload, and pointing out the issue of collaborative motivational design.

C3 -Methodological Framework for Collaborative Gamified Crowdsourcing

This thesis describes a custom-built experimental platform to support team-effort real-time crowdsourcing with familiar services from traditional game-playing environments, including chat, levels, and leaderboards, as well as configurable gamification such as scores and feedback. It serves as a reusable methodological test platform to investigate synchronous crowd collaboration and includes a combined evaluation pipeline of behavioural performance outcome measures and engagement questionnaires, NASA-TLX, and the Big Five Personality Inventory.

This methodological framework provides a strong foundation to examine the critical role that distinct motivational and collaborative configurations have on the engagement of evidence's reliability.

C4 – Empirical and Practical Insights into Collaborative Gamification Impact

This work empirically demonstrates that collaborative gamification significantly improves worker performance (operationalised as labelling accuracy and workers' agreement), and engagement, while reducing perceived workload on key dimensions in image-labelling tasks relative to non-gamified settings. It shows how integrating game design features that fulfil the psychological needs of autonomy, competence, and relatedness (SDT) can support internal motivation while reducing perceived workload. In application, C4 informs the construction of motivation human-in-the-loop AI systems by illustrating how collaborative motivational scaffolding can support the validity of data for training in ML. Moreover, it could help designers of the next generation of crowdsourcing platforms to design for enjoyment on the one hand and fairness and productivity in a massive-scale online collaboration on the other. In summary, these contributions:

- Advance theory on how gamification and collaboration interact to influence worker motivation in the collective digital labour market.

- Provide a validated methodological description to assess the reliability of gamified systems (including high levels of internal and the external reliability).
- Provide practical implications for building effective, accurate, and scalable human-AI data pipelines.

These contributions lay the groundwork for next-generation gamified crowdsourcing systems that both enhance human experience and data quality.

9.5 Future Work

The work presented here can be extended to certain promising directions of future research in multiple ways. First, longitudinal and cross-platform studies might explore long-term effects of collaborative gamification on sustained participation, skill development, and trust among crowd workers. Such studies would be particularly useful, as well, in identifying if motivation effects seen in short-term experimental conditions were maintained while novelty effects diminished with time. Furthermore, because collaborative teams were formed dynamically and varied in size (two to five workers), team size was not experimentally controlled in this work, future work may consider team size as an experimental rather than an emergent property of dynamic assignment by comparing fixed compared to variable team sizes and skill-based groupings in order to examine how a team-formation design influences consensus, coordination, and labelling quality. Secondly, the new designs could investigate adaptive gamification, wherein mechanics are dynamically modified to worker performance and emotional states with the support of real-time analytics. Adaptive systems could enable motivational scaffolding to adapt to shifting levels of engagement, workload, and collaboration quality at both the individual and team level. Third, the inclusion of AI-mediated facilitation, for example, automated moderation, intelligent feedback, or a hybrid human–AI collaborative approach, could improve the scalability and inclusivity of complex labelling operations. This is in line with emerging human-in-the-loop and human–AI collaboration paradigms in which automation complements collective human intelligence rather than replaces it.

Furthermore, the transferability of this framework to other task domains (e.g., design ideation) may provide more general insights into collective intelligence and creative collaboration. Qualitative follow-up studies allow qualitative insights into social interaction patterns, communication quality, and shared motivation in gamified teams and thus enhance the nascent discussions in human-centred crowdsourcing. These dynamics, across other task types could help define which boundary conditions the collaborative gamification model is most able to operate under. By demonstrating this empirical and theoretical synergy, this thesis contributes to progress towards human-centric crowdsourcing systems that prioritise both worker experience and quality output. These studies collectively advance the field to modify a new paradigm by which the gamified collaborative actions enable the crowd not only to contribute but also to co-create a crucial step toward realising human-centred crowdsourcing systems and paradigms that would underpin the future of digital work.

9.6 Thesis Conclusion

This thesis has demonstrated that the effective crowdsourcing increasingly lies on enriched, socially and motivationally aligned collaborative environments rather than on isolated task execution. This research systematically addresses the thesis questions through exploratory, conceptual, and experimental studies, demonstrating that engagement, performance, and agreement quality are intricately linked and may be collectively optimised by deliberate system design. Through the integration of empirical study, review of existing literature, and experimental evaluation, this thesis demonstrates that the outcomes of workers are influenced not only by task characteristics but also by the motivational factors and social structures embedded within the crowdsourcing systems.

The results confirm that crowdsourcing systems should move from pure economic models to human-centred approaches that incorporate psychological, social, and technological dimensions. This thesis offers a theoretical framework and a practical guide for developing advanced crowdsourcing and human–AI systems that are efficient, engaging, and meaningful for participants. Instead of treating gamification and collaboration as individual interventions, the results suggest that their effectiveness emerges from how they are designed and integrated to support workers’ needs for competence, autonomy, and relatedness. Engagement emergent not as consequence of a particular design elements, but as product of interaction between the task features, feedback mechanisms, social participation opportunities, and motivational support together.

Generative AI technologies have rapidly altered the realm of crowdsourcing, especially in the context of data annotation, content creation, and human-AI collaboration workflows. Tasks were traditionally handled by crowd workers, such as image labelling, text classification, and producing content which could now be fully or partially automated through advanced, large language models and vision systems. This evolution does not reduce the importance of crowdsourcing, but it has redefined its function. Human workers are increasingly positioned as validators, collaborators, and coordinators of AI-generated outputs rather than being the sole producers of data. In this context, the findings presented in this thesis gain additional significance. The highlighted significance of engagement, collaboration, and agreement strength within human teams directly applies to hybrid human–AI systems, where improving motivation, trust, and coordination are essential. The same motivational issues that we addressed throughout this thesis, involving engagement, supporting trust, organizing coordination, and encouraging effective participation, remain critical when humans and AI technology play collaborative roles.

Moreover, the thesis presents a promising model for merging human intelligence with generative AI systems through the development of collaborative gamified environments. For example, these systems could be enhanced to facilitate AI-assisted labelling processes, resolve disagreements between human- and AI-generated outputs, or collectively validate the created content. Such environments would benefit from the complementary strength of human judgment, collective intelligence, and AI, rather than treating these

capabilities as competing alternatives. This indicates that the future of crowdsourcing should concentrate on developing frameworks where human motivation and collective intelligence work in tandem with ever-advancing generative models rather than simply substituting human labour with AI.

In conclusion, this thesis argues that the of crowdsourcing becomes not about a decision between human workers and intelligent machines, but about designing systems in which the two can effectively work hand in hand together. Through the confluence of motivational theory, collaborative systems design, and empirical analysis, the thesis offers a basis for the next generation of crowdsourcing and human–AI collaboration contexts that are meaningful and productive for their participants. In summary, this thesis shows that collaborative gamification may turn crowdsourcing from fragmented individual labour into meaningful, socially connected collaboration. The interplay between motivational design and cooperative task structures shows that engagement and performance do not have to be mutually exclusive: in fact, they are mutually reinforcing.

References

- ABID, S. K.;ROOSLI, R.;NAZIR, U. & KAMARUDIN, N. S. 2025. Ai-enhanced crowdsourcing for disaster management: Strengthening community resilience through social media. *Int J Emerg Med*, 18(1), pp. 201. <https://doi.org/10.1186/s12245-025-01009-9>
- AGUINIS, H.;VILLAMOR, I. & RAMANI, R. S. 2021. Mturk research: Review and recommendations. *Journal of management*, 47(4), pp. 823-837. <https://doi.org/10.1177/014920632096>
- AIPE, A. & GADIRAJU, U. 2018. Similarhits: Revealing the role of task similarity in microtask crowdsourcing. *Proceedings of the 29th on hypertext and social media*.
- AITAMURTO, T. 2015. Motivation factors in crowdsourced journalism: Social impact, social change, and peer learning. *International Journal of Communication*, 9(pp. 3523-3543.
- ALBARQOUNI, S.;MATL, S.;BAUST, M.;NAVAB, N. & DEMIRCI, S. Playsourcing: A novel concept for knowledge creation in biomedical research. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016. 269-277.
- ALDWAIRI, M. & SHUHAIBER, A. 2019. Exploring the integration of cdio, crowdsourcing and gamification into information security courses. *World Transactions on Engineering and Technology Education*, 17(3), pp. 237-243.
- ALJANAKI, A.;WIERING, F. & VELTKAMP, R. C. 2016. Studying emotion induced by music through a crowdsourcing game. *Information Processing & Management*, 52(1), pp. 115-128. <https://doi.org/10.1016/j.ipm.2015.03.004>
- ALLAHBAKSH, M.;ARBABI, S.;GALAVII, M.;DANIEL, F. & BENATALLAH, B. 2018. Crowdsourcing planar facility location allocation problems. *Computing*, 101(3), pp. 237-261. <https://doi.org/10.1007/s00607-018-0670-1>
- ALTMAYER, M.;LESSEL, P. & KRÜGER, A. Expense control: A gamified, semi-automated, crowd-based approach for receipt capturing. *International Conference on Intelligent User Interfaces, Proceedings IUI*, 2016. 31-42.
- AMORIM, A. M.;RIBEIRO, A.;AROUCA, M. G.;MEIJON, I.;CAVALHEIRO, V.;PESTANA, M. C. & VIEIRA, V. 2024. Gamification techniques and contribution filtering in crowdsourcing micro-task applications. *Journal on Interactive Systems*, 15(1), pp. 401-416. <https://doi.org/10.5753/jis.2024.3727>
- ANDRÉ, P.;KRAUT, R. E. & KITTUR, A. Effects of simultaneous and sequential work structures on distributed collaborative interdependent tasks. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014. 139-148.
- APOSTOLOPOULOS, K.;GELI, M.;PETRELLI, P.;POTSIU, C. & IOANNIDIS, C. 2016. A new model for cadastral surveying using crowdsourcing. *Survey Review*, 50(359), pp. 122-133. <https://doi.org/10.1080/00396265.2016.1253522>
- ARECHAR, A. A.;GACHTER, S. & MOLLEMAN, L. 2018. Conducting interactive experiments online. *Exp Econ*, 21(1), pp. 99-131. <https://doi.org/10.1007/s10683-017-9527-2>
- ARELLANO-VERDEJO, J. & LAZCANO-HERNANDEZ, H. E. Two algorithms for sargassum segmentation in citizen science images. *2023 Mexican International Conference on Computer Science (ENC)*, 2023. IEEE, 1-8.

- ARKSEY, H. & O'MALLEY, L. 2005. Scoping studies: Towards a methodological framework. *International journal of social research methodology*, 8(1), pp. 19-32. <https://doi.org/10.1080/1364557032000119616>
- ARONSON, E. 1969. The theory of cognitive dissonance: A current perspective. *In: BERKOWITZ, L. (ed.) Advances in experimental social psychology volume 4*. Academic Press.
- B.V, E. (2022). Available at: <https://www.scopus.com/> (Accessed:)
- BAER, M. F. & PURVES, R. S. 2022. Window expeditions: A playful approach to crowdsourcing natural language descriptions of everyday lived landscapes. *Applied Geography*, 148(pp. 102802. <https://doi.org/10.1016/j.apgeog.2022.102802>
- Bär, M. F. 2022. *Capturing perceived everyday lived landscapes through gamification and active crowdsourcing*. University of Zurich.
- BASTANFARD, A.; SHAHABIPOUR, M. & AMIRKHANI, D. 2024. Crowdsourcing of labeling image objects: An online gamification application for data collection. *Multimedia Tools and Applications*, 83(7), pp. 20827-20860. <https://doi.org/10.1007/s11042-023-16325-6>
- BEDERSON, B. B. & QUINN, A. J. (2011). Web workers unite! Addressing challenges of online laborers. In *Proceedings of CHI '11 Extended Abstracts on Human Factors in Computing Systems* (pp. 97-106). Association for Computing Machinery, <https://doi.org/10.1145/1979742.1979606>
- BEHL, A.; SAMPAT, B. & RAJ, S. 2021. Productivity of gig workers on crowdsourcing platforms through artificial intelligence and gamification: A multi-theoretical approach. *The TQM Journal*, 38(1), pp. <https://doi.org/10.1108/TQM-07-2021-0201>
- BIEGEL, B.; BECK, F.; LESCH, B. & DIEHL, S. Code tagging as a social game. *Proceedings - 30th International Conference on Software Maintenance and Evolution, ICSME 2014, 2014*. 411-415.
- BIN HUSSEIN, M. I. H.; FONG, J. H.; LIM, C. X.; LEE, J. S. A.; TAN, C. T. & NG, Y. Y. Mobile application for crowdsourced gamification of automated external defibrillator (aed) locations. *HealthMedia 2019 - Proceedings of the 4th International Workshop on Multimedia for Personal Health and Health Care, co-located with MM 2019, 2019*. 24-31.
- BOELL, S. K. & CECEZ-KECMANOVIC, D. 2015. On being 'systematic' in literature reviews in is. *Journal of Information Technology*, 30(2), pp. 161-173. <https://doi.org/10.1057/jit.2014.26>
- BORROMEO, R. M.; LAURENT, T.; TOYAMA, M. & AMER-YAHIA, S. Fairness and transparency in crowdsourcing. *International Conference on Extending Database Technology (EDBT), 2017*.
- BOUNOV, D.; DEROSI, A.; MENARINI, M.; GRISWOLD, W. G. & LERNER, S. Inferring loop invariants through gamification. *Conference on Human Factors in Computing Systems - Proceedings, 2018*.
- BOWSER, A.; HANSEN, D.; HE, Y.; BOSTON, C.; REID, M.; GUNNELL, L. & PREECE, J. (2013). Using gamification to inspire new citizen science volunteers. In *Proceedings of Proceedings of the First International Conference on Gameful Design, Research, and Applications* (pp. 18-25). Association for Computing Machinery, <https://doi.org/10.1145/2583008.2583011>
- BRABHAM, D. C. Crowdsourcing. *The international encyclopedia of organizational communication*.
- BRABHAM, D. C. 2008. Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence*, 14(1), pp. 75-90. <https://doi.org/10.1177/1354856507084420>

- BRENNER, M.;MIRZA, N. & IZQUIERDO, E. People recognition using gamified ambiguous feedback. *Proceedings of the First International Workshop on Gamification for Information Retrieval*, 2014. 22-26.
- BROSSARD, D.;LEWENSTEIN, B. & BONNEY, R. 2005. Scientific knowledge and attitude change: The impact of a citizen science project. *International Journal of Science Education*, 27(9), pp. 1099-1121. <https://doi.org/10.1080/09500690500069483>
- BROVELLI, M. A.;CELINO, I.;FIANO, A.;MOLINARI, M. E. & VENKATACHALAM, V. 2018. A crowdsourcing-based game for land cover validation. *Applied Geomatics*, 10(1), pp. 1-11. <https://doi.org/10.1007/s12518-017-0201-3>
- BUHRMESTER, M.;KWANG, T. & GOSLING, S. D. 2016. Amazon's mechanical turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science*, 6(pp. 3-5. <https://doi.org/10.1177/174569161039398>
- BUJARI, A.;CIMAN, M.;GAGGI, O. & PALAZZI, C. E. 2016. Using gamification to discover cultural heritage locations from geo-tagged photos. *Personal and Ubiquitous Computing*, 21(2), pp. 235-252. <https://doi.org/10.1007/s00779-016-0989-6>
- CÁCERES, P.;CUESTA, C. E.;VELA, B.;CAVERO, J. M. & SIERRA, A. 2019. Smart data at play: Improving accessibility in the urban transport system. *Behaviour & Information Technology*, 39(6), pp. 681-694. <https://doi.org/10.1080/0144929x.2019.1652852>
- CAO, Z.;FENG, H. & WILES, M. A. 2023. When do marketing ideation crowdsourcing contests create shareholder value? The effect of contest design and marketing resource factors. *Journal of Marketing*, 88(2), pp. 99-120. <https://doi.org/10.1177/00222429231191446>
- CARLIER, A.;SALVADOR, A.;CABEZAS, F.;GIRO-I-NIETO, X.;CHARVILLAT, V. & MARQUES, O. 2015. Assessment of crowdsourcing and gamification loss in user-assisted object segmentation. *Multimedia Tools and Applications*, 75(23), pp. 15901-15928. <https://doi.org/10.1007/s11042-015-2897-6>
- CASCAES CARDOSO, M. (2017). The onboarding effect. In *Proceedings of Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 263-267). Association for Computing Machinery, <https://doi.org/10.1145/3027063.3027128>
- CATALLO, I. & MARTINENGHI, D. The dimensions of crowdsourcing task design. *International Conference on Web Engineering*, 2017. Springer, 394-402.
- CHAFETZ, J. 2005. It's the aggregation, stupid (reviewing james surowiecki, the wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations (2004). HeinOnline.
- CHALWE, C.;CHANDA, C.;MUZYAMBA, L.;MWAPE, J. & PHIRI, L. Quantitative analysis of zambian wikipedia contributions: Assessing awareness, willingness, motivation, and the impact of gamified leaderboards and badges. *Annual Conference of South African Institute of Computer Scientists and Information Technologists*, 2024. Springer, 30-44.
- CHAMBERLAIN, J. The annotation-validation (av) model: Rewarding contribution using retrospective agreement. *Proceedings of the First International Workshop on Gamification for Information Retrieval*, 2014. 12-16.
- CHANDLER, J.;MUELLER, P. & PAOLACCI, G. 2014. Nonnaivete among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behav Res Methods*, 46(1), pp. 112-30. <https://doi.org/10.3758/s13428-013-0365-7>
- CHEN, J.;RIBA, P.;FORNES, A.;MAS, J.;LLADOS, J. & PUJADAS-MORA, J. M. Word-hunter: A gamesourcing experience to validate the transcription of historical manuscripts.

- Proceedings of International Conference on Frontiers in Handwriting Recognition, ICFHR, 2018. 528-533.
- CHOI, J.;CHOI, H.;SO, W.;LEE, J. & YOU, J. A study about designing reward for gamified crowdsourcing system. International Conference of Design, User Experience, and Usability, 2014. Springer, 678-687.
- CLARKE, V. & BRAUN, V. 2014. Thematic analysis. *Encyclopedia of critical psychology*. Springer.
- COHEN, J. 1992. Statistical power analysis. *Current directions in psychological science*, 1(3), pp. 98-101. <https://doi.org/10.1111/1467-8721.ep10768783>
- COHEN, J. 2016. A power primer. *American Psychological Association*, 112(1), pp. 155-159. <https://doi.org/10.1037/14805-018>
- COOPER, S.;KHATIB, F.;TREUILLE, A.;BARBERO, J.;LEE, J.;BEENEN, M.;LEAVER-FAY, A.;BAKER, D.;POPOVIC, Z. & PLAYERS, F. 2010. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307), pp. 756-60. <https://doi.org/10.1038/nature09304>
- CORDOVA MANCHENO, E. A. 2017. *Gamification techniques oriented to a crowdsourcing application to improve quality of participation through motivational affordances*. MASTER'S DEGREE.
- CRESWELL, J. W. & CLARK, V. L. P. 2017. *Designing and conducting mixed methods research*, Sage publications.
- CRESWELL, J. W. & CRESWELL, J. D. 2017. *Research design: Qualitative, quantitative, and mixed methods approaches*, Sage publications.
- CSIKSZENTMIHALYI, M. 1990. *Flow: The psychology of optimal experience*, Harper & Row New York.
- CSIKSZENTMIHALYI, M. 1997. Flow and the psychology of discovery and invention. HarperPerennial, New York, 39(pp).
- CUI, L.;ZHAO, X.;LIU, L.;YU, H. & MIAO, Y. Learning complex crowdsourcing task allocation strategies from humans. *Proceedings of the 2nd International Conference on Crowd Science and Engineering*, 2017. 33-37.
- DAI, P.;RZESZOTARSKI, J. M.;PARITOSH, P. & CHI, E. H. And now for something completely different: Improving crowdsourcing workflows with micro-diversions. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2015. 628-638.
- DANIEL, F.;KUCHERBAEV, P.;CAPPIELLO, C.;BENATALLAH, B. & ALLAHBAKHSH, M. 2018. Quality control in crowdsourcing. *ACM Computing Surveys*, 51(1), pp. 1-40. <https://doi.org/10.1145/3148148>
- DARADKEH, M. & ATALLA, S. 2022. Innovation crowdsourcing mechanisms and innovation performance: An empirical study of a business intelligence community. *World Journal of Entrepreneurship, Management and Sustainable Development*, 18(5), pp. 633-651. <https://doi.org/10.47556/J.WJEMSD.18.5.2022.5>
- DE FRANGA, F. A.;VIVACQUA, A. S. & CAMPOS, M. L. M. Designing a gamification mechanism to encourage contributions in a crowdsourcing system. *Proceedings of the 2015 IEEE 19th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2015*, 2015. 462-466.
- DE, L.;LI, X. & SANTHANAM, R. 2013. Digital games and beyond: What happens when players compete? *MIS Quarterly*, 37(1), pp. 111-124. <https://doi.org/10.25300/MISQ/2013/37.1.05>
- DECI, E. L. & RYAN, R. M. 1985. Conceptualizations of intrinsic motivation and self-determination. *Intrinsic motivation and self-determination in human behavior*. Springer.

- DENG, J.;DONG, W.;SOCHER, R.;LI, L.-J.;LI, K. & FEI-FEI, L. Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition, 2009. Ieee, 248-255.
- DENG, J.;KRAUSE, J.;STARK, M. & FEI-FEI, L. 2016. Leveraging the wisdom of the crowd for fine-grained recognition. *IEEE Trans Pattern Anal Mach Intell*, 38(4), pp. 666-76. <https://doi.org/10.1109/TPAMI.2015.2439285>
- DENZIN, N. K. 2017. *The research act: A theoretical introduction to sociological methods*, Routledge.
- DERGOUSOFF, K. & MANDRYK, R. L. Mobile gamification for crowdsourcing data collection: Leveraging the freemium model. *Conference on Human Factors in Computing Systems - Proceedings*, 2015. 1065-1074.
- DESMET, P.;VASTENBURG, M. & ROMERO, N. 2016. Pick-a-mood manual. Delft University of Technology, Delft, pp.
- DETERDING, S. 2015. The lens of intrinsic skill atoms: A method for gameful design. *Human-Computer Interaction*, 30(3-4), pp. 294-335. <https://doi.org/10.1080/07370024.2014.993471>
- DETERDING, S.;DIXON, D.;KHALED, R. & NACKE, L. From game design elements to gamefulness: Defining" gamification". *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, 2011. 9-15.
- DIFALLAH, D.;FILATOVA, E. & IPEIROTIS, P. (2018). Demographics and dynamics of mechanical turk workers. In *Proceedings of Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 135-143). Association for Computing Machinery, <https://doi.org/10.1145/3159652.3159661>
- DIFALLAH, D. E.;CATASTA, M.;DEMARTINI, G.;IPEIROTIS, P. G. & CUDRÉ-MAUROUX, P. 2015. The dynamics of micro-task crowdsourcing: The case of amazon mturk. *Proceedings of the 24th international conference on world wide web*. International World Wide Web Conferences Steering Committee.
- DOAN, A.;RAMAKRISHNAN, R. & HALEVY, A. Y. 2011. Crowdsourcing systems on the world-wide web. *Communications of the ACM*, 54(4), pp. 86-96.
- DUGGAN, N. M.;JIN, M.;DURAN MENDICUTI, M. A.;HALLISEY, S.;BERNIER, D.;SELAME, L. A.;ASGARI-TARGHI, A.;FISCHETTI, C. E.;LUCASSEN, R.;SAMIR, A. E.;DUHAIME, E.;KAPUR, T. & GOLDSMITH, A. J. 2024. Gamified crowdsourcing as a novel approach to lung ultrasound data set labeling: Prospective analysis. *J Med Internet Res*, 26(pp. e51397. <https://doi.org/10.2196/51397>
- DUHAIME, E. P.;JIN, M.;MOULTON, T.;WEBER, J.;KURTANSKY, N. R.;HALPERN, A. & ROTEMBERG, V. Nonexpert crowds outperform expert individuals in diagnostic accuracy on a skin lesion diagnosis task. 2023 IEEE 20th international symposium on biomedical imaging (ISBI), 2023. IEEE, 1-5.
- DUMITRACHE, A.;AROYO, L.;WELTY, C.;SIPS, R.-J. & LEVAS, A. Dr. Detective”: Combining gamification techniques and crowdsourcing to create a gold standard in medical text. *Proceedings of the 1st International Workshop on Crowdsourcing the Semantic Web (CrowdSem 2013)*, 12th International Semantic Web Conference, 2013. Citeseer.
- DZIEDZIC, D. 2016. Use of the free to play model in games with a purpose: The robocorp game case study. *Bio-Algorithms and Med-Systems*, 12(4), pp. 187-197. <https://doi.org/10.1515/bams-2016-0020>

- DZOMBAK, R.;MOUAKKAD, S. & MEHTA, K. 2016. Motivations of women participating in a technology-based social entrepreneurship program. *Advances in Engineering Education*, 5(1), pp. 28.
- EICKHOFF, C.;HARRIS, C. G.;DE VRIES, A. P. & SRINIVASAN, P. Quality through flow and immersion: Gamifying crowdsourced relevance assessments. *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, 2012. 871-880.
- ERNESTIVITA, G.;KUMAR, V. & ANISAH, T. N. 2024. Role of gamification in crowdsourcing. *Marketing and gamification*. Routledge.
- ERYİĞİT, G.;ŞENTAŞ, A. & MONTI, J. 2021. Gamified crowdsourcing for idiom corpora construction. *Natural Language Engineering*, 29(4), pp. 909-941. <https://doi.org/10.1017/S1351324921000401>
- ESTELLÉS-AROLAS, E. & GONZÁLEZ-LADRÓN-DE-GUEVARA, F. 2012. Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2), pp. 189-200. <https://doi.org/10.1177/0165551512437638>
- FAUL, F.;ERDFELDER, E.;LANG, A. G. & BUCHNER, A. 2007. G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods*, 39(2), pp. 175-91. <https://doi.org/10.3758/bf03193146>
- FEYISETAN, O. & SIMPERL, E. 2019. Beyond monetary incentives: Experiments in paid microtask contests. *ACM Transactions on Social Computing*, 2(2), pp. 1-31. <https://doi.org/10.1145/3321700>
- FEYISETAN, O.;SIMPERL, E.;VAN KLEEK, M. & SHADBOLT, N. (2015). Improving paid microtasks through gamification and adaptive furtherance incentives. In *Proceedings of Proceedings of the 24th International Conference on World Wide Web* (pp. 333-343). International World Wide Web Conferences Steering Committee, <https://doi.org/10.1145/2736277.2741639>
- FIELD, A. 2013. *Discovering statistics using ibm spss statistics*. sage London.
- FONG, C. L.;HOH, D. Z.;SU, H.;CHEN, P. Y.;TSAL, C. C.;TSENG, K. W. H.;HUANG, H. C.;WU, J. Y.;NOZAWA, Y. & CHAN, B. K. K. 2025. Crowdsourcing conservation: Unveiling taiwan's sea turtle foraging grounds, emerging threats, and residency with broad societal engagement. *BMC Ecol Evol*, 25(1), pp. 27. <https://doi.org/10.1186/s12862-025-02354-2>
- FORT, K.;GUILLAUME, B.;PILATTE, Y. A.;CONSTANT, M. & LEFÈBVRE, N. Rigor mortis: Annotating mwes with a gamified platform. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, 2020. 4395-4401.
- FRERING, L.;STEINBAUER-WAGNER, G. & HOLZINGER, A. 2025. Integrating belief-desire-intention agents with large language models for reliable human–robot interaction and explainable artificial intelligence. *Engineering Applications of Artificial Intelligence*, 141(pp. 109771. <https://doi.org/10.1016/j.engappai.2024.109771>
- GADIRAJU, U. & DIETZE, S. Improving learning through achievement priming in crowdsourced information finding microtasks. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 2017. 105-114.
- GADIRAJU, U.;FETAHU, B.;KAWASE, R.;SIEHNDEL, P. & DIETZE, S. 2017. Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Trans. Comput. Hum. Interact.*, 24(4), pp. 30:1-30:26. <https://doi.org/10.1145/3119930>
- GADIRAJU, U.;KAWASE, R. & DIETZE, S. A taxonomy of microtasks on the web. *Proceedings of the 25th ACM conference on Hypertext and social media*, 2014. 218-223.

- GADIRAJU, U.;KAWASE, R.;DIETZE, S. & DEMARTINI, G. (2015). Understanding malicious behavior in crowdsourcing platforms. In Proceedings of Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 1631-1640). Association for Computing Machinery, <https://doi.org/10.1145/2702123.2702443>
- GALLUS, J. 2017. Fostering public good contributions with symbolic awards: A large-scale natural field experiment at wikipedia. *Management Science*, 63(12), pp. 3999-4015. <https://doi.org/10.1287/mnsc.2016.2540>
- GAMAGE, G.;PERERA, I.;MEEDENIYA, D. & WELIVITA, A. 2020. A crowdsourced gameplay for whole-genome assembly via short reads. *International Journal of Online and Biomedical Engineering (iJOE)*, 16(08), pp. 68-84. <https://doi.org/10.3991/ijoe.v16i08.14821>
- GEIGER, D.;ROSEMAN, M. & FIELT, E. (2011). Crowdsourcing information systems—a systems theory perspective. In Proceedings of Australasian Conference on Information Systems ACIS 2011 (pp. Association of Information System,
- GEIGER, D.;ROSEMAN, M.;FIELT, E. & SCHADER, M. 2012. *Crowdsourcing information systems-definition typology, and design*, AISel.
- GEIGER, D. & SCHADER, M. 2014. Personalized task recommendation in crowdsourcing information systems — current state of the art. *Decision Support Systems*, 65(pp. 3-16. <https://doi.org/10.1016/j.dss.2014.05.007>
- GENG, Y. (2024). A review of crowdsourcing applications in artificial intelligence. In Proceedings of Proceedings of the 2024 8th International Conference on Big Data and Internet of Things (pp. 387-391). Association for Computing Machinery, <https://doi.org/10.1145/3697355.3697419>
- GONCALVES, J.;HOSIO, S.;FERREIRA, D. & KOSTAKOS, V. Game of words: Tagging places through crowdsourcing on public displays. *Proceedings of the 2014 conference on Designing interactive systems*, 2014. 705-714.
- GONG, X.;SONG, M.;WANG, S.;JIAO, R. J. & HELO, P. 2025. Crowdsourced manufacturing in industry 4.0: Implications and prospects. *Systems*, 13(3), pp. 183. <https://doi.org/10.3390/systems13030183>
- GOOD, B. M.;LOGUERCIO, S.;GRIFFITH, O. L.;NANIS, M.;WU, C. & SU, A. I. 2014. The cure: Design and evaluation of a crowdsourcing game for gene selection for breast cancer survival prediction. *JMIR Serious Games*, 2(2), pp. e7. <https://doi.org/10.2196/games.3350>
- GORJUP, G.;GEREZ, L. & LIAROKAPIS, M. 2021. Leveraging human perception in robot grasping and manipulation through crowdsourcing and gamification. *Front Robot AI*, 8(pp. 652760. <https://doi.org/10.3389/frobt.2021.652760>
- GRABOVIY, P. 2016. Methods of motivation improvement and effectiveness increase on the example of construction industry enterprises. *Procedia engineering*, 165(pp. 1520-1528. <https://doi.org/10.1016/j.proeng.2016.11.888>
- GUO, J.;NI, Q.;WU, W. & DU, D.-Z. 2023. Multi-task diffusion incentive design for mobile crowdsourcing in social networks. *IEEE Transactions on Mobile Computing*, 23(5), pp. 5740-5754. <https://doi.org/10.1109/TMC.2023.3310383>
- GUO, J.;WANG, J. & YAN, Z. Motivation and factors effecting the participation behavior in the urban crowdsourcing logistics: Evidence from china. *Proceedings of the 10th International Conference on E-Education, E-Business, E-Management and E-Learning*, 2019. 334-341.
- GUTIÉRREZ PÁEZ, N. F.;GÓMEZ-CAÑÓN, J. S.;PORCARO, L.;SANTOS, P.;HERNÁNDEZ-LEO, D. & GÓMEZ, E. Emotion annotation of music: A citizen science approach. *Lecture Notes in*

- Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2021. 51-66.
- GUY, I.;PERER, A.;DANIEL, T.;GREENSPAN, O. & TURBAHN, I. Guess who? Enriching the social graph through a crowdsourcing game. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2011. 1373-1382.
- HAJARIAN, M.;DIAZ, P. & AEDO, I. 2024. Overgamified? An exploratory study on gamification and crowdsourcing for fake accounts detection in social networks. International Journal of Human-Computer Interaction, 41(15), pp. 1-27. <https://doi.org/10.1080/10447318.2024.2427989>
- HAJARIAN, M. & HEMMATI, S. A crowdsourcing and gamification based product ranking method for e-commerce. 2021 7th International Conference on Web Research, ICWR 2021, 2021. 197-201.
- HALLIFAX, S.;SERNA, A.;MARTY, J. C.;LAVOUÉ, G. & LAVOUÉ, E. Factors to consider for tailored gamification. CHI PLAY 2019 - Proceedings of the Annual Symposium on Computer-Human Interaction in Play, 2019. 559-572.
- HAMARI, J.;KOIVISTO, J. & SARSA, H. Does gamification work?--a literature review of empirical studies on gamification. 2014 47th Hawaii international conference on system sciences, 2014. Ieee, 3025-3034.
- HAMARI, J.;SJÖKLINT, M. & UKKONEN, A. 2016. The sharing economy: Why people participate in collaborative consumption. Journal of the association for information science and technology, 67(9), pp. 2047-2059. <https://doi.org/10.1002/asi.23552>
- HAMMON, L. & HIPPER, H. 2012. Crowdsourcing. Business & Information systems engineering, 4(3), pp. 163-166. <https://doi.org/10.1007/s12599-012-0215-7>
- HANCHETT, E. & LISTWON, B. 2018. *Vue. Js in action*, Simon and Schuster.
- HANTKE, S.;APPEL, T. & SCHULLER, B. The inclusion of gamification solutions to enhance user enjoyment on crowdsourcing platforms. 2018 1st Asian Conference on Affective Computing and Intelligent Interaction, ACII Asia 2018, 2018.
- HAO, A.;LEE, C. S. & TAN, W. C. W. Fun with aeds: Examining the effects of a gamified mobile crowdsourcing application. ACM International Conference Proceeding Series, 2019. 207-211.
- HARA, K.;ADAMS, A.;MILLAND, K.;SAVAGE, S.;CALLISON-BURCH, C. & BIGHAM, J. P. (2018). A data-driven analysis of workers' earnings on amazon mechanical turk. In Proceedings of Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-14). Association for Computing Machinery, <https://doi.org/10.1145/3173574.3174023>
- HARA, K.;ADAMS, A.;MILLAND, K.;SAVAGE, S.;HANRAHAN, B. V.;BIGHAM, J. P. & CALLISON-BURCH, C. Worker demographics and earnings on amazon mechanical turk: An exploratory analysis. Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, 2019. 1-6.
- HARRIS, C. G. The beauty contest revisited: Measuring consensus rankings of relevance using a game. Proceedings of the First International Workshop on Gamification for Information Retrieval, 2014. 17-21.
- HARRIS, C. G. Finding the best job applicants for a job posting: A comparison of human resources search strategies. IEEE International Conference on Data Mining Workshops, ICDMW, 2017. 189-194.

- HARRIS, C. G. & SRINIVASAN, P. 2019. My word! Machine versus human computation methods for identifying and resolving acronyms. *Computación y Sistemas*, 23(3), pp. 893-904. <https://doi.org/10.13053/cys-23-3-3249>
- HART, S. G. Nasa-task load index (nasa-tlx); 20 years later. *Proceedings of the human factors and ergonomics society annual meeting*, 2006. Sage publications Sage CA: Los Angeles, CA, 904-908.
- HART, S. G. & STAVELAND, L. E. 1988. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*. Elsevier.
- HASSAN, L. 2016. Governments should play games. *Simulation & Gaming*, 48(2), pp. 249-267. <https://doi.org/10.1177/1046878116683581>
- HATTIE, J. & TIMPERLEY, H. 2007. The power of feedback. *Review of educational research*, 77(1), pp. 81-112. <https://doi.org/10.3102/003465430298487>
- HAWORTH, B.;USMAN, M.;SCHAUMANN, D.;CHAKRABORTY, N.;BERSETH, G.;FALOUTSOS, P. & KAPADIA, M. 2021. Gamification of crowd-driven environment design. *IEEE Comput Graph Appl*, 41(4), pp. 107-117. <https://doi.org/10.1109/MCG.2020.2965069>
- HE, J.;BRON, M.;AZZOPARDI, L. & DE VRIES, A. Studying user browsing behavior through gamified search tasks. *Proceedings of the First International Workshop on Gamification for Information Retrieval*, 2014. 49-52.
- HE, Z. & ZHANG, X. Multi-party auction-based incentive mechanism of data market in mobile crowdsourcing. *2024 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, 2024. IEEE, 878-882.
- HO, C.-J.;SLIVKINS, A.;SURI, S. & VAUGHAN, J. W. Incentivizing high quality crowdwork. *Proceedings of the 24th International Conference on World Wide Web*, 2015. 419-429.
- HO, C.-J. & YIN, M. 2018. Working in pairs: Understanding the effects of worker interactions in crowdwork. *arXiv preprint arXiv:1810.09634*, pp. <https://doi.org/10.48550/arXiv.1810.09634>
- HOLOVATY, A. & KAPLAN-MOSS, J. 2009. *The definitive guide to django: Web development done right*, Apress.
- HONG, Y.;KWAK, H.;BAEK, Y. & MOON, S. Tower of babel: A crowdsourcing game building sentiment lexicons for resource-scarce languages. *Proceedings of the 22nd International Conference on World Wide Web*, 2013. 549-556.
- HOQUE, M. M. 2024. Crowdfunding for innovation: A comprehensive empirical review. *Future Business Journal*, 10(1), pp. 102. <https://doi.org/10.1186/s43093-024-00387-5>
- HOWE, J. 2006a. Crowdsourcing: A definition. *Wired Magazine*.
- HOWE, J. 2006b. The rise of crowdsourcing. *Wired magazine*.
- HOWE, J. 2008. *Crowdsourcing: How the power of the crowd is driving the future of business*, Random House.
- HOWELL, D. C. 2011. Chi-square test: Analysis of contingency tables.
- HUMAYUN, H.;GHAZALI, M.;MALIK, M. N.;VERONICA, S. M.;VENICA, L.;INSIYAH, J.;ASTUTIK, S.;SOEHONO, L. A.;HEGAZY, N. Y. & KHAFAGY, M. H. 2023. The role of motivational theories in the success of crowdsourcing engagement models: A review. *Journal of Theoretical and Applied Information Technology*, 101(2), pp. 30.
- HUMAYUN, H.;MALIK, M. N. & GHAZALI, M. 2024. Analysis of motivational theories in crowdsourcing using long tail theory: A systematic literature review. *International Journal of Crowd Science*, 8(1), pp. 10-27. <https://doi.org/10.26599/ijcs.2023.9100010>

- HUOTARI, K. & HAMARI, J. 2017. A definition for gamification: Anchoring gamification in the service marketing literature. *Electronic Markets*, 27(1), pp. 21-31. <https://doi.org/10.1007/s12525-015-0212-z>
- HUTTER, K.;HAUTZ, J.;FÜLLER, J.;MUELLER, J. & MATZLER, K. 2011. Communitition: The tension between competition and collaboration in community-based design contests. *Creativity and Innovation Management*, 20(1), pp. 3-21. <https://doi.org/10.1111/j.1467-8691.2011.00589.x>
- IMRAN, M.;MITRA, P. & CASTILLO, C. (2016). Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. In *Proceedings of Tenth International Conference on Language Resources and Evaluation* (pp. 1638–1643). European Language Resources Association (ELRA),
- IPEIROTIS, P. G. 2010a. Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, 17(2), pp. 16-21. <https://doi.org/10.1145/1869086.1869094>
- IPEIROTIS, P. G. 2010b. Demographics of mechanical turk. pp.
- IPEIROTIS, P. G. & GABRILOVICH, E. Quizz: Targeted crowdsourcing with a billion (potential) users. *Proceedings of the 23rd international conference on World wide web*, 2014. 143-154.
- ITOKO, T.;ARITA, S.;KOBAYASHI, M. & TAKAGI, H. Involving senior workers in crowdsourced proofreading. *International Conference on Universal Access in Human-Computer Interaction*, 2014. Springer, 106-117.
- JACKSON, S. A. & EKLUND, R. C. 2002. Assessing flow in physical activity: The flow state scale–2 and dispositional flow scale–2. *Journal of sport and exercise psychology*, 24(2), pp. 133-150. <https://doi.org/10.1123/jsep.24.2.133>
- JACOB, A.;FAATZ, A.;KNÜPPE, L. & TEUTEBERG, F. 2022. Understanding the effectiveness of gamification in an industrial work process: An experimental approach. *Business Process Management Journal*, 28(pp. 784–806. <https://doi.org/https://doi.org/10.1108/BPMJ-08-2021-0564>
- JAIN, A.;SARMA, A. D.;PARAMESWARAN, A. & WIDOM, J. 2017. Understanding workers, developing effective tasks, and enhancing marketplace dynamics. *VLDB Endowment*, 10(7), pp. 829-840. <https://doi.org/10.14778/3067421.3067431>
- JAWORSKI, R.;SELJAN, S. & DUNĐER, I. 2023. Four million segments and counting: Building an english-croatian parallel corpus through crowdsourcing using a novel gamification-based platform. *Information*, 14(4), pp. 226. <https://doi.org/10.3390/info14040226>
- JIANG, X.;ZHANG, J.;WU, M. & SHENG, V. S. 2024. Tifedcrowd: Federated crowdsourcing with time-controlled incentive. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 9(2), pp. 1514-1526. <https://doi.org/10.1109/TETCI.2024.3423346>
- JOHNSON, D. W. 2003. Social interdependence: Interrelationships among theory, research, and practice. *Am Psychol*, 58(11), pp. 934-945. <https://doi.org/10.1037/0003-066X.58.11.934>
- JOHNSON, D. W. & JOHNSON, R. T. 2005. New developments in social interdependence theory. *Genet Soc Gen Psychol Monogr*, 131(4), pp. 285-358. <https://doi.org/10.3200/MONO.131.4.285-358>
- JOHNSON, R. B.;ONWUEGBUZIE, A. J. & TURNER, L. A. 2007. Toward a definition of mixed methods research. *Journal of mixed methods research*, 1(2), pp. 112-133. <https://doi.org/10.1177/1558689806298224>
- KACORRI, H.;SHINKAWA, K. & SAITO, S. Capcap: An output-agreement game for video captioning. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015. 2814-2818.

- KALOKYRI, V.;BORGIDA, A. & MARIAN, A. One of us: A multiplayer web-based game for digital evidence acquisition of scripts through crowdsourcing. *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, 2023. 187-196.
- KANKAM, P. K. 2019. The use of paradigms in information research. *Library & Information Science Research*, 41(2), pp. 85-92. <https://doi.org/10.1016/j.lisr.2019.04.003>
- KAPENEKAKIS, I. & CHORIANOPOULOS, K. 2017. Citizen science for pedestrian cartography: Collection and moderation of walkable routes in cities through mobile gamification. *Human-centric Computing and Information Sciences*, 7(1), pp. <https://doi.org/10.1186/s13673-017-0090-9>
- KASHIAN, A.;RAJABIFARD, A. & RICHTER, K. F. Roadplex: A mobile vgi game to collect and validate data for pois. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2014. 55-60.
- KASHIMA, H.;OYAMA, S.;ARAI, H. & MORI, J. 2024. Trustworthy human computation: A survey. *Artificial Intelligence Review*, 57(12), pp. 322. <https://doi.org/10.1007/s10462-024-10974-1>
- KAUFMANN, N.;SCHULZE, T. & VEIT, D. (2011). More than fun and money. Worker motivation in crowdsourcing—a study on mechanical turk. In *Proceedings of Americas Conference on Information Systems (AMCIS)*. (pp. Association for Information Systems (AIS),
- KAVALIOVA, M.;VIRJEE, F.;MAEHLE, N. & KLEPPE, I. A. 2016. Crowdsourcing innovation and product development: Gamification as a motivational driver. *Cogent Business & Management*, 3(1), pp. 1128132. <https://doi.org/10.1080/23311975.2015.1128132>
- KAWAJIRI, R.;SHIMOSAKA, M. & KASHIMA, H. Steered crowdsensing: Incentive design towards quality-oriented place-centric crowdsensing. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2014. 691-701.
- KAZAI, G.;KAMPS, J. & MILIC-FRAYLING, N. Worker types and personality traits in crowdsourcing relevance labels. *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011. 1941-1944.
- KAZAI, G.;KAMPS, J. & MILIC-FRAYLING, N. The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. *Proceedings of the 21st ACM international conference on Information and knowledge management*, 2012. 2583-2586.
- KERBY, D. S. 2014. The simple difference formula: An approach to teaching nonparametric correlation1. *Comprehensive Psychology*, 3(pp. 11.IT.3.1. <https://doi.org/10.2466/11.It.3.1>
- KHOSLA, A.;JAYADEVAPRAKASH, N.;YAO, B. & LI, F.-F. Novel dataset for fine-grained image categorization: Stanford dogs. *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, 2011. Citeseer.
- KIM, J. & MONROY-HERNANDEZ, A. Storia: Summarizing social media content based on narrative theory using crowdsourcing. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 2016. 1018-1027.
- KIRKHAM, R.;SHEPHERD, C. & PLÖTZ, T. Blobsnake: Gamification of feature extraction for 'plug and play' human activity recognition. *ACM International Conference Proceeding Series*, 2015. 74-81.
- KITTUR, A.;NICKERSON, J. V.;BERNSTEIN, M.;GERBER, E.;SHAW, A.;ZIMMERMAN, J.;LEASE, M. & HORTON, J. The future of crowd work. *Proceedings of the 2013 conference on Computer supported cooperative work*, 2013. 1301-1318.

- KOIVISTO, J. & HAMARI, J. 2019. The rise of motivational information systems: A review of gamification research. *International Journal of Information Management*, 45(pp. 191-210. <https://doi.org/10.1016/j.ijinfomgt.2018.10.013>
- KONG, F.;SHI, B. & LU, Y. 2025. Incentive mechanism for task diffusion in social network-based mobile crowdsourcing. *IEEE Transactions on Computational Social Systems*, pp. 2140 - 2151. <https://doi.org/10.1109/TCSS.2025.3610409>
- KöSE, D. B.;MORSCHHEUSER, B. & HAMARI, J. 2019. Is it a tool or a toy? How user's conception of a system's purpose affects their experience and use. *International Journal of Information Management*, 49(pp. 461-474. <https://doi.org/10.1016/j.ijinfomgt.2019.07.016>
- KRATCHANOV, I. & LUCHEV, D. Gamified crowdsourcing in support of the semantic transformation of digital libraries. *Proceedings of the International Conference on Computer Systems and Technologies 2024*, 2024. 113-118.
- KUMATONGO, B. & MUZATA, K. K. 2021. Research paradigms and designs with their application in education. *Journal of Lexicography and Terminology* (Online ISSN 2664-0899. Print ISSN 2517-9306). 5(1), pp. 16-32.
- L'HEUREUX, A.;GROLINGER, K.;HIGASHINO, W. A. & CAPRETZ, M. A. M. A gamification framework for sensor data analytics. *Proceedings - 2017 IEEE 2nd International Congress on Internet of Things, ICIOT 2017*, 2017. 74-81.
- LACITY, M. C. & HIRSCHHEIM, R. 1993. The information systems outsourcing bandwagon. *MIT Sloan management review*, 35(1), pp. 73.
- LAKENS, D. 2022. Sample size justification. *Collabra: psychology*, 8(1), pp. 33267. <https://doi.org/10.1525/collabra.33267>
- LASO BAYAS, J.;SEE, L.;FRITZ, S.;STURN, T.;PERGER, C.;DÜRAUER, M.;KARNER, M.;MOORTHY, I.;SCHEPASCHENKO, D.;DOMIAN, D. & MCCALLUM, I. 2016. Crowdsourcing in-situ data on land cover and land use using gamification and mobile technology. *Remote Sensing*, 8(11), pp. <https://doi.org/10.3390/rs8110905>
- LATOZA, T. D. Crowdsourcing in software engineering: Models, motivations, and challenges. 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), 25-31 May 2019 2019. 301-301.
- LAW, E. & AHN, L. V. 2011. *Human computation*.
- LAW, E.;YIN, M.;GOH, J.;CHEN, K.;TERRY, M. A. & GAJOS, K. Z. (2016). Curiosity killed the cat, but makes crowdwork better. In *Proceedings of Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 4098-4110). Association for Computing Machinery, <https://doi.org/10.1145/2858036.2858144>
- LE, H. H.;NGUYEN, D. M. H.;BHATTI, O. S.;KOPACSI, L.;NGO, T. P.;NGUYEN, B. T.;BARZ, M. & SONNTAG, D. 2025. I-mpn: Inductive message passing network for efficient human-in-the-loop annotation of mobile eye tracking data. *Sci Rep*, 15(1), pp. 14192. <https://doi.org/10.1038/s41598-025-94593-y>
- LEAL, J. P. (2023). A game with a purpose for building crowdsourced semantic relations datasets for named entities. In *Proceedings of Science and Information Conference* (pp. Springer, Cham, https://doi.org/10.1007/978-3-031-37963-5_30
- LEE, J. J.;CEYHAN, P.;JORDAN-COOLEY, W. & SUNG, W. 2013a. Greenify: A real-world action game for climate change education. *Simulation & Gaming*, 44(2-3), pp. 349-365. <https://doi.org/10.1177/1046878112470539>
- LEE, T. Y.;DUGAN, C.;GEYER, W.;RATCHFORD, T.;RASMUSSEN, J.;SHAMI, N. S. & LUPUSHOR, S. Experiments on motivational feedback for crowdsourced workers.

- Proceedings of the International AAAI Conference on Web and Social Media, 2013b. 341-350.
- LENART-GANSINIEC, R.;CZAKON, W.;SUŁKOWSKI, Ł. & POCEK, J. 2022. Understanding crowdsourcing in science. *Review of Managerial Science*, 17(8), pp. 2797-2830. <https://doi.org/10.1007/s11846-022-00602-z>
- LENDAK, I. Simulation-based evaluation of a crowdsourced expert peer review system. 2019 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2019, 2019. 77-82.
- LEVAC, D.;COLQUHOUN, H. & O'BRIEN, K. K. 2010. Scoping studies: Advancing the methodology. *Implement Sci*, 5(pp. 69. <https://doi.org/10.1186/1748-5908-5-69>
- LI, Y.;LI, H.;HUANG, X.;XU, J.;HAN, Y. & XU, M. 2022. Utility-aware dynamic ridesharing in spatial crowdsourcing. *IEEE Transactions on Mobile Computing*, 23(2), pp. 1066-1079. <https://doi.org/10.1109/TMC.2022.3232215>
- LICHTENBERG, S.;LEMBCKE, T. B.;BRENING, M.;BRENDL, A. B. & TRANG, S. Can gamification lead to increase paid crowdworkers output? Proceedings of the 15th International Conference on Business Information Systems 2020 "Developments, Opportunities and Challenges of Digitization", WIRTSCHAFTSINFORMATIK 2020, 2020.
- LIN, Y. L. & DING, N. D. 2023. Competitive gamification in crowdsourcing-based contextual-aware recommender systems. *Int J Hum Comput Stud*, 177(pp. 103083. <https://doi.org/10.1016/j.ijhcs.2023.103083>
- LINARES, M.;POSTIGO, M.;CUADRADO, D.;ORTIZ-RUIZ, A.;GIL-CASANOVA, S.;VLADIMIROV, A.;GARCIA-VILLENA, J.;NUNEZ-ESCOBEDO, J. M.;MARTINEZ-LOPEZ, J.;RUBIO, J. M.;LEDESMA-CARBAYO, M. J.;SANTOS, A.;BASSAT, Q. & LUENGO-OROZ, M. 2019. Collaborative intelligence and gamification for on-line malaria species differentiation. *Malar J*, 18(1), pp. 21. <https://doi.org/10.1186/s12936-019-2662-9>
- LING, T. Y.;TEH, Y. B.;KIAT TING, G. K.;WONG, W. Z. & MAHMOOD, N. H. Project kappa: An augmented-reality based sensor data crowdsourcing platform for environmental monitoring. 2021 IEEE International Conference on Sensors and Nanotechnology, SENNANO 2021, 2021. 17-20.
- LINTOTT, C. J.;SCHAWINSKI, K.;SLOSAR, A.;LAND, K.;BAMFORD, S.;THOMAS, D.;RADDICK, M. J.;NICHOL, R. C.;SZALAY, A. & ANDREESCU, D. 2008. Galaxy zoo: Morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 389(3), pp. 1179-1189. <https://doi.org/10.1111/j.1365-2966.2008.13689.x>
- LITMAN, L.;ROBINSON, J. & ABBERBOCK, T. 2017. Turkprime.Com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behav Res Methods*, 49(2), pp. 433-442. <https://doi.org/10.3758/s13428-016-0727-z>
- LIU, A.;GUERRA, S.;FUNG, I.;MATUTE, G.;KAMAR, E. & LASECKI, W. Towards hybrid human-ai workflows for unknown unknown detection. Proceedings of The Web Conference 2020, 2020. 2432-2442.
- LIU, H.;NASIRIANY, S.;ZHANG, L.;BAO, Z. & ZHU, Y. 2025. Robot learning on the job: Human-in-the-loop autonomy and learning during deployment. *The International Journal of Robotics Research*, 44(10-11), pp. 1727-1742. <https://doi.org/10.1177/02783649241273901>
- LIU, J.;ZHU, H. & CHEN, X. Complicated-skills-based task assignment in spatial crowdsourcing. *International Conference on Web-Age Information Management*, 2016. Springer, 211-223.

- LIU, Y.;ALEXANDROVA, T.;NAKAJIMA, T. & LEHDONVIRTA, V. Mobile image search via local crowd: A user study. 2011 IEEE 17th International Conference on Embedded and Real-Time Computing Systems and Applications, 2011. IEEE, 109-112.
- LOCKE, E. A. & LATHAM, G. P. 1990. *A theory of goal setting & task performance*, Prentice-Hall, Inc.
- LU, Y.;MAO, X.;ZHOU, M.;ZHANG, Y.;LI, Z.;WANG, T.;YIN, G. & WANG, H. 2021. Motivation under gamification: An empirical study of developers' motivations and contributions in stack overflow. *IEEE transactions on software engineering*, 48(12), pp. 4947-4963. <https://doi.org/10.1109/TSE.2021.3130088>
- LUMSDEN, J.;SKINNER, A.;WOODS, A. T.;LAWRENCE, N. S. & MUNAFO, M. 2016. The effects of gamelike features and test location on cognitive test performance and participant enjoyment. *PeerJ*, 4(7), pp. e2184. <https://doi.org/10.7717/peerj.2184>
- MACAKOĞLU, Ş. S.;ÇİNAR, B. A. & PEKER, S. 2022. Kiyaslio: A gamified mobile crowdsourcing application for tracking price dispersion in the grocery retail market. *International Journal of Web Information Systems*, 18(1), pp. 55-75. <https://doi.org/10.1108/IJWIS-09-2021-0089>
- MACFARLAND, T. W. & YATES, J. M. 2016. Mann–whitney u test. *Introduction to nonparametric statistics for the biological sciences using r*. Springer.
- MADDALENA, E.;IBÁÑEZ, L.-D.;REEVES, N. & SIMPERL, E. 2023. Qrowdsmith: Enhancing paid microtask crowdsourcing with gamification and furtherance incentives. *ACM Transactions on Intelligent Systems and Technology*, 14(5), pp. 1-26. <https://doi.org/10.1145/3604940>
- MAHMOUDI-DEHAKI, M. & NASR-ESFAHANI, N. 2025. Automated vs. Manual linguistic annotation for assessing pragmatic competence in english classes. *Research Methods in Applied Linguistics*, 4(3), pp. 100253. <https://doi.org/10.1016/j.rmal.2025.100253>
- MAIRITTHA, N.;MAIRITTHA, T.;LAGO, P. & INOUE, S. 2021. Crowdact. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1), pp. 1-32. <https://doi.org/10.1145/3432222>
- MALATINI, S.;KLOPFENSTEIN, L. C. & BOGLIOLO, A. Gamification for crowdsourced data collection in mobile usability field studies. *ACM International Conference Proceeding Series*, 2017.
- MALONE, T. W.;LAUBACHER, R. & DELLAROCAS, C. 2009. Harnessing crowds: Mapping the genome of collective intelligence. pp.
- MARANI, F.;MARANI & KARKAL 2019. *Practical django 2 and channels 2*, Springer.
- MARTELARO, N.;LAKDAWALA, T.;CHEN, J. & HAMMER, J. Leveraging the twitch platform and gamification to generate home audio datasets. *Designing Interactive Systems Conference 2021*, 2021. 1765-1782.
- MARTELLA, R.;KRAY, C. & CLEMENTINI, E. 2015. A gamification framework for volunteered geographic information. *Lecture Notes in Geoinformation and Cartography*, 217(pp. 73-89. https://doi.org/10.1007/978-3-319-16787-9_5
- MARTIN, D.;CARPENDALE, S.;GUPTA, N.;HOßFELD, T.;NADERI, B.;REDI, J.;SIAHAAN, E. & WECHSUNG, I. 2017. Understanding the crowd: Ethical and practical matters in the academic use of crowdsourcing. *Evaluation in the crowd. Crowdsourcing and human-centered experiments*. Springer.
- MARTÍN-MARTÍN, A.;ORDUNA-MALEA, E.;THELWALL, M. & LÓPEZ-CÓZAR, E. D. 2018. Google scholar, web of science, and scopus: A systematic comparison of citations in 252

- subject categories. *Journal of informetrics*, 12(4), pp. 1160-1177. <https://doi.org/10.1016/j.joi.2018.09.002>
- MASON, A. D.;MICHALAKIDIS, G. & KRAUSE, P. J. Tiger nation: Empowering citizen scientists. 2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST), 2012. IEEE, 1-5.
- MASON, W. & SURI, S. 2012. Conducting behavioral research on amazon's mechanical turk. *Behav Res Methods*, 44(1), pp. 1-23. <https://doi.org/10.3758/s13428-011-0124-6>
- MASSUNG, E.;COYLE, D.;CATER, K. F.;JAY, M. & PREIST, C. Using crowdsourcing to support pro-environmental community activism. *Proceedings of the SIGCHI Conference on human factors in Computing systems*, 2013. 371-380.
- MATIN, A.;MADURO, M.;DE LEON PEREIRA, R. & TREMBLAY-SAVARD, O. Effect of timer, top score and leaderboard on performance and motivation in a human computing game. *ACM International Conference Proceeding Series*, 2020.
- MATSCHKE, C.;MOSKALIUK, J.;BOKHORST, F.;SCHÜMMER, T. & CRESS, U. 2014. Motivational factors of information exchange in social information spaces. *Computers in Human Behavior*, 36(pp. 549-558. <https://doi.org/10.1016/j.chb.2014.04.044>
- MAXWELL, J. A. & LOOMIS, D. M. 2003. Mixed methods design: An alternative approach. *Handbook of mixed methods in social and behavioral research*.
- MCAULEY, E.;DUNCAN, T. & TAMMEN, V. V. 1989. Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: A confirmatory factor analysis. *Research quarterly for exercise and sport*, 60(1), pp. 48-58.
- MCKIGHT, P. E. & NAJAB, J. 2010. Kruskal-wallis test. *The corsini encyclopedia of psychology*. John Wiley & Sons, Ltd.
- MCNEIL, A. J.;PARKS, K.;LIU, X.;JIANG, B.;COCO, J.;MCCOOL, K.;FABBRI, D.;DUHAIME, E. P.;DAWANT, B. M. & TKACZYK, E. R. 2023. Crowdsourcing skin demarcations of chronic graft-versus-host disease in patient photographs: Training versus performance study. *JMIR Dermatol*, 6(pp. e48589. <https://doi.org/10.2196/48589>
- MEISNER, A.;CASELDEN, D.;HUMPHREYS, A.;ROBBINS, G.;SCHAPER, E.;KIRKPATRICK, J. D.;SCHNEIDER, A.;JOHNSON, L. C.;KUCHNER, M. & FAHERTY, J. 2024. Cool neighbors: Combining artificial intelligence and citizen science to chart the sun's cosmic neighborhood. *Citizen Science: Theory and Practice*, 9(1), pp. <https://doi.org/10.5334/cstp.737>
- MEKLER, E. D.;BRÜHLMANN, F.;TUCH, A. N. & OPWIS, K. 2017. Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in human behavior*, 71(pp. 525-534. <https://doi.org/10.1016/j.chb.2015.08.048>
- MELENHORST, M.;NOVAK, J.;MICHEEL, I.;LARSON, M. & BOECKLE, M. Bridging the utilitarian-hedonic divide in crowdsourcing applications. *CrowdMM 2015 - Proceedings of the 4th International Workshop on Crowdsourcing for Multimedia*, co-located with MM 2015, 2015. 9-14.
- MENDOZA, K.;VILLALOBOS-DANIEL, V. E.;JÁUREGUI, A.;VALERO-MORALES, I.;HERNÁNDEZ-ALCARAZ, C.;ZACARÍAS-ALEJANDRO, N.;ALARCON-GUEVARA, R. O. & BARQUERA, S. 2024. Development of a crowdsourcing-and gamification-based mobile application to collect epidemiological information and promote healthy lifestyles in mexico. *Scientific Reports*, 14(1), pp. 6174. <https://doi.org/10.1038/s41598-024-56761-4>

- MENNIS, E. A. 2006. The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations. *Business Economics*, 41(4), pp. 63-65.
- MESBAH, S.; AROUS, I.; YANG, J. & BOZZON, A. Hybrideval: A human-ai collaborative approach for evaluating design ideas at scale. *Proceedings of the ACM Web Conference 2023*, 2023. 3837-3848.
- MICHAIL, A. M. & GAVALAS, D. Bucketfood: A crowdsourcing platform for promoting gastronomic tourism. *2019 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2019*, 2019. 9-14.
- MIYATA, A.; MURAYAMA, Y.; FURUTA, A.; OKUGAWA, K.; OCHIAI, K. & MURAYAMA, Y. Gamification strategies to improve the motivation and performance in accessibility information collection. *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 2022. 1-7.
- MOLASY, M.; ROSIENKIEWICZ, M.; HELMAN, J. & CHOLEWA, M. Gamification-based crowdsourcing as a tool for new product development in manufacturing companies. *International Conference on Intelligent Systems in Production Engineering and Maintenance*, 2023. Springer, 368-379.
- MORENO-ACEVEDO, S. A.; VASQUEZ-CORREA, J. C.; MARTÍN-DOÑAS, J. M. & ÁLVAREZ, A. Stream-based active learning for speech emotion recognition via hybrid data selection and continuous learning. *International Conference on Text, Speech, and Dialogue*, 2024. Springer, 105-117.
- MORENO-ALCAYDE, Y.; RUOTSALO, T.; LEIVA, L. A. & TRAVER, V. J. 2025. Affective annotation of videos from eeg-based crowdsourcing. *Pattern Analysis and Applications*, 28(3), pp. 146. <https://doi.org/10.1007/s10044-025-01476-z>
- MORGAN, D. L. 2007. Paradigms lost and pragmatism regained: Methodological implications of combining qualitative and quantitative methods. *Journal of mixed methods research*, 1(1), pp. 48-76. <https://doi.org/10.1177/2345678906292462>
- MORGAN, D. L. 2014. Pragmatism as a paradigm for social research. *Qualitative inquiry*, 20(8), pp. 1045-1053. <https://doi.org/10.1177/1077800413513733>
- MORRIS, R. R.; DONTCHEVA, M. & GERBER, E. M. 2012. Priming for better performance in microtask crowdsourcing environments. *IEEE Internet Computing*, 16(5), pp. 13-19. <https://doi.org/10.1109/MIC.2012.68>
- MORSCHHEUSER, B.; HAMARI, J. & KOIVISTO, J. Gamification in crowdsourcing: A review. *2016 49th Hawaii international conference on system sciences (HICSS)*, 2016. IEEE, 4375-4384.
- MORSCHHEUSER, B.; HAMARI, J.; KOIVISTO, J. & MAEDCHE, A. 2017. Gamified crowdsourcing: Conceptualization, literature review, and future agenda. *International Journal of Human-Computer Studies*, 106(pp. 26-43. <https://doi.org/10.1016/j.ijhcs.2017.04.005>
- MORSCHHEUSER, B.; HAMARI, J. & MAEDCHE, A. 2019. Cooperation or competition – when do people contribute more? A field experiment on gamification of crowdsourcing. *International Journal of Human-Computer Studies*, 127(pp. 7-24. <https://doi.org/10.1016/j.ijhcs.2018.10.001>
- NA, K. & HAN, K. 2023. How leaderboard positions shape our motivation: The impact of competence satisfaction and competence frustration on motivation in a gamified crowdsourcing task. *Internet Research*, 33(7), pp. 1-18. <https://doi.org/10.1108/INTR-12-2021-0897>

- NAGAI, Y.;HIYAMA, A.;MIURA, T. & HIROSE, M. T-echo: Promoting intergenerational communication through gamified social mentoring. *International Conference on Universal Access in Human-Computer Interaction*, 2014. Springer, 582-589.
- NG, L. H. X.;JAIDKA, K.;TAY, K. Y. & CHHAYA, N. 2025. Improving user behavior prediction: Leveraging annotator metadata in supervised machine learning models. *Proceedings of the ACM on Human-Computer Interaction*, 9(7), pp. 1-35. <https://doi.org/10.1145/3757484>
- NICHOLSON, S. 2014. A recipe for meaningful gamification. *Gamification in education and business*. Springer.
- NOSE, T. & HISHIYAMA, R. Analysis of self-tagging during conversational chat in multilingual gaming simulation. *Second International Conference on Future Generation Communication Technologies (FGCT 2013)*, 2013. IEEE, 81-86.
- NUMMENMAA, T.;LAATO, S.;CHAMBERS, P.;YRTTIMAA, T.;VASTARANTA, M.;BURUK, O. T. & HAMARI, J. 2025. Employing gamified crowdsourced close-range sensing in the pursuit of a digital twin of the earth. *International Journal of Human-Computer Interaction*, 41(8), pp. 1-17. <https://doi.org/10.1080/10447318.2024.2352922>
- O'BRIEN, H. L.;CAIRNS, P. & HALL, M. 2018. A practical approach to measuring user engagement with the refined user engagement scale (ues) and new ues short form. *International Journal of Human-Computer Studies*, 112(pp. 28-39. <https://doi.org/10.1016/j.ijhcs.2018.01.004>
- O'BRIEN, H. L. & TOMS, E. G. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American society for Information Science and Technology*, 59(6), pp. 938-955.
- OGAWA, H.;NISHIKAWA, H.;TOKUNAGA, T. & YOKONO, H. Gamification platform for collecting task-oriented dialogue data. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, 2020. 7084-7093.
- OLINSKY, S.;DESAI, P. M.;TURKAY, S.;HEITKEMPER, E. M.;MITCHELL, E. G.;MAMYKINA, L. & HWANG, M. L. Meals for monsters: A mobile application for the feasibility of gaming and social mechanisms. *Conference on Human Factors in Computing Systems - Proceedings*, 2021.
- OLTEANU, A.;CASTILLO, C.;DIAZ, F. & VIEWEG, S. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. *Proceedings of the international AAAI conference on web and social media*, 2014. 376-385.
- ORLIKOWSKI, W. J. & IACONO, C. S. 2001. Research commentary: Desperately seeking the “it” in it research—a call to theorizing the it artifact. *Information systems research*, 12(2), pp. 121-134. <https://doi.org/10.1287/isre.12.2.121.9700>
- ORTIZ, T. & TANG, V. Improving urban accessibility data collection through enhanced user experience in a crowdsourcing web application. *2024 4th International Conference on Information Communication and Software Engineering (ICICSE)*, 2024. IEEE, 55-59.
- PACKHAM, S. & SULEMAN, H. Crowdsourcing a text corpus is not a game. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015. 225-234.
- PALMA, L. D.;DIAO, Y. & LIU, A. 2024. Efficient version space algorithms for human-in-the-loop model development. *ACM Transactions on Knowledge Discovery from Data*, 18(3), pp. 1-49. <https://doi.org/10.1145/3637443>
- PAN, Y.;GAO, J.;DUAN, J.;SHI, J.;GUO, B.;LIANG, Y. & HU, Y. 2024. Pioneering cooperative air-ground instant delivery using uavs and crowdsourced couriers. *Proceedings of the ACM*

- on Interactive, Mobile, Wearable and Ubiquitous Technologies, 8(4), pp. 1-26.
<https://doi.org/10.1145/3699722>
- PAOLACCI, G. & CHANDLER, J. 2014. Inside the turk: Understanding mechanical turk as a participant pool. *Current directions in psychological science*, 23(3), pp. 184-188.
<https://doi.org/10.1177/0963721414531598>
- PAOLACCI, G.;CHANDLER, J. & IPEIROTIS, P. G. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5), pp. 411-419.
<https://doi.org/10.1017/S1930297500002205>
- PATEL, C.;HUSAIRI, M. A.;HAON, C. & OBEROI, P. 2023. Monetary rewards and self-selection in design crowdsourcing contests: Managing participation, contribution appropriateness, and winning trade-offs. *Technological Forecasting and Social Change*, 191(pp. 122447.
<https://doi.org/10.1016/j.techfore.2023.122447>
- PAULINO, D.;FERREIRA, J.;NETTO, A.;CORREIA, A.;RIBEIRO, J.;GUIMARÃES, D.;BARROSO, J. & PAREDES, H. Probing into the usage of task fingerprinting in web games to enhance cognitive personalization: A pilot gamified experience with neurodivergent participants. 2024 IEEE 12th International Conference on Serious Games and Applications for Health (SeGAH), 2024. IEEE, 1-8.
- PEER, E.;BRANDIMARTE, L.;SAMAT, S. & ACQUISTI, A. 2017. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70(pp. 153-163. <https://doi.org/10.1016/j.jesp.2017.01.006>
- PINTO, J. P. & VIANA, P. Improving youtube video retrieval by integrating crowdsourced timed metadata. *Journal of Intelligent and Fuzzy Systems*, 2019. 7207-7221.
- PINTO, L. F. S. & DOS SANTOS, C. D. 2018. Motivations of crowdsourcing contributors. *Innovation & Management Review*, 15(1), pp. 58-72. <https://doi.org/10.1108/inmr-02-2018-004>
- POSCH, L.;BLEIER, A.;LECHNER, C. M.;DANNER, D.;FLÖCK, F. & STROHMAIER, M. 2019. Measuring motivations of crowdworkers: The multidimensional crowdworker motivation scale. *ACM Transactions on Social Computing*, 2(2), pp. 1-34.
<https://doi.org/10.1145/3335081>
- POTHINENI, D.;MISHRA, P.;RASHEED, A. & SUNDARARAJAN, D. Incentive design to mould online behavior: A game mechanics perspective. *Proceedings of the First International Workshop on Gamification for Information Retrieval*, 2014. 27-32.
- PRANDI, C.;MELIS, A.;PRANDINI, M.;DELNEVO, G.;MONTI, L.;MIRRI, S. & SALOMONI, P. 2018. Gamifying cultural experiences across the urban environment. *Multimedia Tools and Applications*, 78(3), pp. 3341-3364. <https://doi.org/10.1007/s11042-018-6513-4>
- PRANDI, C.;ROCCETTI, M.;SALOMONI, P.;NISI, V. & NUNES, N. J. 2016. Fighting exclusion: A multimedia mobile app with zombies and maps as a medium for civic engagement and design. *Multimedia Tools and Applications*, 76(4), pp. 4951-4979.
<https://doi.org/10.1007/s11042-016-3780-9>
- PREIST, C.;MASSUNG, E. & COYLE, D. Competing or aiming to be average? Normification as a means of engaging digital volunteers. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 2014. 1222-1233.
- PRESTOPNIK, N.;CROWSTON, K. & WANG, J. 2017. Gamers, citizen scientists, and data: Exploring participant contributions in two games with a purpose. *Computers in Human Behavior*, 68(pp. 254-268. <https://doi.org/10.1016/j.chb.2016.11.035>

- PRPIĆ, J.;SHUKLA, P. P.;KIETZMANN, J. H. & MCCARTHY, I. P. 2015. How to work a crowd: Developing crowd capital through crowdsourcing. *Business Horizons*, 58(1), pp. 77-85. <https://doi.org/10.1016/j.bushor.2014.09.005>
- PURITAT, K. 2019. A gamified mobile-based approach with web monitoring for a crowdsourcing framework designed for urban problems related smart government: A case study of chiang mai, thailand. *International Journal of Interactive Mobile Technologies (iJIM)*, 13(12), pp. 55-66. <https://doi.org/10.3991/ijim.v13i12.10989>
- PURNOMO W.P, Y. S.;INDRIASARI, T. D.;ANINDITO, K.;ANDREAN, Y. & PRASETYO, J. G. 2019. Crimeid: Towards crime prevention and community safety in indonesia using mobile and web technology. *International Journal of Interactive Mobile Technologies (iJIM)*, 13(09), pp. 52-65. <https://doi.org/10.3991/ijim.v13i09.10172>
- QIU, S.;BOZZON, A.;BIRK, M. & GADIRAJU, U. 2021. Using worker avatars to improve microtask crowdsourcing. *ACM on Human-Computer Interaction*, 5(CSCW2), pp. 1-28. <https://doi.org/10.1145/3476063>
- QIU, S.;GADIRAJU, U. & BOZZON, A. Improving worker engagement through conversational microtask crowdsourcing. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020. 1-12.
- QUINN, A. J. & BEDERSON, B. B. Human computation: A survey and taxonomy of a growing field. *Proceedings of the SIGCHI conference on human factors in computing systems*, 2011. 1403-1412.
- RADDICK, J.;BRACEY, G. L. & GAY, P. L. (2010). Motivations of citizen scientists participating in galaxy zoo: A more detailed look. In *Proceedings of American Astronomical Society Meeting Abstracts #215* (pp. 467.04).
- RAMMSTEDT, B. & JOHN, O. P. 2007. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1), pp. 203-212. <https://doi.org/10.1016/j.jrp.2006.02.001>
- RAND, D. G. 2012. The promise of mechanical turk: How online labor markets can help theorists run behavioral experiments. *J Theor Biol*, 299(pp. 172-9. <https://doi.org/10.1016/j.jtbi.2011.03.004>
- REINHARD, P.;LI, M. M.;DICKHAUT, E.;REH, C.;PETERS, C. & LEIMEISTER, J. M. A conceptual model for labeling in reinforcement learning systems: A value co-creation perspective. *International Conference on Design Science Research in Information Systems and Technology*, 2023. Springer, 123-137.
- REMELHE, E.;CERQUEIRA, M.;FARIA, P. M. & PAIVA, S. 2022. Sustainable smart parking solution in a campus environment. *EAI Endorsed Transactions on Energy Web*, 9(39), pp. e2-e2. <https://doi.org/10.4108/ew.v9i39.1191>
- RENARD, D. & DAVIS, J. G. 2019. Social interdependence on crowdsourcing platforms. *Journal of Business Research*, 103(pp. 186-194. <https://doi.org/10.1016/j.jbusres.2019.06.033>
- RETELNY, D.;ROBASZKIEWICZ, S.;TO, A.;LASECKI, W. S.;PATEL, J.;RAHMATI, N.;DOSHI, T.;VALENTINE, M. & BERNSTEIN, M. S. Expert crowdsourcing with flash teams. *Proceedings of the 27th annual ACM symposium on User interface software and technology*, 2014. 75-85.
- RIAR, M.;MORSCHHEUSER, B.;ZARNEKOW, R. & HAMARI, J. 2022. Gamification of cooperation: A framework, literature review and future research agenda. *International Journal of Information Management*, 67(pp. 102549. <https://doi.org/10.1016/j.ijinfomgt.2022.102549>

- RIEGLER, M.;EG, R.;CALVET, L.;LUX, M.;HALVORSEN, P. & GRIWODZ, C. Playing around the eye tracker: A serious game based dataset. CEUR Workshop Proceedings, 2015. 34-40.
- RINNER, C.;KITTLER, H.;ROSENDAHL, C. & TSCHANDL, P. 2020. Analysis of collective human intelligence for diagnosis of pigmented skin lesions harnessed by gamification via a web-based training platform: Simulation reader study. J Med Internet Res, 22(1), pp. e15597. <https://doi.org/10.2196/15597>
- ROENGSAUT, B.;KUWABARA, K. & HUANG, H. H. Toward gamification of knowledge base construction. INISTA 2015 - 2015 International Symposium on Innovations in Intelligent SysTems and Applications, Proceedings, 2015.
- ROGSTADIUS, J.;KOSTAKOS, V.;KITTUR, A.;SMUS, B.;LAREDO, J. & VUKOVIC, M. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. Proceedings of the international AAAI conference on web and social media, 2011. 321-328.
- ROJAS, J. M.;WHITE, T. D.;CLEGG, B. S. & FRASER, G. Code defenders: Crowdsourcing effective tests and subtle mutants with a mutation testing game. Proceedings - 2017 IEEE/ACM 39th International Conference on Software Engineering, ICSE 2017, 2017. 677-688.
- ROKICKI, M.;CHELARU, S.;ZERR, S. & SIERSDORFER, S. Competitive game designs for improving the cost effectiveness of crowdsourcing. Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, 2014. 1469-1478.
- RÖNNEBERG, M. & KETTUNEN, P. 2023. A gamified map application utilising crowdsourcing engaged citizens to refine the quality and accuracy of cadastral index map border markers. International Journal of Digital Earth, 16(2), pp. 4726-4748. <https://doi.org/10.1080/17538947.2023.2279673>
- ROSS, J.;IRANI, L.;SILBERMAN, M. S.;ZALDIVAR, A. & TOMLINSON, B. 2010. Who are the crowdworkers? Shifting demographics in mechanical turk. *Chi'10 extended abstracts on human factors in computing systems*.
- ROUSE, A. C. (2010). A preliminary taxonomy of crowdsourcing. In Proceedings of 21st Australasian Conference on Information Systems (pp.
- RYAN, R. M. & DECI, E. L. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol, 55(1), pp. 68-78. <https://doi.org/10.1037//0003-066x.55.1.68>
- RYAN, R. M.;RIGBY, C. S. & PRZYBYLSKI, A. 2006. The motivational pull of video games: A self-determination theory approach. Motivation and emotion, 30(4), pp. 344-360. <https://doi.org/10.1007/s11031-006-9051-8>
- RZESZOTARSKI, J.;CHI, E.;PARITOSH, P. & DAI, P. Inserting micro-breaks into crowdsourcing workflows. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 2013.
- SADEGHIANASL, S.;HOFSTEDE, A. H. M. T.;SURIADI, S. & TURKAY, S. Collaborative and interactive detection and repair of activity labels in process event logs. Proceedings - 2020 2nd International Conference on Process Mining, ICPM 2020, 2020. 41-48.
- SADEGHIANASL, S.;HOFSTEDE, A. H. M. T.;WYNN, M. T.;TURKAY, S. & MYERS, T. 2021. Process activity ontology learning from event logs through gamification. IEEE Access, 9(pp. 165865-165880. <https://doi.org/10.1109/access.2021.3134915>

- SADEGHIANASL, S.;TER HOFSTEDE, A. H.;WYNN, M. T. & TÜRKAY, S. 2024. Humans-in-the-loop: Gamifying activity label repair in process event logs. *Engineering Applications of Artificial Intelligence*, 132(pp. 107875. <https://doi.org/10.1016/j.engappai.2024.107875>
- SAEED, A.;SPATHIS, D.;OH, J.;CHOI, E. & ETEMAD, A. 2025. Learning under label noise through few-shot human-in-the-loop refinement. *Sci Rep*, 15(1), pp. 4276. <https://doi.org/10.1038/s41598-025-87046-z>
- SAILER, M. & HOMNER, L. 2020. The gamification of learning: A meta-analysis. *Educational psychology review*, 32(1), pp. 77-112. <https://doi.org/10.1007/s10648-019-09498-w>
- SAITO, S.;WATANABE, T.;KOBAYASHI, M. & TAKAGI, H. Skill development framework for micro-tasking. *International Conference on Universal Access in Human-Computer Interaction*, 2014. Springer, 400-409.
- SAKAMOTO, M. & NAKAJIMA, T. Gamifying social media to encourage social activities with digital-physical hybrid role-playing. *International Conference on Social Computing and Social Media*, 2014. Springer, 581-591.
- SAMMUT, C. & WEBB, G. I. 2011. *Encyclopedia of machine learning*, Springer Science & Business Media.
- SARAVANOS, A.;ZERVOUDAKIS, S.;ZHENG, D.;STOTT, N.;HAWRYLUK, B. & DELFINO, D. (2021). The hidden cost of using amazon mechanical turk for research. In *Proceedings of International Conference on Human-Computer Interaction* (pp. https://doi.org/10.1007/978-3-030-90238-4_12
- SAUNDERS, B.;SIM, J.;KINGSTONE, T.;BAKER, S.;WATERFIELD, J.;BARTLAM, B.;BURROUGHS, H. & JINKS, C. 2018. Saturation in qualitative research: Exploring its conceptualization and operationalization. *Qual Quant*, 52(4), pp. 1893-1907. <https://doi.org/10.1007/s11135-017-0574-8>
- SEABORN, K. & FELS, D. I. 2015. Gamification in theory and action: A survey. *International Journal of human-computer studies*, 74(pp. 14-31. <https://doi.org/10.1016/j.ijhcs.2014.09.006>
- SEGUNDO DÍAZ, R. L.;ROVELO RUIZ, G.;BOUZOUITA, M. & CONINX, K. 2022. Building blocks for creating enjoyable games—a systematic literature review. pp. <https://doi.org/10.1016/j.ijhcs.2021.102758>
- SEONG, E. & KIM, S. Designing a crowdsourcing system for the elderly: A gamified approach to speech collection. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020. 1-9.
- SHADISH, W. R.;COOK, T. D. & CAMPBELL, D. T. 2002. *Experimental and quasi-experimental designs for generalized causal inference*, Houghton, Mifflin and Company.
- SHANK, D. B. 2016. Using crowdsourcing websites for sociological research: The case of amazon mechanical turk. *The American Sociologist*, 47(1), pp. 47-55. <https://doi.org/10.1007/s12108-015-9266-9>
- SHAPIRO, S. S. & WILK, M. B. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4), pp. 591-611. <https://doi.org/10.1093/biomet/52.3-4.591>
- SHENG, L. Y. Modelling learning from ingress (google's augmented reality social game). 2013 IEEE 63rd annual conference international council for education media (ICEM), 2013. IEEE, 1-8.
- SIAMTANIDOU, E.;VRYSIS, L.;VRYZAS, N. & DIMOULAS, C. A. 2025. Gamified engagement for data crowdsourcing and ai literacy: An investigation in affective communication through speech emotion recognition. *Societies*, 15(3), pp. 54. <https://doi.org/10.3390/soc15030054>

- SIGALA, M. 2015. The application and impact of gamification funware on trip planning and experiences: The case of tripadvisor's funware. *Electronic markets*, 25(pp. 189-209. <https://doi.org/10.1007/s12525-014-0179-1>
- SNIJDERS, R.;DALPIAZ, F.;BRINKKEMPER, S.;HOSSEINI, M.;ALI, R. & ÖZÜM, A. Refine: A gamified platform for participatory requirements engineering. 1st International Workshop on Crowd-Based Requirements Engineering, CrowdRE 2015 - Proceedings, 2015. 1-6.
- SNOW, R.;O'CONNOR, B.;JURAFSKY, D. & NG, A. Y. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. *Proceedings of the 2008 conference on empirical methods in natural language processing*, 2008. 254-263.
- SOLIMAN, W. & TUUNAINEN, V. K. 2015. Understanding continued use of crowdsourcing systems: An interpretive study. *Journal of Theoretical and Applied Electronic Commerce Research*, 10(1), pp. 1-18. <https://doi.org/10.4067/S0718-18762015000100002>
- SONG, D.;TAVARES, A.;PINTO, S. & XU, H. 2017. Setting engineering students up for success in the 21st century: Integrating gamification and crowdsourcing into a cdio-based web design course. *EURASIA Journal of Mathematics, Science and Technology Education*, 13(7), pp. 3565-3585. <https://doi.org/10.12973/eurasia.2017.00745a>
- SPATHARIOTI, S. E. 2020. *Designing effective interfaces for motivating engagement in crowdsourced image labeling*. Northeastern University.
- SPICHER, N.;WESEMEYER, T. & DESERNO, T. M. 2024. Crowdsourcing image segmentation for deep learning: Integrated platform for citizen science, paid microtask, and gamification. *Biomed Tech (Berl)*, 69(3), pp. 293-305. <https://doi.org/10.1515/bmt-2023-0148>
- SPINDELDREHER, K. & SCHLAGWEIN, D. What drives the crowd? A meta-analysis of the motivation of participants in crowdsourcing. *PACIS*, 2016. 119.
- SPITZ, R.;PEREIRA JUNIOR, C.;QUEIROZ, F.;LEITE, L. C.;DAM, P. & REZENDE, A. C. 2018. Gamification, citizen science, and civic technologies: In search of the common good. *Strategic Design Research Journal*, 11(3), pp. 263-273. <https://doi.org/10.4013/sdrj.2018.113.11>
- STAMATIADOU, M. E.;VRYZAS, N.;VRYSSIS, L. & DIMOULAS, C. Beemate the game: A hunting treasure serious game for raising awareness and audience engagement in air pollution monitoring. 2023 8th International Conference on Smart and Sustainable Technologies (SpliTech), 2023. IEEE, 1-4.
- STARKWEATHER, J. & MOSKE, A. K. 2011. Multinomial logistic regression.
- STEWART, N.;CHANDLER, J. & PAOLACCI, G. 2017. Crowdsourcing samples in cognitive science. *Trends Cogn Sci*, 21(10), pp. 736-748. <https://doi.org/10.1016/j.tics.2017.06.007>
- STRAUB, T.;GIMPEL, H.;TESCHNER, F. & WEINHARDT, C. 2015. How (not) to incent crowd workers. *Business & Information Systems Engineering*, 57(3), pp. 167-179. <https://doi.org/10.1007/s12599-015-0384-2>
- SUEN, T. Y.;CHEUNG, S. K.;WANG, F. L. & HUI, J. Y. 2022. Effects of intrinsic and extrinsic motivational factors on employee participation in internal crowdsourcing initiatives in china. *Sustainability*, 14(14), pp. 8878. <https://doi.org/10.3390/su14148878>
- SWEETSER, P. & WYETH, P. 2005. Gameflow: A model for evaluating player enjoyment in games. *Computers in Entertainment (CIE)*, 3(3), pp. 3-3. <https://doi.org/10.1145/1077246.1077253>
- SWELLER, J. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2), pp. 257-285.
- TAHMASBI, N. & STERNBERG, H. Get the crowd going - an empirical analysis toward a crowdsourcing engagement model. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2019. 729-738.

- TANG, W.;YIN, M. & HO, C.-J. Leveraging peer communication to enhance crowdsourcing. The World Wide Web Conference, 2019. 1794-1805.
- TAPSCOTT, D. & WILLIAMS, A. D. 2008. *Wikinomics: How mass collaboration changes everything*, Portfolio.
- TINATI, R.;LUCZAK-ROESCH, M.;SIMPERL, E. & HALL, W. 2017. An investigation of player motivations in eyewire, a gamified citizen science project. *Computers in Human Behavior*, 73(pp. 527-540. <https://doi.org/10.1016/j.chb.2016.12.074>
- TOCCHETTI, A.;TALENTI, S. M. & BRAMBILLA, M. " My lockdown escape": A data collection approach based on gamification and crowdsourcing for subjective perspectives, self-empathy, and memories about past experiences. *Proceedings of the 3rd Empathy-Centric Design Workshop: Scrutinizing Empathy Beyond the Individual*, 2024. 14-20.
- TSAI, T. H. & LEE, L. C. 2017. Location-based augmented reality for exploration of cultural heritage. *ICIC Express Letters, Part B: Applications*, 8(12), pp. 1577-1583. <https://doi.org/10.24507/icicelb.08.12.1577>
- TSVETKOVA, M.;YASSERI, T.;MEYER, E. T.;PICKERING, J. B.;ENGEL, V.;WALLAND, P.;LÜDERS, M.;FØLSTAD, A. & BRAVOS, G. 2017. Understanding human-machine networks: A cross-disciplinary survey. *ACM Computing Surveys (CSUR)*, 50(1), pp. 1-35. <https://doi.org/10.1145/3039868>
- UHLGREN, V.-V.;LAATO, S.;HAMARI, J. & NUMMENMAA, T. Gamification to motivate the crowdsourcing of dynamic nature data: A field experiment in northern europe. *Proceedings of the 27th International Academic Mindtrek Conference*, 2024. 230-234.
- USTALOV, D.;SAVAGE, S.;VAN BERKEL, N. & LIU, Y. 4th crowd science workshop-candle: Collaboration of humans and learning algorithms for data labeling. *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 2023. 1268-1268.
- UZUN, A.;LEHMANN, L.;GEISMAR, T. & KÜPPER, A. Turning the openmobilenetwork into a live crowdsourcing platform for semantic context-aware services. *Proceedings of the 9th International Conference on Semantic Systems*, 2013. 89-96.
- VALENTINE, M. A.;RETELNY, D.;TO, A.;RAHMATI, N.;DOSHI, T. & BERNSTEIN, M. S. Flash organizations: Crowdsourcing complex work by structuring crowds as organizations. *Proceedings of the 2017 CHI conference on human factors in computing systems*, 2017. 3523-3537.
- VANDEN BERGHEN, B.;VANERMEN, I. & VRANKEN, L. 2024. Citizen scientists: Unveiling motivations and characteristics influencing initial and sustained participation in an agricultural project. *PLoS One*, 19(5), pp. e0303103. <https://doi.org/10.1371/journal.pone.0303103>
- VASILESCU, B.;SEREBRENIK, A.;DEVANBU, P. & FILKOV, V. How social q&a sites are changing knowledge sharing in open source software communities. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 2014. 342-354.
- VERGARA-BORGE, F.;LÓPEZ-DE-IPÍÑA, D.;EMALDI, M.;OLIVARES-RODRÍGUEZ, C.;KHAN, Z. & SOOMRO, K. 2025. Gamifying engagement in spatial crowdsourcing: An exploratory mixed-methods study on gamification impact among university students. *Systems*, 13(7), pp. 519. <https://doi.org/10.3390/systems13070519>
- VON AHN, L. Human computation. *2008 IEEE 24th international conference on data engineering*, 2008. IEEE, 1-2.

- WALTER, V.;KÖLLE, M. & COLLMAR, D. A gamification approach for the improvement of paid crowd-based labelling of geospatial data. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2022. 113-120.
- WANG, J.;IPEIROTIS, P. G. & PROVOST, F. 2017. Cost-effective quality assurance in crowd labeling. *Information Systems Research*, 28(1), pp. 137-158. <https://doi.org/10.1287/isre.2016.0661>
- WANG, J.;LIU, Y. & CHEN, Y. Forecast aggregation via peer prediction. *Proceedings of the AAAI conference on human computation and crowdsourcing*, 2021. 131-142.
- WANG, W. S. & LEE, L. C. Development of a fake news checking crowdsourcing platform consisting of smart contracts combined with gamification. 2021 IEEE International Conference on Consumer Electronics-Taiwan, ICCE-TW 2021, 2021.
- WEBER, M.;RIAR, M. & MORSCHHEUSER, B. (2023). Is adaptive gamification just a theoretical fairytale? An experiment in a text-based adventure game for data crowdsourcing. In *Proceedings of Hawaii International Conference on System Sciences* (pp. <https://doi.org/10.24251/HICSS.2023.138>
- WEI, L. K. & ANWAR, T. Analysis of motivation approach in mobile crowdsensing application. 6th ICT International Student Project Conference: Elevating Community Through ICT, ICT-ISPC 2017, 2017. 1-4.
- WELINDER, P.;BRANSON, S.;PERONA, P. & BELONGIE, S. (2010). The multidimensional wisdom of crowds. In *Proceedings of Advances in neural information processing systems* (pp. Curran Associates / NeurIPS,
- WENG, J.;XIE, H.;FENG, Y.;WANG, R.;YE, Y.;HUANG, P. & ZHENG, X. Effects of gamification elements on crowdsourcing participation: The mediating role of justice perceptions. *Proceedings of the International Conference on Electronic Business (ICEB)*, 2019. 199-211.
- WERNBACHER, T.;PFEIFFER, A.;GEBETSROITHER-GERINGER, E.;GOELS, M.;WORSTER, J.;MEIBNER, E.;GRAF, A.;STOLLNBERGER, R.;GEYER, R.;SCHMIDT, R. R. & SERADA, A. Hotcity—a gamified token system for reporting waste heat sources. *Lecture Notes in Networks and Systems*, 2022. 15-27.
- WIETHOF, C.;ROOCKS, T. & BITTNER, E. A. Gamifying the human-in-the-loop: Toward increased motivation for training ai in customer service. *International Conference on Human-Computer Interaction*, 2022. Springer, 100-117.
- WILLETT, K. W.;LINTOTT, C. J.;BAMFORD, S. P.;MASTERS, K. L.;SIMMONS, B. D.;CASTEELS, K. R.;EDMONDSON, E. M.;FORTSON, L. F.;KAVIRAJ, S. & KEEL, W. C. 2013. Galaxy zoo 2: Detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 435(4), pp. 2835-2860. <https://doi.org/10.1093/mnras/stt1458>
- WU, Q.;LI, Y.;YAN, J.;ZHANG, M.;XU, J. & XU, M. 2024. Adaptive task assignment in spatial crowdsourcing: A human-in-the-loop approach. *IEEE Transactions on Mobile Computing*, 24(4), pp. 2726 - 2739. <https://doi.org/10.1109/TMC.2024.3501734>
- WYSS, E. L.;SAAL, J.;RAPP, A.;SCHMUNK, S.;DIETZ, N.;DUNKELMANN, L.;KORB, N. & NAUNHEIM, A. 2025. The dynamics of processing knowledge in citizen science projects. Enhancing participation in public citizen science formats. *Citizen Science: Theory and Practice*, 10(1), pp. 24. <https://doi.org/10.5334/cstp.779>
- XIN, D.;MA, L.;LIU, J.;MACKE, S.;SONG, S. & PARAMESWARAN, A. Accelerating human-in-the-loop machine learning: Challenges and opportunities. *Proceedings of the second workshop on data management for end-to-end machine learning*, 2018. 1-4.

- XU, G.;DING, W.;TANG, J.;YANG, S.;HUANG, G. Y. & LIU, Z. Learning effective embeddings from crowdsourced labels: An educational case study. 2019 IEEE 35th International Conference on Data Engineering (ICDE), 2019. IEEE, 1922-1927.
- XU, Q. & ZHENG, R. Mobibee: A mobile treasure hunt game for location-dependent fingerprint collection. UbiComp 2016 Adjunct - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2016. 1472-1477.
- XU, R.;LI, G. & SHENG, V. S. 2025a. Detecting worker collusion in crowdsourcing: Methods, datasets, and challenges. Expert Systems with Applications, 294(pp. 128767. <https://doi.org/10.1016/j.eswa.2025.128767>
- XU, S.;ZHANG, Y.;RESNICK, P. & SCHOENEBECK, G. Spot check equivalence: An interpretable metric for information elicitation mechanisms. Proceedings of the ACM Web Conference 2024, 2024. 276-287.
- XU, Z.;SCHEID, B. H.;CONRAD, E. C.;DAVIS, K. A.;GANGULY, T.;GELFAND, M. A.;GUGGER, J. J.;JIANG, X.;LAROCQUE, J. J.;OJEMANN, W. K. S.;SINHA, S. R.;WALDMAN, G. J.;WAGENAAR, J.;SINHA, N. & LITT, B. 2025b. Annotating neurophysiologic data at scale with optimized human input. J Neural Eng, 22(4), pp. <https://doi.org/10.1088/1741-2552/ade402>
- YACOBSON, E.;TODA, A. M.;CRISTEA, A. I. & ALEXANDRON, G. (2021). Participatory design of feedback mechanism in a physics blended-learning environment. In Proceedings of ECTEL 2021: AI for Blended Learning: Empowering Teachers in Real Classrooms (pp. CEUR Workshop Proceedings, Bozen-Bolzano, Italy.
- YAKUSHIN, D. & LEE, J.-H. Cooperative robot software development through the internet. 2014 IEEE/SICE International Symposium on System Integration, 2014. IEEE, 577-582.
- YANG, Y.;SHIN, A.;KIM, N.;WOO, H.;CHUNG, J. J. Y. & SONG, J. Y. Find the bot!: Gamifying facial emotion recognition for both human training and machine learning data collection. Proceedings of the CHI Conference on Human Factors in Computing Systems, 2024. 1-20.
- YANG, Y.;TRUONG, N. D.;MAHER, C.;NIKPOUR, A. & KAVEHEI, O. 2022. Continental generalization of a human-in-the-loop ai system for clinical seizure recognition. Expert Systems with Applications, 207(pp. 118083. <https://doi.org/10.1016/j.eswa.2022.118083>
- YE, Y.;LIU, C. H.;DAI, Z.;ZHAO, J.;YUAN, Y.;WANG, G. & TANG, J. Exploring both individuality and cooperation for air-ground spatial crowdsourcing by multi-agent deep reinforcement learning. 2023 IEEE 39th International Conference on Data Engineering (ICDE), 2023. IEEE, 205-217.
- YU, H.;LIU, Y.;WEI, X.;ZHENG, C.;CHEN, T.;YANG, Q. & PENG, X. Fair and explainable dynamic engagement of crowd workers. IJCAI, 2019. 6575-6577.
- ZHAI, H.;YAN, H. Y.;ZHOU, J. Y.;LIU, J.;XIE, Q. W.;SHEN, L. J.;CHEN, X. & HAN, H. 2025. Intebomb: Integrating generic object tracking and segmentation with pose estimation for animal behavior analysis. Zool Res, 46(2), pp. 355-369. <https://doi.org/10.24272/j.issn.2095-8137.2024.268>
- ZHANG, S.;MU, H. & LIU, T. Improving accuracy and generalizability via multi-modal large language models collaboration. 2024 International Joint Conference on Neural Networks (IJCNN), 2024. IEEE, 1-8.
- ZHAO, M. & VAN DER HOEK, A. A brief perspective on microtask crowdsourcing workflows for interface design. 2015 IEEE/ACM 2nd International Workshop on CrowdSourcing in Software Engineering, 2015. IEEE, 45-46.

- ZHAO, Y.;GONG, X. & MAO, S. Truthful incentive mechanism for federated learning with crowdsourced data labeling. IEEE INFOCOM 2023-IEEE Conference on Computer Communications, 2023. IEEE, 1-10.
- ZHAO, Y.;ZHENG, K.;YIN, H.;LIU, G.;FANG, J. & ZHOU, X. 2020. Preference-aware task assignment in spatial crowdsourcing: From individuals to groups. IEEE Transactions on Knowledge and Data Engineering, 34(7), pp. 3461-3477. <https://doi.org/10.1109/TKDE.2020.3021028>
- ZHAO, Y. & ZHU, Q. 2014a. Evaluation on crowdsourcing research: Current status and future direction. Information Systems Frontiers, 16(3), pp. 417-434. <https://doi.org/10.1007/s10796-012-9350-4>
- ZHAO, Y. C. & ZHU, Q. 2014b. Effects of extrinsic and intrinsic motivation on participation in crowdsourcing contest. Online Information Review, 38(7), pp. 896-917. <https://doi.org/10.1108/OIR-08-2014-0188>
- ZHENG, H.;LI, D. & HOU, W. 2011. Task design, motivation, and participation in crowdsourcing contests. International Journal of Electronic Commerce, 15(4), pp. 57-88. <https://doi.org/10.2753/JEC1086-4415150402>
- ZHOU, M. & WANG, H. 2018. The role of rationality in motivating participation in social movements: The case of anti-japanese demonstrations in china. Rationality and Society, 30(1), pp. 155-186. <https://doi.org/10.1177/1043463117754078>
- ZHOU, X.;TANG, J.;WANG, T. & MA, Y. Investigating the impacts of task characteristics in gamified citizen science. Proceedings of the 21st Pacific Asia Conference on Information Systems: "Societal Transformation Through IS/IT", PACIS 2017, 2017.
- ZHUANG, M. & GADIRAJU, U. (2019). In what mood are you today? In Proceedings of Proceedings of the 10th ACM Conference on Web Science (pp. 373-382). Association for Computing Machinery, <https://doi.org/10.1145/3292522.3326010>
- ZUCHOWSKI, O.;POSEGGA, O.;SCHLAGWEIN, D. & FISCHBACH, K. 2016. Internal crowdsourcing: Conceptual framework, structured review, and research agenda. Journal of Information Technology, 31(2), pp. 166-184. <https://doi.org/10.1057/jit.2016.14>

Appendix A

A.1 Ethics Approval for Study 1

11/28/25, 7:49 PM

CIS Ethics Approval System – Computer and Information Sciences – local.cis



Computer and Information Sciences - local.cis

departmental information for staff and students

[Home](#) | [Ethics](#) | [Events](#) | [Safety](#) | [IT Support](#) | [Teaching](#) | [Utilities](#)

[Browse Home](#) / [Utilities](#) / CIS Ethics Approval System

CIS Ethics Approval System

You are Huda Banuqitah (Research Student - 202084148)

[Return to Main](#)

Application ID: 1399

Title of research:
Inequality of contribution in crowdsourcing

Summary of research (short overview of the background and aims of this study):

As part of my PhD, I am investigating the inequality of contribution in crowdsourcing workers tasks. Classification of Twitter content, Image classification, and Survey are tasks which are part of an ongoing research that examine and analyze the users' contributions in different task types on crowdsourcing platforms. This can help us to determine the inequality of contribution that can be in such platforms of participants and then develop a strategy to retain the workers, so not to drop out of Crowdsourcing platforms. The tasks that the Participants or workers from Amazon Mechanical Turk (AMT) and Appen crowdsourcing platform will perform are:

- ☐ Classification task is asking the workers to categorize each tweet into one of the nine categories.
- ☐ Image Classification task is asking the workers to categorize dog images into its breeds.
- ☐ Survey task contains two part the first part asking the participant about demographic information such as gender, age, field of study and employment and the second part is asking questions about the E-learning effect on some practical courses in current pandemic

All of the three tasks will be outsourcing to participant in crowdsourcing AMT and Appen platforms website. The tasks were chosen as representative of the most common tasks on crowdsourcing platforms.

How will participants be recruited?

Limited number of participant (around 100 participants) are needed to conduct each tasks in this study. For ecological validity, the participant will be recruited to the study in the normal way via crowdsourcing service. The tasks will be posting in the crowdsourcing website stating the detailed of the task. Participants who will accept the task posted on the website can complete the task and the researchers will be notify from the platform via email.

We have estimated that the tasks will take about 10 minutes to complete per participant. After discussion with colleagues, we have agreed US\$8.50 a fair rate of payment that is within the realms of normal payment by a University team. This equates to US\$1.5 per task including a little reading time. To safeguard against spamming attempts a minimal performance level is typical in crowdsourcing – we have agreed a low level for payment and will pay all workers who correctly answer at least ¼ of questions.

there are no explicit inclusion or exclusion criterias except AMT rules restrict to adults (18 and over) who registered on a crowdsourcing platform such as AMT or Appen :

https://www.mturk.com/help/who_completes_the_tasks_and_how

Can you add some detail about the inclusion and exclusion criteria please of participants so I can get a sense of which population you are working with to ensure there are no specific considerations to review?

What will the participants be told about the proposed research study? Either upload or include a copy of the briefing notes issued to participants. In particular this should include details of yourself, the context of the study and an overview of the data that you plan to collect, your supervisor, and contact details for the Departmental Ethics Committee.

PDF File: [View document](#)

On Amazon Mechanical Turk and Appen Platforms, the following will be used to advertise the tasks :

1. Classification Task: The goal of this task is to categorize each tweet into one of the nine categories given below. These tweets were collected during some crisis emergency situation. Please, read the text carefully and decide only based on the content of the tweet, you do not need to access any of the URLs in tweets. The task will be followed by a short questionnaire as part of our research.
2. Image Classification Tasks: The goal of this task is to view different dogs images and categorize them each based on their breeds. The task will be followed by a short questionnaire as part of our research.
3. Survey Task: The goal of this survey is to analyze the student's point of view related to E-Learning of some practical course in pandemic situation. The task will be followed by a short additional questionnaire as part of our research.

The link For the short Questionnaire on Qualtrics :https://strathsci.qualtrics.com/jfe/form/SV_0UsBWQ9H5uDGfDU

How will consent be demonstrated? Either upload or include here a copy of the consent form/instructions issued to participants. It is particularly important that you make the rights of the participants to freely withdraw from the study at any point (if they begin to feel stressed for example), nor feel under any pressure or obligation

<https://local.cis.strath.ac.uk/wp/extras/ethics/?view=1399>

1/2

to complete the study, answer any particular question, or undertake any particular task. Their rights regarding associated data collected should also be made explicit.
PDF File: None.

The first page of the Tasks will include the consent details, which will make it clear for participate that:

- Their participation they will be anonymized and kept confidential.
- Their participation is voluntary.
- Participants have the right to withdraw at any time before submitting their responses.
- The survey questions should not cause any discomfort. Yet, participant have the right not to answer any question that they feel uncomfortable with.
- The electronic data of the tasks will be stored on crowdsourcing platforms servers. The survey data will be stored on Qualtrics server.
- If participants have any concern regarding their participation in this study, they can contact the researcher or the Department Ethics Committee via email

What will participants be expected to do? Either upload or include a copy of the instructions issued to participants along with a copy of or link to the survey, interview script or task description you intend to carry out.

PDF File: None.

PDF File: None.

The participant from crowdsourcing platforms will be allowed to submit up to 20 tasks plus one post-task questionnaire.

What data will be collected and how will it be captured and stored? In particular indicate how adherence to the Data Protection Act and the General Data Protection Regulation (GDPR) will be guaranteed and how participant confidentiality will be handled.

- No sensitive personal or identifiable information will be collected except for the worker's platform ID. An anonymous data set will be created once the study is complete by replacing worker IDs with randomly generated ones.
- To allow deletion upon request a table of real platform IDs to random IDs will be kept on secure department storage and deleted at end of PhD.
- Identifiable data will be stored on Qualtrics server and Crowd platform servers until the end of the study.

How will the data be processed? (e.g. analysed, reported, visualised, integrated with other data, etc.) Please pay particular attention to describing how personal or sensitive data will be handled and how GDPR regulations will be met.

- The collected data will be used only for this investigation.
- The data will be secure and restricted to the researcher and the researcher's supervisors.
- The data will be processed analyzed based on the researcher goals, using statistical tools as SPSS, there will be no attempt to deanonymize the data.
- The electronic data of the tasks will be stored crowdsourcing platforms servers. The survey data will be stored on Qualtrics server and a copy will be stored on the researcher's strathcloud for analysis purposes.
- The data will be compared with other sources such as related results from related research papers.
- A report of finding of this study might be published as a journal article or conference proceeding. The collected data will not include any information that could identify individual participant.

How and when will data be disposed of? Either upload a copy of your data management plan or describe how data will be disposed.

PDF File: None.

Crowd platform and Qualtrics (the non-anonymous data) will be deleted within 30 days of the end of the study.

The translation table for worker IDs to random IDs will be deleted at end of study on deletion of department account.

No personal data is recorded in the anonymous data thus no disposal procedure is required.

Supervisor's name (students need to complete/staff should probably put N/A).

☐ I confirm that my supervisor has seen and approved both my planned study and this associated ethics application.

A.2 Users ‘Consent Form of Study 1

I am a Ph.D. student at Strathclyde University (Scotland, UK), aiming to make crowdsourcing platforms safe and fair. My research aims to improve the overall quality of crowdsourcing platforms by developing techniques that help to engage the workers in the crowdsourcing platforms.

Your collaboration is highly appreciated, and your input is very important to us for scientific purposes. We will ask you to answer ten short, single question tasks and to complete a short survey, the tasks are described in the instructions along with examples of questions and answers.

Your participation is entirely voluntary, and you are free to discontinue at any point. Payment will be made using standard TURK policy. We have included minimal standards for responses so please complete the tasks as accurately as possible. (If your answers are below the minimum number of correct answers or your Qualtrics-survey code is invalid then we will not make a payment). It will take around five minutes to complete everything and can be done on a mobile phone, table or personal or computer. All existing Amazon Mechanical Turk workers are welcome to participate.

The collected data will be used only for this investigation, and details may be used in academic publications, but your identity will be protected. No sensitive personal, or identifiable information will be collected except for your worker’s platform ID. An anonymous data set will be created once the study is complete by replacing worker IDs with randomly generated ones. You can request the deletion of your data within two months by post or direct email to me. After this your data will have been anonymised, so that it will not then be possible to identify your data in order to delete it.

A.3 Survey Form of Study 1

General information about you

Q1: Gender: Are you...

- ☐ Male
- ☐ Female
- ☐ Non-Binary
- ☐ Prefer to Self-Describe
- ☐ Prefer not to say

Q2: Age: What is your age?

- ☐ 18-24 years old
- ☐ 25-34 years old
- ☐ 35-44 years old
- ☐ 45-54 years old
- ☐ 55-64 years old
- ☐ Over 64 years old
- ☐ Prefer not to say

Q3: Please specify your ethnicity.

- ☐ White
- ☐ Hispanic or Latino
- ☐ Black or African American
- ☐ Asian
- ☐ Native Hawaiian or Pacific Islander
- ☐ Other _____
- ☐ Prefer not to say

Q4: Education: What is the highest level of education you have completed?

- ☐ If currently enrolled, highest degree received.
- ☐ No schooling completed
- ☐ Nursery school to 8th grade
- ☐ Some high school, no diploma
- ☐ High school graduate, diploma or the equivalent
- ☐ Some college credit, no degree
- ☐ Trade/technical/vocational training
- ☐ Bachelor's degree
- ☐ Master's degree
- ☐ Professional degree
- ☐ Doctorate degree
- ☐ Prefer not to say

Q5: What is your personal status?

- ☐ Single, never married
- ☐ Married or in a domestic partnership
- ☐ Widowed
- ☐ Separated
- ☐ Divorced
- ☐ Prefer not to say
- ☐ Other (please specify)

Q6: What is your employment status?

- ☐ Employed for wages
- ☐ Self-employed
- ☐ Out of work and looking for work
- ☐ Out of work but not currently looking for work
- ☐ A homemaker
- ☐ A student
- ☐ Military
- ☐ Retired
- ☐ Unable to work
- ☐ Prefer not to say

As a Worker in Amazon Mechanical Turk or Appen platforms please answer the following questions

Q7: How long have you been working on crowdsourcing tasks?

- ☐ Less than one year
- ☐ 1-2 year
- ☐ 3-4 years
- ☐ Over 4 years
- ☐ Prefer not to say

Q8: What level of expertise do you have in crowdsourcing tasks?

- ☐ Beginner
- ☐ Intermediate
- ☐ Advanced

Q9: How many different types of these tasks have you participated in on the platform

- ☐ This is my first task
- ☐ 2-5 tasks
- ☐ 6-10 tasks
- ☐ 11-20 tasks
- ☐ More than 20

Q10: How often do you use the following types of devices for performing crowdsourcing tasks?

	Always	Frequently	Occasionally	Rarely	Very Rarely	Never
Desktop (laptop or personal computer)	☐	☐	☐	☐	☐	☐
Mobile (tablet or smartphone)	☐	☐	☐	☐	☐	☐
Other (please specify)	☐	☐	☐	☐	☐	☐

Q11: What best describes your current location

- ☐ Office
- ☐ Home
- ☐ Mobile
- ☐ Other, (please specify)

Q12: How many times have you visited the Mrs Planet platform?

- ☐ 0
- ☐ 5
- ☐ 10
- ☐ 15
- ☐ 20

Q13: Is crowdsourcing the main source of your income?

- ☐ No
- ☐ Yes
- ☐ Prefer not to say

Q14: Rank in order of preference the following criteria that matter to you when choosing to perform this task? (required) Use Drag and Drop (1 being your most favourite criteria and 9 your least favourite)

- _____ Reward
- _____ Time required for completion
- _____ The most interesting
- _____ Include Creative work
- _____ The platform chooses the task for me
- _____ Similar to task that you have done before
- _____ Levels of difficulty
- _____ Includes professional development, education, career path advancement
- _____ Other, please specify

Q15: The following statements ask you to reflect on your experience of engaging with this task. For each statement, please use the following scale to indicate what is most true for you.

	Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
I lost myself in this experience	☐	☐	☐	☐	☐
The time I spent doing the task just slipped away.	☐	☐	☐	☐	☐
I was absorbed in this experience.	☐	☐	☐	☐	☐
I felt frustrated while doing this task.	☐	☐	☐	☐	☐
I found this task confusing to do.	☐	☐	☐	☐	☐
This task was taxing	☐	☐	☐	☐	☐
This task was attractive.	☐	☐	☐	☐	☐
This task was aesthetically appealing.	☐	☐	☐	☐	☐
This task appealed to my senses	☐	☐	☐	☐	☐
Doing this task was worthwhile.	☐	☐	☐	☐	☐
My experience was rewarding.	☐	☐	☐	☐	☐
I felt interested in this experience.	☐	☐	☐	☐	☐

Q16: How well do the following statement describe your personality: I see myself as someone who

	Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
.. is reserved	☐	☐	☐	☐	☐
.. is generally trusting	☐	☐	☐	☐	☐
..tends to be lazy	☐	☐	☐	☐	☐
..is relaxed, handles stress well	☐	☐	☐	☐	☐
..has few artistic interests	☐	☐	☐	☐	☐
..is outgoing, sociable	☐	☐	☐	☐	☐
..tends to find fault with others	☐	☐	☐	☐	☐
..does a thorough job	☐	☐	☐	☐	☐
..gets nervous easily	☐	☐	☐	☐	☐
..has an active imagination	☐	☐	☐	☐	☐

Q17: This scale questions are regarding how do you found the task where 1 is (very not very) and 7 is (very).

	1	2	3	4	5	6	7
How mentally demanding was the task?	☐	☐	☐	☐	☐	☐	☐
How physically demanding was the task?	☐	☐	☐	☐	☐	☐	☐
How hurried or rushed was the pace of the task?	☐	☐	☐	☐	☐	☐	☐
How successful were you in accomplishing what you were asked to do?	☐	☐	☐	☐	☐	☐	☐
How hard did you have to work to accomplish your level of performance?	☐	☐	☐	☐	☐	☐	☐
How insecure, discouraged, stressed and annoyed were you?	☐	☐	☐	☐	☐	☐	☐

Q18: Is there any feedback, opinion, or question about this survey you want to share with us?

Q19: Important: Please write your Amazon worker' ID here to get your bonus

Here is your ID [\\${c://Field/Random%20ID}](#)

Please copy and paste this into MTurk

When you have copied this ID, please click the 'next' button to submit your survey.

A.4 Demographic Details of Participants in Study 1

This appendix presents the extended demographic distributions of Amazon Mechanical Turk workers who participated in the three task types (text, image, and survey). These charts complement the summary figures presented in Section 4.8.1.

In summary, the extended demographic results confirm the representativeness of the participant pool, showing a predominance of workers aged 25–34, with a bachelor’s degree, working primarily from home on a desktop device. Approximately half identified as intermediate-level Amazon Mechanical Turk workers with 3–4 years of experience as shown in Figures A.1–A.5.

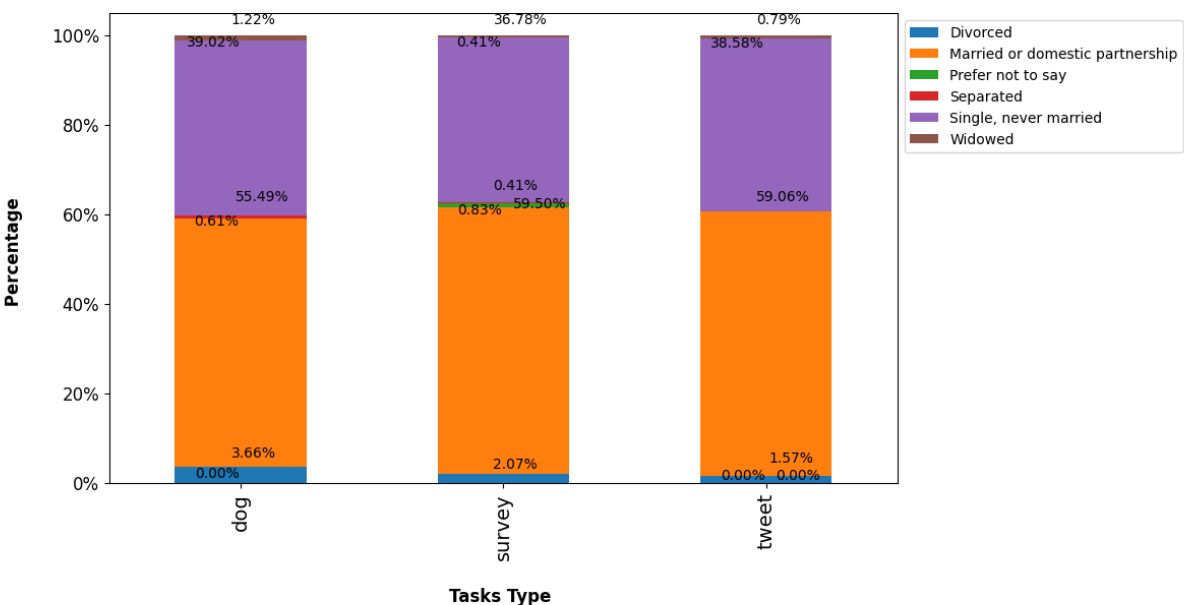


Figure A.1: Marital distribution across the tasks

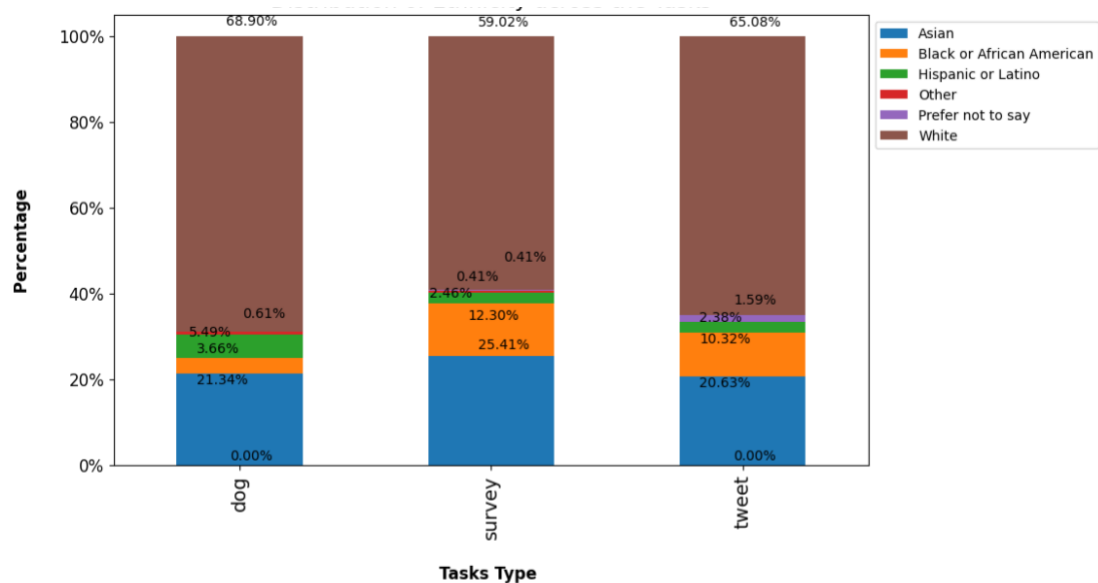


Figure A.2: Ethnicity distribution across the tasks

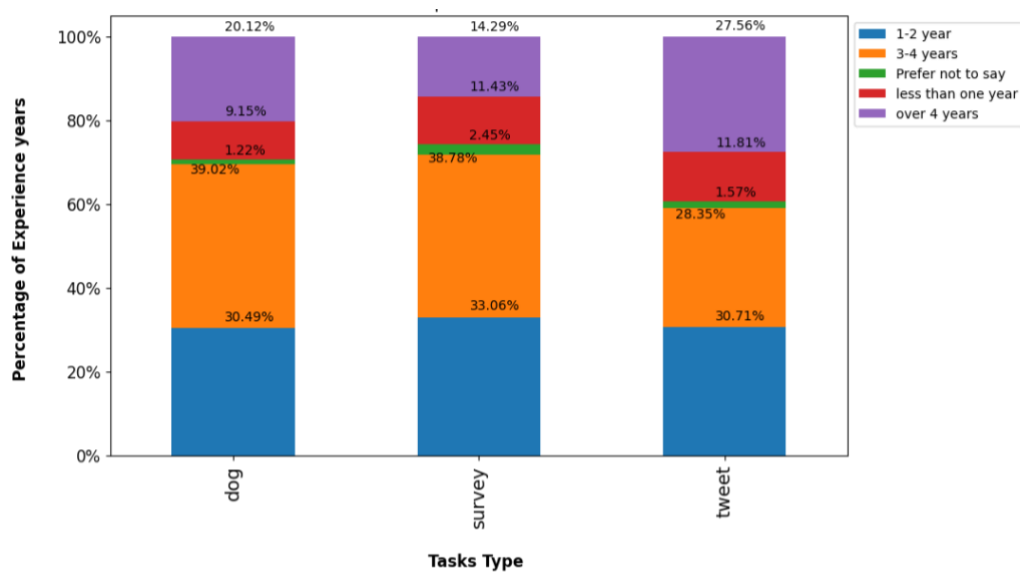


Figure A.3: Years of experience across the tasks

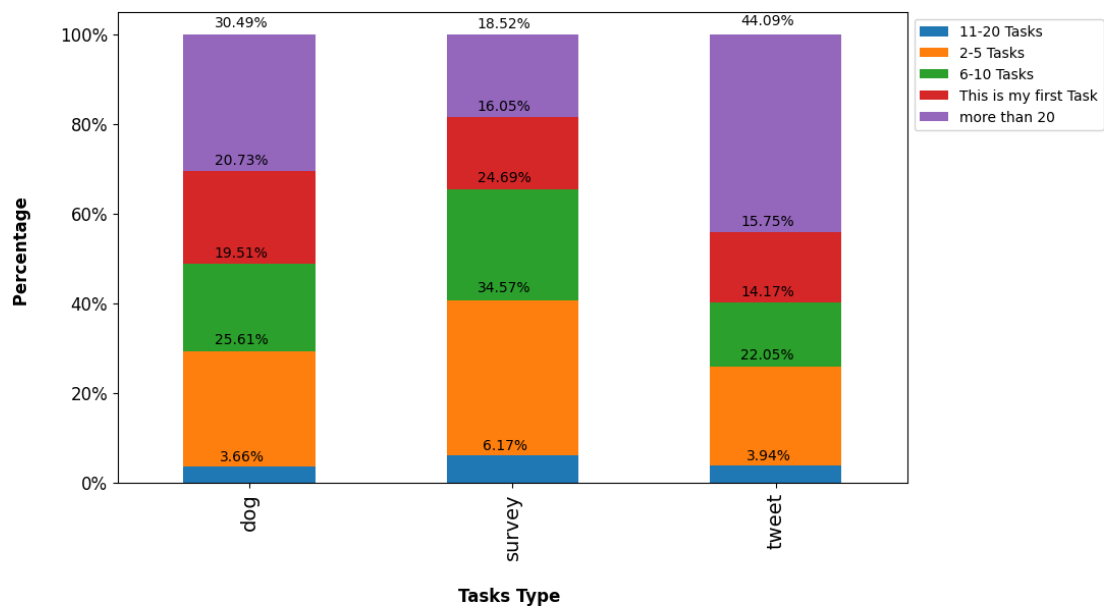


Figure A.4: Worker experience by number of tasks

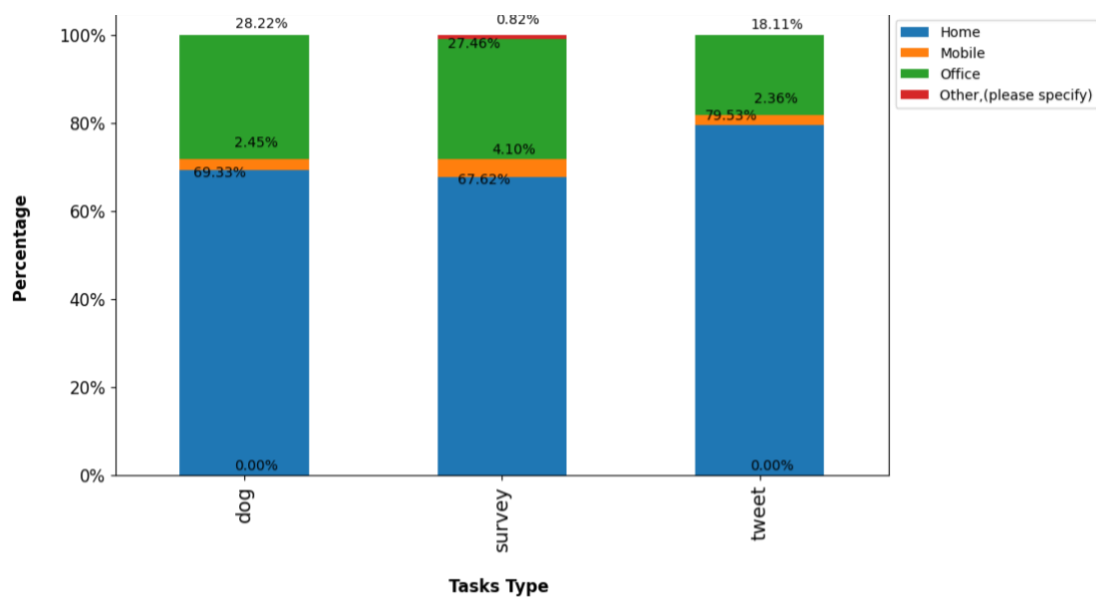


Figure A.5: Worker location during crowdsourcing tasks

Appendix B

B.1 Ethics Approval for Study 3

1/5/26, 12:41 PM

CIS Ethics Approval System – Computer and Information Sciences – local.cis



Computer and Information Sciences - local.cis

departmental information for staff and students

[Home](#) | [Ethics](#) | [Events](#) | [Safety](#) | [IT Support](#) | [Teaching](#) | [Utilities](#)

Browse: [Home](#) / [Utilities](#) / CIS Ethics Approval System

CIS Ethics Approval System

You are Huda Banuqitah (Research Student - 202084148)

[Return to Main](#)

Application ID: 2728

Parent Application ID: [2548](#)

Title of research:

Motivating Engagement and Performance in Crowdsourced Image Labelling Through Gamification

Summary of research (short overview of the background and aims of this study):

As part of my PhD, I am investigating the effect of gamification on engagement and performance in the Image classification. The Participants or workers from Amazon mechanical Turk and Appen crowdsourcing platform will be asked to do:

- ☐ Perform a gamified image classification task classifying galaxy images according to their morphology.
- ☐ Complete a follow-up Survey with multiple choice questions about demographic information such as gender, age, field of study and employment, workload, engagement gamification elements evaluation questions and personality traits questions.

The gamification task webpage will be loaded and rendered to the MTurk workers dashboards and Appen workers on an iframe display.

Our previous study in crowdsourcing platforms with Ethical approval ID (1399) was to investigate the engagement and performance of crowdsourcing workers in different task types. This study will help us to determine the effect of gamification framework model on improving workers' engagement and performance of contribution in labelling.

How will participants be recruited?

Around 250 participants will be invited to conduct the task in this study. For ecological validity, the participant will be recruited via a crowdsourcing in the normal way. The survey will be posted on the crowdsourcing website stating the details of the task. When participants who accept the task complete it, the researchers will be notified from the platform via email.

On Mturk and Appen the participants are paid based on the tasks done, we have estimated that the tasks will take about 20 minutes to complete mandatory task with the post survey per participant. After discussion with colleagues, we have agreed that for the US minimum wage of \$8.50/hr, a fair rate of payment that is within the average payment by university team on MTurk. I will pay the workers 2.85\$/task on MTurk (as we calculated by minimum wage per hour). The participation information has been included in the consent information form.

It is unclear from the above whether you are paying per hour or per task. Please clarify.

Please also define "task". Is it e.g. completing one label or doing all 20 labels+survey?

It should also be made clearer in the Appendices how points relate to tasks.

Will paying the Appen participants a different rate according to their age (as it looks here) bias the recruitment? Consider paying all participants the same, or clarify that this is already the case.

The mandatory task appears to be 20 labels plus completing the survey. Clarify whether the 20 minutes mentioned here applies to this. Or is it 20 minutes on the labelling then 10 on the survey? 20 minutes is not mentioned anywhere in the Appendices. The approximate time that it will take to perform the mandatory task plus the survey should be stated on the Appendices including the Consent form.

What will the participants be told about the proposed research study? Either upload or include a copy of the briefing notes issued to participants. In particular this should include details of yourself, the context of the study and an overview of the data that you plan to collect, your supervisor, and contact details for the Departmental Ethics Committee.

PDF File: [View document](#)

I have amendment a pdf based on the feedback as follow:

At the consent form we clarify more for participant that we will pay them \$ 2.5 for submitting the minimum of 20 labels and also we clarify at the post survey in a few consent sentences that we will pay them an extra for completing the post survey and we encourage them to provide as much feedback as possible with mention that answering every question is not mandatory."

<https://local.cis.strath.ac.uk/wp/extras/ethics/?view=2728>

1/3

On Amazon Mechanical Turk and Appen Platforms, the task will appear on the dashboards, as task title with estimated payment and duration time, the workers could click on the preview button to read the task instruction and description and decide whether to participate or not. The consent statement with the game instruction of participating will be display to the workers at the hit preview (see Attached Document)

Image Classification Task: As part of our **research** into behaviour of crowd-workers, the goal of this task is to view different galaxies images and categorize them based on the numbers of arms the galaxy presents in the image.

The collected data may be used in academic publications, but your identity will be protected. No sensitive personal, or identifiable information will be collected except for your worker's platform ID. An anonymous data set will be created once the study is complete by replacing worker IDs with randomly generated ones.

The participant will be told that:

You can request the deletion of your data within two months by post or email to me. After this your data will have been anonymized, so that it will not be possible to identify your data in order to delete it.

In each task we will do the following:

- In the first task, the participants will be told that by completing and submitting the tasks they indicate consent to participate in the study.
- Participants will be told that their participation is voluntary and there is no harm or risk associated with collaboration to this study.
- Responses will not be shared with anyone besides the **researchers**, but their analysis may be published in a scientific venue.
- All answers will be confidential and anonymous.
- Participants have the right to withdraw at any time before submitting the survey.
- Participants have the right to not answer any question they feel uncomfortable with.
- If participants have any concern regarding their participation in this study, they can contact the **researcher** or the Department Ethics Committee via email: ethics@cis.strath.ac.uk.
- The Qualtrics survey link: https://strathsci.qualtrics.com/jfe/form/SV_6zKtHAN9wVuZ3IW

It should be made clearer in Appendix 1 Consent form whether the collaboration with other workers on the task is live/synchronous and that communication with other workers via a Chat function may be involved. It should be made clear that this is anonymous.

In the survey, when you ask for participants' ID, make it clear that they should not enter any other identifying information.

(note a typo on Page 2: "Monterey" should be monetary)

How will consent be demonstrated? Either upload or include here a copy of the consent form/instructions issued to participants. It is particularly important that you make the rights of the participants to freely withdraw from the study at any point (if they begin to feel stressed for example), nor feel under any pressure or obligation to complete the study, answer any particular question, or undertake any particular task. Their rights regarding associated data collected should also be made explicit.

PDF File: [View document](#)

I have amendment a pdf based on the feedback as follow:

At the instruction set of the task, we clarify for workers the minimum set of tasks required for the payment. Moreover we encourage them we mention out that the tasks will payment will be based on the points gained during the tasks. Then we mention the conversion of accumulated Points to monetary Rewards will be by end of the task. " Your total points will determine your final payment, ensuring fair compensation for your efforts."

More over we introduced 2 more bounses for participants in addition to the honeypot quetions, this based on their collaboration oand agreement during the labeling tasks, we clarify this by encourage the participant at the instruction and consent from, the correct answers to the honeypot questions will boost thier scores based on the number of teams that agree on the correct answer. " teamwork is crucial, and your team's performance affects your individual and collective scores. Pay attention to the accuracy as too many incorrect labels may impact your bonus score." For majority agreement we increased the workers hwo agree with other by 6 cent that's appear as 6 points in their score, so If the majority of participant (at least 3 out of 5) agree on the same label, each of them will earn an additional 6 points per image. And for full agreement, If all the workers agree on the same label, all team members will eam an additional 7 points per image which equal to 7 cent.

The first page of the Tasks will include the consent details as shown (See Attached Document), which will make it clear for participate that:

- Their participation will be anonymised and kept confidential.
- Their participation is voluntary.
- Participants have the right to withdraw at any time before submitting their responses.
- The survey questions should not cause any discomfort. Yet, participant have the right not to answer any question that they feel uncomfortable with.
- The electronic data of the tasks will be stored on crowdsourcing platforms servers. The survey data will be stored on Qualtrics server.
- If participants have any concern regarding their participation in this study, they can contact the **researcher** or the Department Ethics Committee via email.

Again, "Monterey" should be monetary

What will participants be expected to do? Either upload or include a copy of the instructions issued to participants along with a copy of or link to the survey, interview script or task description you intend to carry out.

PDF File: [View document](#)

PDF File: None.

I have amendment a pdf of the instructions and consent based on what we discuss in previous 2 sections.

The participant from crowdsourcing platforms could submitting minimum of 20 from 50 labels task with post survey, or he/she could choose to continue and get more reward or stop by doing only mandatory task. Then the workers' answers will back to the **researchers**. See below section for more details

What data will be collected and how will it be captured and stored? In particular indicate how adherence to the Data Protection Act and the General Data Protection Regulation (GDPR) will be guaranteed and how participant confidentiality will be handled.

- No sensitive personal or identifiable information will be collected.
- The **researcher** will abide by the provision of the General Data Protection Regulation (GDPR).
- The data will be processed and analysed fairly, with limited purposes, and not kept longer than necessary.
- The data that will be gathered from the task is workers nickname and mood image character at the login page (Figure1 in the attached file). The starting and end

time of each game level, the answer selection of each labelling task and the text messages between the workers. This data will be stored at Crowd platform server.

e. The survey data will be divided into parts: first section is about demographic features such as gender, age, education, devices type used in the task, income level from the crowd platforms. The second parts Likert scale question to measure engagement. The third section about the NASA task load followed by the fourth part which is the Likert scale questions about the workers' measure for each mechanism used in the game. Finally, the Big Five personality traits questions. This will be stored on Qualtrics server.

g. The analysis of the data will be on a **researcher** OneDrive.

h. The data will be secure and restricted to the **researcher** and the **researcher's** supervisors.

What is "mood" in part d? Where is this collected?

How will the data be processed? (e.g. analysed, reported, visualised, integrated with other data, etc.) Please pay particular attention to describing how personal or sensitive data will be handled and how GDPR regulations will be met.

a. The collected data will be used only for this investigation.

b. The data will be analysed based on the **researcher** goals, using statistical tools as python and SPSS to measure the workers engagement, workload, personality traits and analyse the relationship between this to their labelling accuracy and different demographic features, there will be no attempt to de-anonymize the data.

c. The electronic data of the tasks will be stored crowdsourcing platforms servers. The anonymous survey data will be stored on Qualtrics server (by enabling anonymous settings in Qualtrics), and a copy will be stored on the **researcher's** One Drive for analysis purposes.

d. The data will be compared with other sources such as related results from related **research** papers.

e. A report of the findings of this study will be published as a journal article or conference paper. The collected data will not include any information that could identify individual participants.

You should state that Qualtrics will be used in anonymous mode.

How and when will data be disposed of? Either upload a copy of your data management plan or describe how data will be disposed.

PDF File: None.

No personal data is recorded except the workers' IDs that will be stored on Amazon Mechanical Turk and Qualtrics accounts. Within Fourteen days from the end of the study, all the data will be downloaded into OneDrive and deleted from the AMT server and Qualtrics server. Once it downloaded the data will be anonymised to be used in further analysis and optional publication in Pure system. Any data that is not published through Pure will be securely destroyed upon the completion of my PhD degree.

"Any data that will not be published through Pure will be removed when I complete my PhD degree".

Make clear that this data is being destroyed.

Supervisor's name (students need to complete/staff should probably put N/A).

☒ I confirm that my supervisor has seen and approved both my planned study and this associated ethics application.

B.2 Consent Form for Study 3

This research is being conducted by myself (Huda Banuqitah (huda.banuqitah@strath.ac.uk)) under the supervision of Dr Mark Dunlop (mark.dunlop@strath.ac.uk) and Dr Sotirios Terzis (sotirios.terzis@strath.ac.uk) – all members of the Computer and Information Sciences Department, University of Strathclyde. I am currently a Ph.D. student at the University. My research aims to improve the overall quality of crowdsourcing platforms by developing techniques that help to retain and improve worker engagement and performance in image-labelling crowdsourcing. Your collaboration is highly appreciated, and your input is very important to us. Thank you for considering participating in our study. We have designed a collaborative game that aims to encourage worker engagement and performance, where you will work collaboratively and anonymously (by Nickname) with a team of workers to label images of galaxies based on the number of arms in each image. Your participation will contribute significantly to our research efforts in improving task design in crowdsourcing platforms. The task will involve labelling up to a total of 50 images. For your participation in this study, you will be compensated in accordance with common US minimum wage levels \$8.50/h. To be eligible for payment, you are required to complete a minimum of 20 labels and fill in a follow-up survey which will take around 20 minutes. Your payment of the minimum tasks will be based on the minimum wage levels \$8.50/h for both the labelling and post survey tasks. However, we encourage you to label more images, as increasing the number of labels completed will result in higher compensation, especially if you collaborate live with your team to improve the accuracy of the labels. The more correct image labels, the higher your reward will be. The research was granted ethical approval by the University of Strathclyde Computer and Information Sciences Ethics Committee. The collected data may be used in academic publications, but your identity will be protected. No personal, or identifiable information will be collected except for your worker's platform ID. An anonymous data set of the tasks and a post survey data set will be created once the study is complete by replacing worker IDs with randomly generated ones. Your contributions are invaluable to the success of our study, and we appreciate your time and effort in advance. If you have any questions or concerns, please feel free to contact us. If you have any questions/concerns during or after the research, or wish to contact an independent person to whom any questions may be directed or further information sought, please contact the Ethics Committee of the Department of Computer and Information Sciences, Livingstone Tower Richmond Street Glasgow G1 1XH email: ethics@cis.strath.ac.uk

B.3 Scripts of Publishing the Task on MTurk

```

4
5 client = boto3.client('mturk', endpoint_url = endpoint_url, region_name='us-east-1', aws_access_key_id="",
6     aws_secret_access_key="")
7
8 from boto.mturk.question import ExternalQuestion
9
10 new_hit = client.create_hit(
11     Title='Collective Image Labelling Task',
12     Description="""Join a collaborative research task where you'll work with a team to
13     label galaxy images. Completing the main labeling task will earn you a base reward of $1,
14     and you'll have the opportunity to earn additional money
15     (up to more than $4.5 total) by completing a short follow-up survey
16     and contributing accurate labels . The more accurate your team's labels,
17     the higher the reward potential!
18     Create Jira Issue
19     Desktop/laptop recommended; Refresh browser if you face any issue""",
20     Reward='1.0',
21     MaxAssignments=180,
22     LifetimeInSeconds=14400,
23     AssignmentDurationInSeconds=60*40,
24     AutoApprovalDelayInSeconds=259200,
25     Question=ExternalQuestion("https://www.collaborativegames.net/game/", frame_height=1000).get_as_xml())
26
27 print("A new HIT has been created. You can preview it here:")
28 print("https://workersandbox.mturk.com/mturk/preview?groupId=" + new_hit['HIT']['HITGroupId'])
29 print("HITID = " + new_hit['HIT']['HITId'] + " (Use to Get Results)")

```

Figure B.1: Script of team-based condition hit publishing on Amazon Mechanical

```

1  import boto3
2
3  endpoint_url = 'https://mturk-requester.us-east-1.amazonaws.com'
4
5  client = boto3.client('mturk', endpoint_url = endpoint_url, region_name='us-east-1', aws_access_key_id="",
6      aws_secret_access_key="")
7
8  from boto.mturk.question import ExternalQuestion
9
10
11  new_hit = client.create_hit(
12      Title=' Galaxy Image Labeling with Performance Rewards!',
13      Description="""Join a research task where you'll work to label galaxy images.
14      Completing the main labeling task (Minimum of 20 images) will earn you a base reward of $ 1,
15      and you'll have the opportunity to earn additional money by contributing accurate labels
16      with survey up to $ 4. The more accurate your labels, the higher the reward potential!""",
17      Reward='1.0',
18      MaxAssignments=90,
19      LifetimeInSeconds=14400,
20      AssignmentDurationInSeconds=1800,
21      AutoApprovalDelayInSeconds=259200,
22      Question=ExternalQuestion("https://www.collaborativegames.net/game/", frame_height=1000).get_as_xml())
23
24  print("A new HIT has been created. You can preview it here:")
25  print("https://workersandbox.mturk.com/mturk/preview?groupId=" + new_hit['HIT']['HITGroupId'])
26  print("HITID = " + new_hit['HIT']['HITID'] + " (Use to Get Results)")
27

```

Figure B.2: Script of individual condition hit publishing on Amazon Mechanical

B.4 Task Instruction of Full-Feature Condition in Study 3

Task Instructions and Rules

Overview

Levels: 5 levels, each with 10 images.

Objective: Earn points for accurate labels to maximize your reward.

Requirements

Complete at least 2 levels + survey for a reward.

Additional reward for completing the survey at the end.

How to label

- **Vote:** Select your answer using the radio buttons beside each image. Teammates' votes will display as avatars.
- **Confirm:** Press "I'm done" when ready. Once all teammates confirm, you'll move to the next image.
- **Timer:** If half the team confirms, a 1-minute timer starts, and answers will auto-submit when time's up.
- **Chat:** Use the chat to discuss and coordinate with your team. Accurate collaboration improves your score and potential rewards!

Scoring

- **Progress:** Each labeled image earns **5 points**.
 - Points convert to rewards at the end of each task.
- **Correct labels:** Each 10 has one image with a known answer (honeypot).
 - The entire team earns bonus points for each correct honeypot label identified by any player. Score scales with level. **First level honeypot image:** 5 points, **second level honeypot image:** 6 points and so on.
- **Agreement**
 - +1 bonus point if the majority of the team selects the accurate label.
 - +2 bonus points if all team members select the accurate label.

Leaderboard

Teams are ranked by total points, so accuracy and teamwork pay off!

Classification example

Your task is to classify galaxy image by count of arms

- | | |
|--|--|
| <ul style="list-style-type: none">•  If galaxy looks like this - it has one arm | <ul style="list-style-type: none">•  This is four-arm galaxy |
| <ul style="list-style-type: none">•  This is two-arm galaxy | <ul style="list-style-type: none">•  If galaxy has more than four arms - select "more than 4" label |
| <ul style="list-style-type: none">•  This is three-arm galaxy | <ul style="list-style-type: none">•  If you're unsure that galaxy has arms - check "can't tell label" |

Ending the task

The task ends upon completing all levels or by pressing "quit". You'll then complete a brief survey to finalize your reward.

Do not close task window until you enter survey code!

Thank you for participating in the Galaxy Classification Task—your input helps improve crowdsourcing!

Good luck and enjoy the experience.

Figure B.3: Full-Feature Task Instruction

B.5 Task Instruction of Chat-Only Condition in Study 3

Task Instructions and Rules

Overview

The task has 50 images

Objective: Make accurate labels to maximize your reward.

Requirements

Complete at least 20 + survey for a reward.

Additional reward for completing the survey at the end.

How to label

- **Vote:** Select your answer using the radio buttons beside each image. Teammates' votes will display as avatars.
- **Confirm:** Press "I'm done" when ready. Once all teammates confirm, you'll move to the next image.
- **Timer:** If half the team confirms, a 1-minute timer starts, and answers will auto-submit when time's up.
- **Chat:** Use the chat to discuss and coordinate with your team. Accurate collaboration improves your score and potential rewards!

Reward

1. **Base:** You'll receive a base payment of \$1 if requirements are met.
2. **Quality bonus:** You'll receive bonus reward based on correct labels for honeypot images.
3. **Agreement bonus:**
 - Bonus for accurate majority agreement among teammates
 - More bonus for accurate agreement among all teammates

Classification example

Your task is to classify galaxy image by count of arms

- | | | | |
|---|--|---|--|
|  | If galaxy looks like this - it has one arm |  | This is four-arm galaxy |
|  | This is two-arm galaxy |  | If galaxy has mor than four arms - select "more than 4" label |
|  | This is three-arm galaxy |  | If your unsrue that galaxy has arms - check "can't tell label" |

Ending the task

The task ends upon completing all levels or by pressing "quit". You'll then complete a brief survey to finalize your reward.

Do not close task window until you enter survey code!

Thank you for participating in the Galaxy Classification Task—your input helps improve crowdsourcing!

Good luck and enjoy the experience.

Figure B.4: Chat-Only Task Instruction

B.6 Task Instruction of No-Element Condition in Study 3

Task Instructions and Rules

Overview

The task has 50 images

Objective: Make accurate labels to maximize your reward.

Requirements

Complete at least 20 + survey for a reward.

Additional reward for completing the survey at the end.

How to label

- **Vote:** Select your answer using the radio buttons beside each image. Teammates' votes will display as avatars.
- **Confirm:** Press "I'm done" when ready. Once all teammates confirm, you'll move to the next image.
- **Timer:** If half the team confirms, a 1-minute timer starts, and answers will auto-submit when time's up.

Reward

1. **Base:** You'll receive a base payment of \$1 if requirements are met.
2. **Quality bonus:** You'll receive bonus reward based on correct labels for honeypot images.
3. **Agreement bonus:**
 - Bonus for accurate majority agreement among teammates
 - More bonus for accurate agreement among all teammates

Classification example

Your task is to classify galaxy image by count of arms

- | | | | |
|---|--|---|--|
| •  | If galaxy looks like this - it has one arm | •  | This is four-arm galaxy |
| •  | This is two-arm galaxy | •  | If galaxy has mor than four arms - select "more than 4" label |
| •  | This is three-arm galaxy | •  | If your unsrue that galaxy has arms - check "can't tell label" |

Ending the task

The task ends upon completing all levels or by pressing "quit". You'll then complete a brief survey to finalize your reward.

Do not close task window until you enter survey code!

Thank you for participating in the Galaxy Classification Task—your input helps improve crowdsourcing!

Good luck and enjoy the experience.

Figure B.5: No-Element Task Instruction

B.7 Task Instruction of Individual Condition in Study 3

Task Instructions and Rules

1. **Classify** images of galaxies based on the number of arms they have.
2. **The task** has 50 images
3. **Complete at least 20 + survey** for reward
4. **Your reward** is based on the minimum wage per hour
5. **Additional reward** for accurate classifications of 'honeypot' images and for completing the post-task survey.
6. **Select** your answer using radio buttons beside each image
7. **Press "Next"** to proceed to the next image

Classification example

Your task is to classify galaxy image by count of arms

-  If galaxy looks like this - it has one arm
-  This is two-arm galaxy
-  This is three-arm galaxy
-  This is four-arm galaxy
-  If galaxy has mor than four arms - select "more than 4" label
-  If your unsrue that galaxy has arms - check "can't tell label"

Ending the task

The task ends once you complete all levels or by pressing "Quit". You'll then complete a brief survey to finalize your reward.

Do not close task window until you enter survey code!

Thank you for participating in the Galaxy Classification Task—your input helps improve crowdsourcing!

Good luck and enjoy the experience.

Figure B.6: Individual Task Instruction

B.8 Demographic Details of Participants in Study 3

Gender analysis showed different distributions across the experimental conditions. The Chat-Only (CO) condition has a relatively balanced gender ratio (Male: 50.79%, Female: 48.41%), with a minor percentage preferring not to reveal their gender (0.79%). The Full-Features (FF) condition also demonstrated a higher proportion of male participants (65.42%), and lower female participation (34.58%). The highest skew in male participation can be seen in No-Elements (NE) and Individual conditions, at (77.78%) and (78.69 %), respectively. Figure B.7 summarises gender distribution across the experimental conditions. Participants' age distribution was concentrated primarily within the age brackets of 25–34 and 35–44 years, which represented the dominant categories across all conditions. Specifically, the higher proportion of participants aged 25–34 years (55.56%) have been shown in the Chat-Only condition, followed by 35–44 years (33.33%). Similar age distributions were observed in the Full-Features (25–34 years: 54.21%, 35–44 years: 27.10%) and No-Elements conditions (25–34 years: 86.11%, 35–44 years: 9.72%). Notably, the individual condition also highlighted a substantial proportion of participants aged 25–34 years (73.77%). Few participants fell within the 18–24 or ≥ 45 -year categories, highlighting that the sample comprised predominantly young to mid-adulthood participants. Figure B.8 visually summarises the age distribution across conditions.

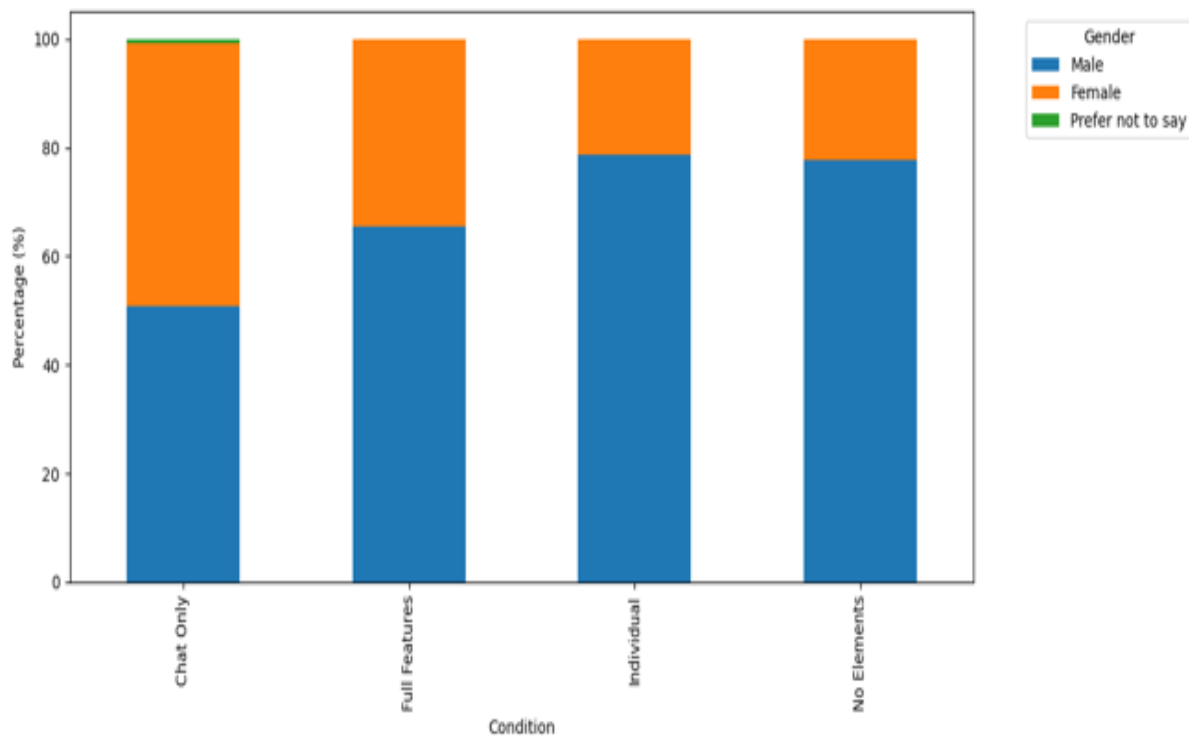


Figure B.7: Gender Distributions Across Conditions

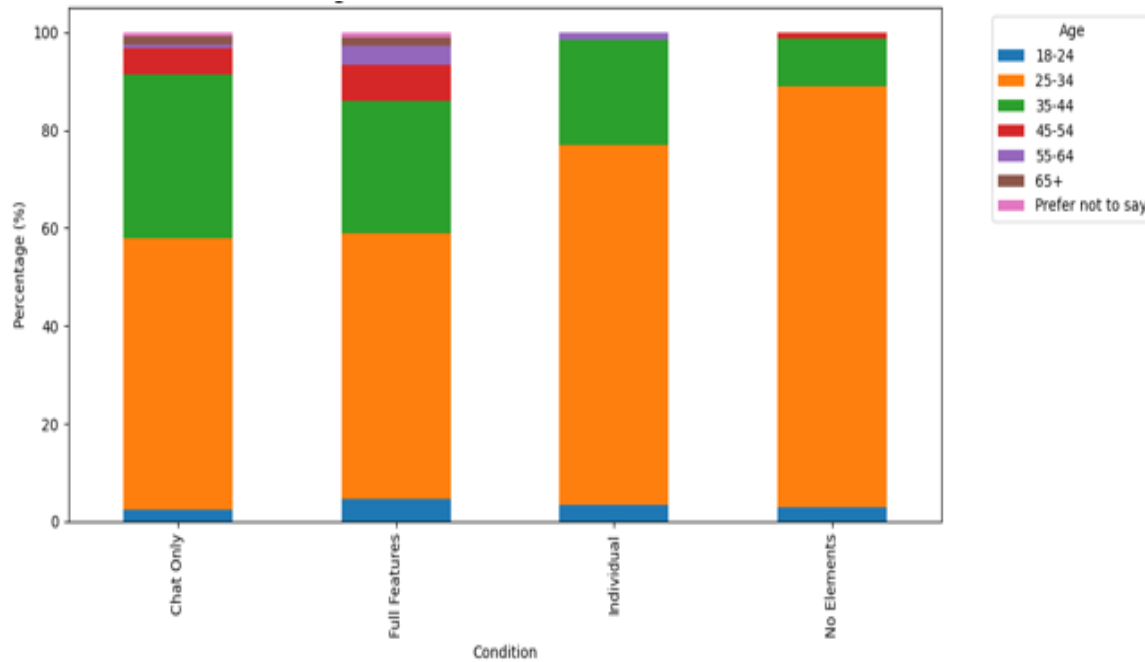


Figure B.8: Age Distributions Across Conditions

Analysis of educational levels has shown that a bachelor's degree was noted as the highest majority educational degree among all the participants in different experimental conditions. The No-Elements condition (88.89%) had the greatest percentage; closely followed were Chat-Only (81.75%), Full-Features (74.77%), and the Individual condition (63.93%). A smaller proportion of participants in the individual condition (26.23%) had a master's degrees. Doctoral-level education was rare across all conditions, consistent with a large sample including graduates with bachelor's or master's degree levels. Figure B.9 shows the exact educational distribution. Analysis of device usage patterns demonstrated significant consistency in desktop usage, with participants frequently reporting they "Always" used desktops, in Chat-Only (58.73%) and Full-Features (55.66%) and around (29%) in other conditions as shown in Figure B.10. In contrast, as shown in Figure B.11 the usage of mobile devices has been more dispersed, with a considerable proportion among all conditions using mobiles "Rarely" or "Very Rarely". Frequent mobile usage was limited across all conditions, indicating a practical preference for desktop interaction, most likely owing to the visual complexity and sustained attention demanded by the image-labelling task. Figure B.12 summarise the devise usage across the experimental conditions.

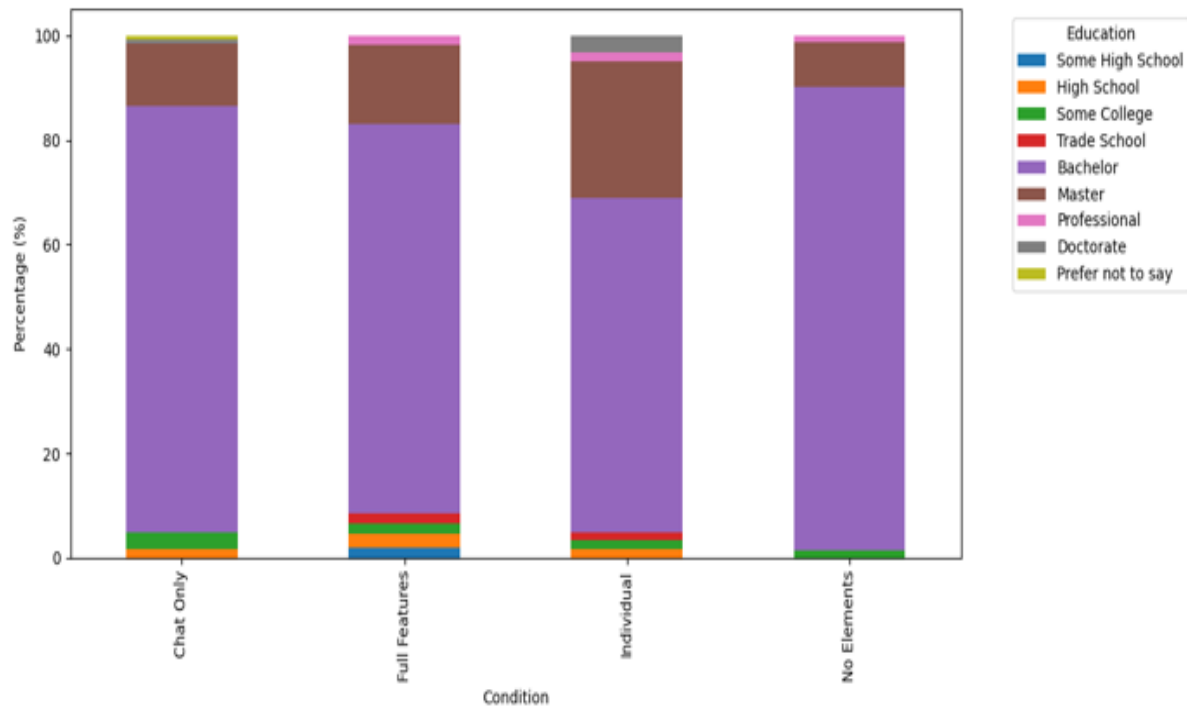


Figure B.9: Educational Levels Distribution Across the Conditions

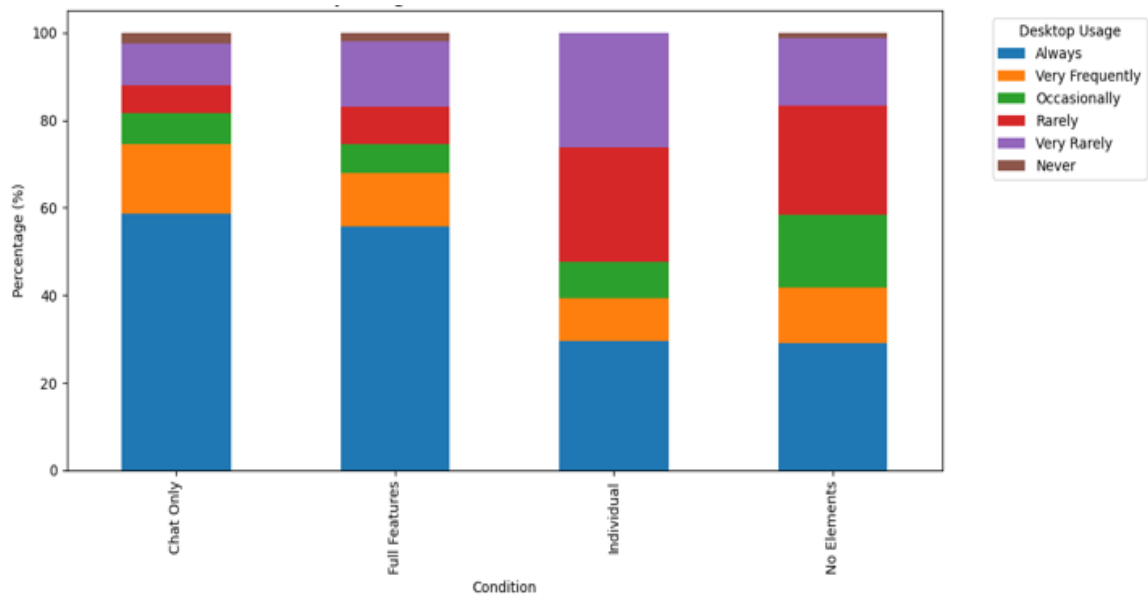


Figure B.10: Desktop Device Usage Across the Conditions

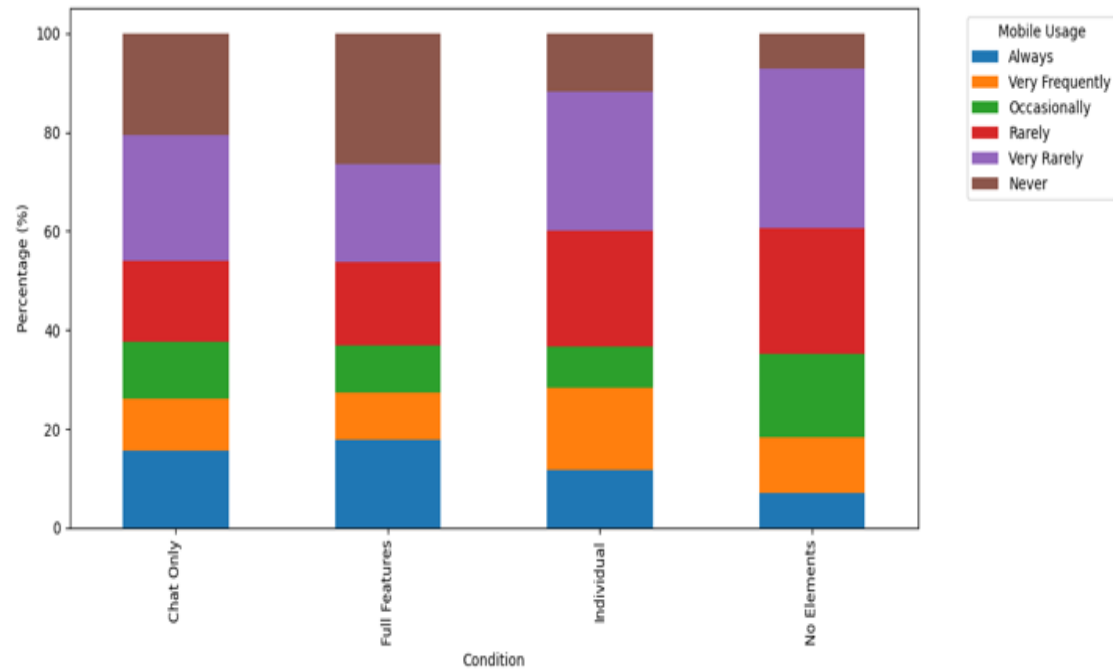


Figure B.11: Mobile Device Usage Across the Conditions

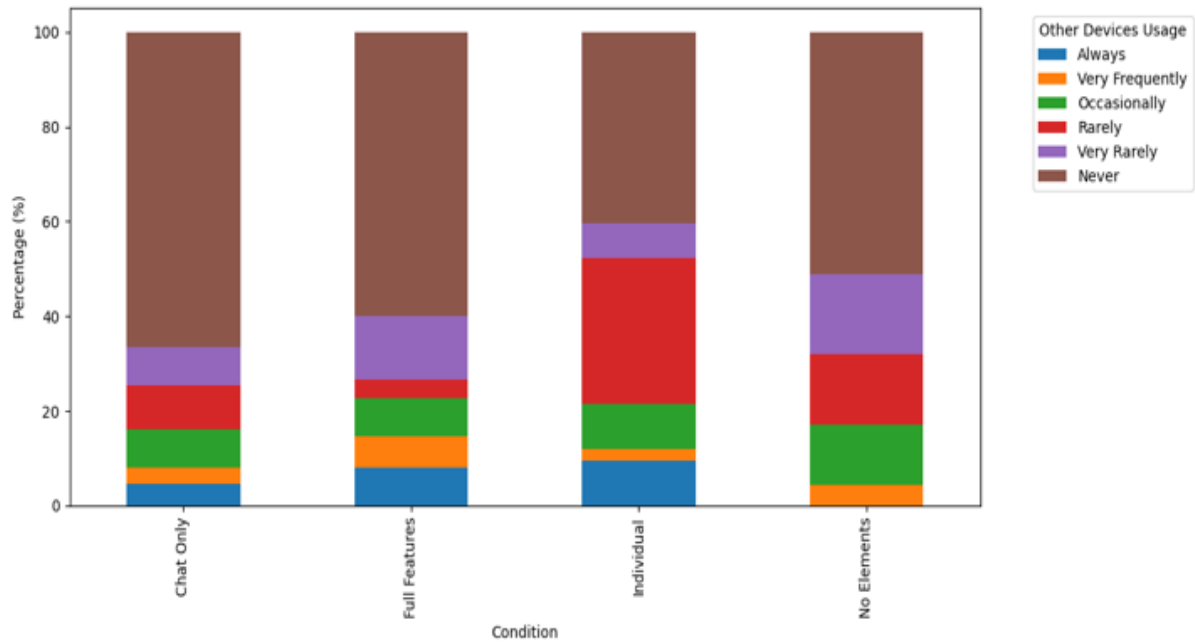


Figure B.12: Other Devices Usage Across the Conditions

The relationship between gender distribution and average accuracy across the four experimental conditions—full features, Chat-Only, Individual and No-Elements—has been evaluated using a set of statistical tests as follows: First, a one-way ANOVA has been done for each condition to observe whether average accuracy varied significantly between various gender types. The results showed no significant gender effect in any condition (Chat-Only: $F = 1.29, p = .270$; Individual: $F = 0.56, p = .500$; No Element: $F = 1.47, p = .230$, and Full-Features: $F = 0.48, p = .500$). Therefore, these results indicated that male and female participants exhibited comparable performance levels within each experimental condition. We then investigated if gender distributions were similarly distributed among the experimental circumstances by means of a chi-square test of independence. The study found a statistically significant difference in the number of males and females ($\chi^2(3) = 20.25, p < .001$), which means that there were more of one gender than the other. This imbalance arises from natural variation arising from random assignment rather than systematic experimental control, as gender was not employed as a stratification factor. The ANOVA results for accuracy dependence on gender across all conditions did not reveal any significance ($F = 0.84, p = .43$).

In conclusion, despite variations in gender composition across conditions, neither within-condition comparisons nor the gender $\times$ condition analysis revealed performance disparities based on gender. Accordingly, the observed effects of the experimental condition on accuracy can be interpreted independently of gender-related confounding influences.

B.9 The scoring and compensation logic.

This appendix section explains the scoring and compensation framework employed in the collaborative image-labelling system to guarantee transparency, reproducibility, and ethical considerations. The reward structure was intended to justly compensating participants as is expected US minimum wage standards, focussing attention on classification accuracy and agreement as opposed to speed only, and to prevent introducing financial pressure or behavioural bias during task execution. To maintain focus on collaborative task performance, monetary value was not introduced on the task interface. Rather than earning points at the point of play, they became a monetary reward after the task has been completed. This abstraction is consistent with previous ethical guidelines regarding paid crowdsourcing experiments.

Structure and Point Allocation.

Every worker participant annotated 50 images; the images were sorted into 5 levels (10 images per level).

Each level contained:

- standard images.
- 1 honeypot image (expert validated ground truth).

Base Scoring Rules.

For each labelled image:

- Standard Image: +5 points.

- Honeypot Image (Correct Label): + (5,6,7,8,9) points (increased by level)

Such a structure gradually increases of reward across different levels.

Agreement-Based Bonuses

To promote consensus and accuracy rather than rapid individual submission, agreement-based bonuses were limited to team-based conditions:

- Majority agreement ($\geq 50\%$ team selects same label): +1 point (total = 6 points).
- Full agreement (all team members select the same label): +2 points (total = 7 points).

They were earned per image and collected at the team level, but individual contributions were still tracked

Conversion to Monetary Reward.

- Point-to-payment rate: 1 point = USD 0.01
- Post-task survey completion bonus: USD 1.00

Together, this model guarantees that the participants who complete the task and survey receive payment which is at least the equivalent of 30 minutes of work based on the US federal minimum wage, in line with ethical payment principles for online work. To operationalise this guarantee, each worker's points-based earnings were manually checked against their actual time on task, benchmarked against the pilot-study average (30 minutes). Where earnings fell short of the US minimum wage equivalent for that time, the difference was paid immediately as a separate bonus through Amazon Mechanical Turk. The Figure B.13 showed a Team-Based scoring Pseudocode.

```

function ROOM_SCORE_FOR_LEVEL_DEFAULT(room, level, question_set):
    N = 5
    base = 5
    Agreement_bonus = 2
    majority_bonus = 1
    accuracy_score = [5, 6, 7, 8, 9][level - 1] # level 1..5

    questions = question_set.question_ids # length 10 (9 normal + 1 honeypot)
    honeypot_idx = question_set.honeypot_indexes # length 1

    room_delta = 0

    for idx, q in enumerate(questions):
        state = votes(room, q) # worker_id -> (answer, confirmed)

        # base points (always)
        room_delta += 5

        # agreement points
        confirmed_answers = [answer for (answer, confirmed) in state.values if confirmed]
        max_count = max_frequency(confirmed_answers)

        if max_count >= 3: # strict majority of 5
            room_delta += (Agreement_bonus + majority_bonus if max_count == N else
majority_bonus)

        # honeypot points (only on honeypot question)
        if idx in honeypot_idx:
            correct_answers = count_confirmed_correct(state, correct_answer(q))
            room_delta += accuracy_score * correct_answers

    return room_delta

```

Figure B.13: Team-Based scoring Pseudocode

B.10: Survey of Study 3

Q1 Gender: Are you...

- ☐ Male
- ☐ Female
- ☐ Non-Binary
- ☐ Prefer to Self-Describe _____
- ☐ Prefer not to say

Q2 Age: What is your age?

- ☐ 18-24 years old
- ☐ 25-34 years old
- ☐ 35-44 years old
- ☐ 45-54 years old
- ☐ 55-64 years old
- ☐ Over 64 years old
- ☐ Prefer not to say

Q3 Education: What is the highest degree or level of school you have completed? If currently enrolled, highest degree received.

- ☐ No schooling completed
- ☐ Nursery school to 8th grade
- ☐ Some high school, no diploma or leaving qualifications
- ☐ High school graduate, diploma, leaving qualifications or equivalent
- ☐ Some college credit, no degree
- ☐ Trade/technical/vocational training
- ☐ Bachelor's degree
- ☐ Master's degree
- ☐ Professional degree
- ☐ Doctorate degree
- ☐ Prefer not to say

Q4 The following statements ask you to reflect on your experience of engaging with this task. For each statement, please use the following scale to indicate what is most applicable for you.

	Strongly Disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat Agree	Agree	Strongly Agree
I did things in the task because they interested me	☐	☐	☐	☐	☐	☐	☐
I felt controlled and pressured to in a certain way.	☐	☐	☐	☐	☐	☐	☐

I am
satisfied
with my
performance
of the task.

☐	☐	☐	☐	☐	☐	☐	☐
-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------

After doing
the task for a
while, I feel
pretty
competent.

☐	☐	☐	☐	☐	☐	☐	☐
-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------

When I visit
the online
community,
I feel
supported
by other
users.

☐	☐	☐	☐	☐	☐	☐	☐
-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------

When I visit
the online
community,
I feel like
other people
care what I
have to say
and what I
do

☐	☐	☐	☐	☐	☐	☐	☐
-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------

I enjoyed
this task
very much.

☐	☐	☐	☐	☐	☐	☐	☐
-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------

This task
didn't hold
my
attention.

☐	☐	☐	☐	☐	☐	☐	☐
-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------

I felt
frustrated
while
visiting this
task site

☐	☐	☐	☐	☐	☐	☐	☐
-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------

I found this
task site
confusing to
use.

☐	☐	☐	☐	☐	☐	☐	☐
-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------	-----------------------

Q5 How frequently do you use these types of devices for performing crowdsourcing tasks?

	Always	Frequently	Occasionally	Rarely	Very Rarely	Never
Desktop (laptop or personal computer)	☐	☐	☐	☐	☐	☐
Mobile (tablet or smartphone)	☐	☐	☐	☐	☐	☐
Other (please specify)	☐	☐	☐	☐	☐	☐

Q6 How many times have you visited the Mars Planet platform?

- ☐ 0
- ☐ 5
- ☐ 10
- ☐ 15
- ☐ 20

Q7 Is crowdsourcing the main source of your income?

- ☐ No
- ☐ Yes
- ☐ Prefer not to say

Q8 How well do the following statement describe your personality: I see myself as someone who ...

	Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
.. is reserved	☐	☐	☐	☐	☐
.. is generally trusting	☐	☐	☐	☐	☐
..tends to be lazy	☐	☐	☐	☐	☐
..is relaxed, handles stress well	☐	☐	☐	☐	☐
..has few artistic interests	☐	☐	☐	☐	☐
..is outgoing, sociable	☐	☐	☐	☐	☐
..tends to find fault with others	☐	☐	☐	☐	☐
..does a thorough job	☐	☐	☐	☐	☐
..gets nervous easily	☐	☐	☐	☐	☐
..has an active imagination	☐	☐	☐	☐	☐

Q9 These scale questions are regarding: How did you find the task? where 1 is 'hardly at all' and 7 is 'very'.

	1	2	3	4	5	6	7
How mentally demanding was the task?	☐	☐	☐	☐	☐	☐	☐
How physically demanding was the task?	☐	☐	☐	☐	☐	☐	☐
How hurried or rushed was the pace of the task?	☐	☐	☐	☐	☐	☐	☐
How successful were you in accomplishing what you were asked to do?	☐	☐	☐	☐	☐	☐	☐
How hard did you have to work to accomplish your level of performance?	☐	☐	☐	☐	☐	☐	☐
How in secure, discouraged, stressed and annoyed were you?	☐	☐	☐	☐	☐	☐	☐

Q10 These scale questions is regarding how do you found the task elements where 1 is (Strongly Disagree) and 7 is (strongly agree).

	Strongyle Disagree	Disagree	Somewhat disagree	Neural	Somewhat Agree	Agree	Strongly agree
I attach great importance to the avatars on the task	☐	☐	☐	☐	☐	☐	☐
I attach great importance to the Nickname in the task.	☐	☐	☐	☐	☐	☐	☐
I attach great importance to being a levels builder in a game.	☐	☐	☐	☐	☐	☐	☐
I attach great importance to expressing my ideas and interacting with others on the task	☐	☐	☐	☐	☐	☐	☐
I attach great importance to having conversations (chats) with others on the task	☐	☐	☐	☐	☐	☐	☐
I attach great importance to following others' advice or being followed	☐	☐	☐	☐	☐	☐	☐
I attach great importance to my score on the task	☐	☐	☐	☐	☐	☐	☐

I often attend to my team score and leaderboard on the task.

☐☐☐☐☐☐☐

I often attend to my team levels on the task.

☐☐☐☐☐☐☐

I have spent more time on the task than expected

☐☐☐☐☐☐☐

I feel addicted to the task

☐☐☐☐☐☐☐

Q11 Is there any feedback, opinion, or question about the task you want to share with us?

Important: Please write your Amazon worker' ID here to get your bonus

Here is your ID [\\${e://Field/Random%20ID}](#)

Please Copy this value into MTurk

When you have copied this ID, please click the 'next' button to submit your survey.