The Wizard of Incentive

A Guiding Tool for the Design of Incentive Formulas in Crowdsourcing

Master Thesis

Macsim Violeta-Mara

Delft University of Technology

TU Delft

The Wizard of Incentive

A Guiding Tool for the Design of Incentive Formulas in Crowdsourcing

by

Macsim Violeta-Mara

to obtain the degree of Master of Science

at the Delft University of Technology,

to be defended publicly on Monday January 1, 2024 at 10:00 AM.

Student number:	5498031	
Project duration:	November 10, 2025 – June 25, 2026	
Thesis committee:	Dr. Ujwal Gadiraju (Advisor & Chair)	TU Delft
	Dr. Myrthe Tielman (External Examiner)	TU Delft

TU Delft Delft University of Technology

Preface

This thesis marks the end of my Master's in Computer Science at Delft University of Technology. When I first encountered the topic of crowdsourcing incentive design, I was drawn to the possibility of working on something with real impact. Behind every AI system being built today, there are people doing the quiet, unglamorous work of labeling data and validating outputs, which has little regulatory protection and few guarantees of fair compensation. Better compensated and more motivated workers produce higher quality data, which in turn leads to more reliable and robust intelligent systems. In this way, improving incentive design is not just about protecting workers — it is also about helping the tech industry build better products. Supporting both sides of this equation felt like a direction worth pursuing.

The process of building and evaluating the Wizard of Incentive was as much a learning experience as it was a research exercise. I am grateful to my supervisor, Dr. Ujwal Gadiraju, for his guidance and patience throughout this project, and to Dr. Myrthe Tielman for her valuable feedback. I'd also like to thank everyone who participated in my user study and helped me collect data for my research, as without you, I wouldn't have been able to finish this chapter.

Macsim Violeta-Mara
Delft, June 2026

Summary

The rapid growth of artificial intelligence has driven demand for large volumes of real-world data, making crowdsourcing an essential practice. However, crowdsourcing remains largely unregulated, with minimal disclosure of compensation practices in academic literature or dataset documentation. This lack of transparency undermines two important goals: collecting high-quality, realistic data for AI systems, and ensuring fair treatment of workers. Without clear guidance on incentive design, it becomes difficult to distinguish between requesters' lack of knowledge and poor practices—a problem that affects both data quality and worker welfare.

To address this gap, a wizard tool was developed to guide requesters through the process of designing payment schemas for crowdsourcing tasks. A user study was conducted to investigate how structured guidance affects incentive design: first, by comparing designs created with and without the tool, and second, by examining whether the tool produces consistency in compensation decisions across different requesters. The study evaluated both the designs participants created and their feedback on the tool itself.

The analysis reveals three primary insights. First, the tool's primary strength lies in structuring the design process rather than fundamentally altering participants' compensation decisions. The extent to which structured guidance benefited participants depended significantly on their prior experience with crowdsourcing, suggesting that the tool's value is contingent on user expertise. Second, the tool produced convergence around a limited set of high-level design elements, though participants used varied implementation approaches within these patterns, such as specific bonus sums. Lastly, the tool was not met with immediate rejection across different user types: participants found the information it provided useful and it notably reduced the time needed to produce a basic incentive design.

These findings indicate that the tool could serve a valuable function in documenting and contextualizing design rationales, capturing the constraints and considerations that shaped dataset creation decisions. However, realizing the tool's full potential as a design aid requires enhancements to customization options and user experience refinement. Despite these limitations, the tool shows promise as an educational resource for introducing beginners to crowdsourcing incentive design, offering a structured entry point into a complex domain.

Contents

1

Introduction

Over the past decade, Artificial Intelligence (AI) has rapidly become a central topic of interest across numerous disciplines [32]. This increasing attention has been accompanied by a growing demand for large volumes of real-world data [69], which are essential for the deployment pipeline of AI systems into everyday life. Researchers have increasingly leveraged the vast number and diversity of online users to collect data for training machine learning models – a process commonly referred to as *crowdsourcing*. In effect, crowdsourced datasets form the foundation of many AI pipelines, supporting tasks such as data labeling, validation, and model fine-tuning that lead to a better generalization of machine learning (ML) systems [73].

Crowdsourcing marketplaces, such as *Amazon Mechanical Turk*[1], *Prolific*[2] or *Appen*[3] (formarly known as *Figure Eight*) marked a shift in crowdsourcing by introducing financial compensation as an incentive for participation, transforming what was once primarily a voluntary activity. Since then, incentive mechanisms have become a standard practice in the design of crowd-based tasks, where a job typically consists of a small, well-defined task or piece of work completed by an online participant.

Within Human Computation and Human–Computer Interaction, researchers are actively investigating how to design crowd computing tasks that maximize mutual benefit for all involved parties [53, 85]. Yet, although incentive structures play a crucial role in shaping participation and job outcomes [18], they are frequently underreported or described only in broad terms [50], limiting the reproducibility and cumulative progress of research in this area.

1.1. Payment - A Taboo Subject

Research aimed at improving crowdsourcing outcomes is extensive, spanning worker–task matching based on skill or task complexity [23, 38, 15], studies on how to formulate clear instructions and design effective user interfaces [72], and even work into additional motivators that complement monetary compensation [1, 47, 75]. Throughout these, one common issue persists: when payment is not the central focus of the study, descriptions of worker compensation are often described vaguely or entirely omitted [65].

This lack of transparency can be observed even in studies where monetary reward likely influences the very phenomena being measured. For instance, Difallah et al. [23] and Bozzon et al. [15] do not report any details about their payment schemes, despite examining task–worker recommendation. In this circumstance, the engagement is still up to the worker's choice of partaking, which could be tied to expected earnings. Such omissions make it difficult to contextualize findings or understand whether compensation may have shaped participation patterns, worker behavior, or data quality.

Most crowdsourcing platforms provide limited guidance on how to design fair or well-structured compensation schemes, and few enforce any form of standardized payment. Because this type of labor remains largely unregulated [61, 30], requesters are not obliged to comply with income standards or to disclose how rewards

[1]Amazon Mechanical Turk: https://www.mturk.com/
[2]Prolific: https://www.prolific.com/
[3]Appen: https://www.appen.com/

are present and how they are determined. The absence of guidance does not stem from an intentional effort to grant utmost freedom to requesters; rather, platforms offer resources for task implementation and setup [4] and best-practice recommendations for improving data quality [6], while refraining from providing comparable support for compensation design.

1.2. Study Motivation

As the ongoing "AI boom" continues to drive demand for large and diverse datasets, reliance on crowdsourcing is expected to grow accordingly [27]; this trend that has been visible since the early stages of web-based crowdsourcing. For many individuals, participation in these platforms represents a meaningful source of income [30], which highlights the importance of payment structures in the design of crowdsourcing jobs. Despite the extensive usage of this form of labor, crowdsourcing remains largely unregulated, leaving workers vulnerable to unfair compensation and inadequate treatment. From a requester's perpective, priority is given more to defining task objectives and data collection strategies over determining how workers should be compensated.

At the same time, crowdsourcing platforms provide limited support for compensation design. Although platforms may enforce minimum payment requirements, they offer little to no guidance on how to design fair and effective compensation schemes. For individuals with limited or no prior experience in crowdsourcing, this lack of structured support can make the process feel overwhelming. The absence of tutorials, practical recommendations, and transparent examples from previously published jobs leaves requesters without clear reference points for informed decision-making. As a result, compensation choices may be made without fully understanding their implications, and this seemingly benign lack of knowledge can easily lead to workers being underpaid, often unintentionally [61].

Although crowdcomputing tasks differ in purpose, they share a common goal: to obtain high-quality data efficiently from human contributors. Motivation therefore plays an essential role in achieving reliable results, and challenges appear when the objectives of requesters and workers do not align. Literature remains divided on whether monetary incentives reliably improve data quality. Yin et al. [87] report that occasional bonuses can enhance worker performance, and Kazushi and Bernstein [42] find that even non-monetary rewards, such as physical goods, can decrease engagement. On the hand, Litman et al. [58] showed that higher pay may attract spammers or bots, ultimately having negative impacts on the crowdcomputing pipeline.

Considering these mixed feelings, monetary compensation cannot yet be dismissed as an important component of job design. Besides having a great impact on output quality, payment plays a significant role in who will participate in a study [62]. Intelligent systems trained on crowdsourced datasets are at risk of propagating false assumptions if the underlying data is uneven, faulty, or prejudiced [7].

Despite its importance, data quality considerations are not consistently reflected in academic literature and reports on crowdsourcing studies, where payment practices are often insufficiently documented, even though compensation constitutes a central element of the experimental design. This lack of clear and structured reporting hinders the reproducibility of prior studies and limits the ability to build reliably on existing findings.

The importance of data quality is also not reflected in academic literature and reports on crowdsourcing studies: they do not consistently document payment practices, even though compensation is a core component of the experimental design. The lack of clear and structured reporting makes it difficult to reproduce prior studies and to build reliably on existing findings.

As of recently, guidance on what should be included in such reporting already exists. S. Kaur [50] provides an overview of how incentives are currently structured in crowdsourcing research and presents a streamlined checklist of components that together form a complete incentive schema. This raises a new set of questions: what might a tool designed to support researchers in creating incentive schemas look like? How could it shape the way that participant rewarding looks like?

1.3. Research Scope

The term incentive encompasses a wide range of mechanisms for motivating participation in crowdsourcing systems (as discussed in section 2.2). Among these, monetary payment remains a central and widely used driver of worker engagement in today's digital platforms. For this reason, this study focuses on how to assist

people in designing paid incentives for workers and how compensation can be expressed in a clearer and more consistent way.

Given the ongoing evolution of crowdsourcing research, this study focuses specifically on compensation design within academic crowdsourcing contexts, with the aim of improving data collection practices and provide bigger worker satisfaction. Due to the wide range of platforms and online environments in which crowdsourcing jobs can be deployed, each with its own requirements, the scope is further restricted to scenarios involving *Prolific*, a platform known for enforcing a minimum compensation rate based on estimated task completion time in order to publish.

The main goal of this work is straightforward: to provide a the foundations for tool that addresses existing gaps in incentive design for crowdsourcing. When considered more broadly, however, this goal encompasses several related objectives. First, the tool aims to help requesters design incentive schemes more efficiently by reducing the effort needed to determine appropriate compensation, while supporting the creation of task setups that lead to higher-quality data. Second, it seeks to encourage fairer and more suitable compensation practices for workers by promoting more deliberate and informed decision-making.

Beyond individual job design, this work also aims to provide a starting point for reducing ambiguity in how incentive structures are reported and for improving transparency around dataset creation through the development of this tool. Empirical validation of the tool's effectiveness in this trajectory is beyond the scope of this study, as it would require a multi-stage user evaluation with more detailed and time-intensive tasks requested from the participants from the initial stage. Due to time constraints and concerns about participant fatigue which can compromise the validity of the results, the evaluation was restricted to a single session focused on users' interactions with the tool.

Based on this, the study aims to answer the following research questions:

RQ1: How do descriptive and prescriptive incentive guidance tools affect payment structures designed by task requesters?

Incentive design in crowdsourcing has traditionally been carried out without formal guidelines, limited input from domain experts, scarce educational resources, and minimal support from crowdsourcing platforms. As a result, requesters often rely primarily on personal judgment and intuition when deciding how to compensate workers. Without a formed mental model to build on, repeated exposure to the same design decisions can gradually consolidate into habit of relying on the same design that has been proven to work before. This raises the question of whether additional guidance during the incentive design process can encourage requesters to consider a broader range of available options and construct more nuanced payment structures. Is this observation a result of a lack of awareness of alternative incentive components and their possibilities, such as bonuses, performance-based rewards, and the incorporation of metrics, or does it simply reflect a preferred method or a simple solution for spending less time on this mandatory component of deployment? For individuals new to this domain, would a simple "nudge" be sufficient to help them achieve better outcomes from the outset, with less need for trial and error learning? This study compares different levels of guidance to determine whether increased support leads to more complex and thoughtfully structured payment schemes, or whether requesters continue to prefer simpler approaches regardless of the assistance provided.

RQ2: How do descriptive and prescriptive incentive guidance tools affect the variability of solutions across different task requesters?

Incentive design in crowdsourcing closely resembles an ill-structured problem — a situation that lacks a single correct solution and does not provide fixed or universally accepted evaluation criteria [46]. Under these conditions, similar tasks can be expected to yield a wide range of compensation strategies, shaped by individual interpretations, assumptions, and judgment. A key reason to study solution variability is therefore to understand whether guidance tools can meaningfully counteract this ill-structuredness, or whether requesters continue to diverge even in their presence. When a tool presents a bounded set of incentive options, it narrows the design space, but it remains an open question whether this actually brings requesters closer together in practice. Would they gravitate toward similar choices when options are constrained, or continue

to spread across the available range regardless? Requesters might still diverge considerably, anchoring on different defaults or interpreting the same guidance in different ways, making it important to empirically test how much of that variability guidance tools can actually absorb.

Beyond the design process itself, variability in incentive schemes has direct consequences for the workers on the receiving end. When two requesters independently design incentives for tasks of equivalent complexity and effort, the resulting schemes should ideally be comparable; yet without guidance, there is little reason to expect they will be. High cross-requester variability means that workers performing equivalent work may be compensated very differently depending on who happened to post the task, making variability a fairness concern independent of whether any individual scheme is well-designed. This inequity also has consequences at the platform level: when incentive structures are inconsistent across requesters, workers cannot reliably anticipate compensation for a given task type, which undermines their ability to make informed participation decisions. Guidance tools that reduce this variability therefore carry implications not only for individual requesters, but for the predictability and fairness of the crowdsourcing ecosystem as a whole.

1.4. Summary of Contributions

This research makes three primary contributions to crowdsourcing research and practice.

First, this research introduces The Wizard of Incentive, a guidance tool designed to support the creation of incentive structures in crowdsourcing. Crowdsourcing platforms typically provide resources for task design and quality management, but offer little guidance on how to design fair compensation schemes. This tool addresses this gap by walking requesters through key decisions about base pay, bonuses, penalties, and payment logic. By doing so, the wizard helps designers surface and articulate design considerations that they might otherwise miss or struggle to put into words.

Second, the study provides an empirical analysis of how incentive guidance tools affect the shape of crowdsourcing incentive designs. Through a user study, it examines how adding a layer of descriptive or prescriptive guidance transforms the way monetary incentives are applied — what the resulting designs look like and to what extent personal interpretation and subjective judgment still drive divergence when the design space is explicitly constrained. This directly addresses the ill-structured nature of incentive design, where the absence of a single correct solution leaves room for individual assumptions to shape outcomes, and raises the question of how much that changes when a tool is introduced.

Third, foundational knowledge is provided on the viability of tool-assisted incentive design. While adoption challenges might persist, as with any new tool introduced into established workflows, this research examines how such guidance is perceived by users and identifies design gaps that limit adoption. The findings establish a foundation for determining whether continued investment in tool development is warranted, and what design improvements would increase adoption and effectiveness across different user groups.

1.5. Thesis Structure

This thesis is structured as follows. chapter 2 presents the theoretical foundations underlying crowdsourcing and incentive design, establishing the conceptual framework for the study. chapter 3 then describes the design and implementation of the tool, outlining its structure, functionality, and the rationale behind its development. chapter 4 details the research design, including the experimental setup, participant procedures, data collection methods, and analytical approach. chapter 5 reports the findings of the study, presenting both qualitative analyses and supporting indicators. chapter 6 interprets these findings in light of the research questions and existing literature, while also addressing the limitations of the study. Finally, chapter 7 summarizes the main contributions and outlines directions for future research.

2

Theoretical Background

This chapter establishes the conceptual foundations underlying incentive design in crowdsourcing. It begins by tracing the evolution from volunteer-driven initiatives to monetized labor markets (section 2.1), then introduces the key elements that comprise incentive structures (section 2.3). Theoretical and empirical research on how incentives shape worker behavior is examined in section 2.2, while section 2.4 surveys existing guidance and tools for job design. By identifying gaps in support for compensation design, the chapter motivates the need for the tool developed in this research.

2.1. From Volunteering to Monetization

The concept of *crowdsourcing* first appeared in 2006 in *Wired Magazine*, where Jeff Howe described it as *"the act of taking a job traditionally performed by an employee or contractor and outsourcing it to an undefined, generally large group of people [41]."* At that time, the concept itself was still not clearly defined, even though it was already widely practiced. Open-source platforms like *Wikipedia* and *Waze* [1], participatory product design studies [14, 13], and early "citizen science" projects [26] were all thriving around the 2000s – well before they were formally recognized as part of the crowdsourcing paradigm.

Participation in these early initiatives was primarily motivated by curiosity, collaboration, or the sense of shared community rather than financial incentives. Contributors were motivated by the desire to support a successful outcome or contribute to a broader societal good. However, the new emphasis on machine learning revealed a potential for practical and commercial value of large-scale human input [58], especially for data collection, labeling, and validation.

The launch of *Amazon Mechanical Turk* in 2005 marked one of the first major steps in integrating this long-practice form of collective work within Web 2.0 [17], making it easier to share paid microtasks with a broad and distributed online crowd [53]. This new scalability brought the concept of *human computation* proposed by von Ahn [2] into practice, paving the way for broader applications of human contribution in AI development. Following this milestone, several crowdsourcing marketplaces were released (e.g. *Clickworker* [2], *Appen*, *Prolific*, etc.), formalizing the practices of this method of online collaboration. Beyond their technological significance, these platforms gave rise to new forms of digital labor and economic opportunities [30].

The crowdsourcing process involves two main actors: the **requester** (task initiator) and the **worker** (participant). It begins when a requester posts a task on an online platform, making it accessible to registered users. The nature of these tasks can vary, commonly including activities such as classification, survey completion, or data annotation [67]. Workers browse available tasks, select those they wish to complete, and carry out the required work in accordance with the requester's instructions. Once submitted, the requester can review the output against predefined criteria or quality benchmarks to determine whether it meets the required standards. Payment is issued to the worker once the task is verified and accepted as successfully completed [43].

[1] **Waze:** https://www.waze.com/company
[2] **Clickworker:** https://www.clickworker.com/

2.2. Incentives in Crowdsourcing

In the current context of online-deployed microtasks, money is a primary motivator [88], but not a sole driver of participation, motivation, retention and quality [10]. This has lead researchers to explore a wide range of options that capture both direct reward structures and indirect behavioral drivers embedded in platform design.

Monetary incentives Such rewards represent the most common form of motivation in online crowdsourcing platforms, which require payment for all tasks (*Prolific* even suggesting a minimum sum based on their definition of fairness [36]). A typical payment structure includes a base pay [33, 16], either set as a fixed amount or derived from an estimated hourly rate, and may be complemented by bonuses that reward specific behaviors or results [87]. These bonus schemes can vary considerably, ranging from random-based awarding to performance-dependent incentives [39] or additional payments for completing particular subtasks.

Material incentives This type of incentive refers to non-monetary, tangible rewards offered to workers in exchange for their contributions. They may include gift cards [42], discount vouchers, symbolic items or merchandise. Such rewards can sometimes be perceived as a more personal or thoughtful expression of appreciation compared to standard monetary payment. Their main limitation, however, stems from the variety of worker preferences; what requesters give as options may be irrelevant or unappealing to them.

Intrinsic motivation This type of motivation attracts workers to participate as they find the activity interesting, enjoyable or personally meaningful, much like the early ethos of volunteer-based crowdsourcing. Examples of sources of motivation can be the desire to learn a new skill [49], enjoyment and challenging tasks or satisfaction sourced from contributing to a meaningful project or social cause [74].

Gamification Elements of play, competition or recreational challenges are integrated into a task to improve worker focus and motivation. This can involve comparison through point systems and leaderboards [25], mission-based goals and achievements [55], progress feedback or framing the task within a compelling story [78]. These aim to make the environment more enjoyable and reduce the sense of obligation and routine labor, supporting sustained motivation [3].

Reputation This form of reward highlights workers through status markers like approval rates, badges, or expertise tiers, which determine their access to tasks. Since platforms have the option to reserve higher-quality or better-paying jobs for workers with strong reputations, these markers provide a powerful incentive to perform well [44].

2.3. Elements of Incentive Formulas

Kaur's literature review [50] highlights the multilayered nature of monetary incentive structures in crowdsourcing. The review identifies two fundamental components: a base pay, which is typically required by most platforms, and bonuses, which are optional and whose complexity depends entirely on the requester's design choices. Given the number of possible variations and design decisions involved, it is unsurprising that many requesters rely solely on base pay and overlook opportunities to construct more tailored or strategic reward schemes.

Base pay The base payment is generally mandatory for publishing a task on crowdsourcing platforms and represents the minimum compensation a worker receives upon completion. Requesters commonly determine this amount in one of two ways: either by using an inferred pay rate based on external standards or heuristics (such as estimated hourly wages or task duration), or by assigning a fixed completion reward independent of performance. Despite its apparent simplicity, the base pay plays a crucial role in shaping expectations and effectively "advertising" a task [62], as it is the amount clearly shown when a job is listed on the platform. Workers typically interpret this value as guaranteed income and use it as a reference point when planning and pursuing their earning goals [40].

Bonuses Bonus rewards are an optional component that introduce additional flexibility into incentive design and tend to be more complex to compute. They can be used to reinforce high-quality work, reward task completion, or even motivate participation before the work begins, as in the case of lottery-style incentives. Performance-based bonuses may take both positive and negative forms: workers may earn additional compensation for strong performance, while cheating or poor-quality submissions may result in reduced bonus payouts (though never reducing the base pay). Performance can be evaluated using predefined thresholds or dynamic scoring systems, and while various metrics can be employed, accuracy remains one of the most commonly used indicators when determining bonus eligibility.

Motivation behind the reward When creating a monetary incentive scheme, the requester must consider not only which elements to include (for example, base pay or bonus structures), but also why each component is being used. Motivation provides the fundamental logic that guides payment design, and thus plays an important role in shaping the incentive formula, despite not being a component itself.

Research shows that different payment choices signal different expectations to workers. Mason and Watts [62] demonstrate that higher base pay attracts more workers and sets an anchor for how much effort workers believe is appropriate. This is particularly relevant for requesters who need rapid recruitment or high task throughput, since a stronger base payment not only increases participation but also encourages workers to take on more tasks. In contrast, if the requester only needs feedback, or the task is short and straightforward, investing heavily in base pay may not meaningfully influence results.

Performance motivation introduces additional considerations. While Mason and Watts find no clear link between higher pay and accuracy, later work suggests that increased pay can draw in a broader range of workers [68]. Depending on the task, this can be beneficial—diversity can improve idea generation or robustness—but it can also mean that workers with lower expertise attempt tasks that require specialized judgment. In such cases, bonuses tied to quality may be more appropriate than raising the base payment, as they provide a targeted mechanism to motivate careful or skilled work without inflating the reward for all participants.

Bonuses themselves can serve multiple motivational roles: encouraging high-quality performance, rewarding extra effort, stimulating competition, or signaling that accuracy matters. Kaur [50] notes that complex bonus schemes are used to make workers more aware of accuracy expectations and more deliberate in their responses. Despite this, only about a third of the reviewed studies explicitly discussed the intended goal behind using such incentives. This suggests that researchers may either overlook the motivational rationale behind bonus design or consider it secondary, despite prior work showing that these intentions can actually influence worker behavior.

Incorporating bonuses into an incentive design therefore requires requesters to decide what type of behavior they aim to encourage (eg. precision, speed, thoroughness, or engagement) and which bonus structure best aligns with that goal. Prompting users to reflect on these goals from the beginning appears to be essential: it establishes a clear rationale that can help them decide what base pay is appropriate, whether a bonus is necessary at all and, if so, how it should be structured to effectively serve its purpose.

Communication of the reward Beyond the design of the reward itself, the way compensation is communicated to workers plays an important role in shaping their behavior and their perception of the task's stakes. Clear and transparent communication has been shown to foster trust between requesters and workers [9]. In a setting where many workers seek to meet self-defined income goals [40], such trust can soften the purely transactional nature of the requester–worker relationship. Workers can also rely on online community platforms, such as Turkopticon[3] and MTurk Crowd[4], to share experiences with requesters and signal fair or unfair practices to others. As a result, treating workers fairly and communicating expectations clearly can enhance a requester's visibility and reach, as positive reputations may encourage broader participation and sustained trust among workers.

Effective communication of compensation details also directly influences worker behavior during task execution. Explicitly stating potential penalties can discourage careless or dishonest responses, while clearly described performance-based bonuses can motivate workers to invest greater effort. Although workers generally expect to receive payment upon task completion, transparency around reward structures, bonus

[3] **Turkopticon:** https://turkopticon.net/
[4] **MTurk Crowd:** https://www.mturkcrowd.com/

conditions, and evaluation criteria helps align expectations and reinforces trust in the task design. Together, these communication practices support both fair treatment of workers and the collection of higher-quality data.

2.4. Research on Monetary Incentive Tuning

It is widely acknowledged that most workers use crowdsourcing platforms as a source of income [5]. In this context, financial incentives have become a central motivational factor, directly influencing worker engagement. This has motivated extensive research into how payment structures can be optimized to meet the needs of both workers and task designers, resulting in a wide range of compensation strategies. Although efforts have also focused on understanding how compensation influences data quality, there is no clear consensus [42].

Apart from figuring out an appropriate base pay, one other trajectory of research was identifying a good way of paying workers: either telling workers during the job that they are getting awarded, or implicitly keeping the after-completion rewarding. Ikeda and Bersnstein [42] found that letting workers know the added payment sum after completing 10 sub-tasks led to 15% more completed tasks.

Some researchers also explore reward uncertainty and risk-based incentives to examine their impact on worker behavior. Rula et al. [76] tested weighted lottery payments, where workers who completed five tasks increased their chances of winning a gift card. They found that participation rose, although the overall quality of submissions was lower than under standard micro-payments. Tie et al. [60] instead implemented an *all-pay* mechanism, in which only a small number of top submissions received rewards. Their results showed that a subset of workers, were notably more motivated and engaged under this scheme, achieving the best results on highly-skilled workers. At the same time, the highest-quality submissions improved, while keeping the required funds for the job low, since only the best contributors needed to be compensated.

There is also lots of research examining how bonuses can motivate workers to produce higher-quality work rather than large quantities of low-value submissions, yet findings remain mixed [39, 87]. Ho et al. [39] reported that workers who received no bonuses performed worse than those who obtained at least some bonus reward. However, simply making workers explicitly aware of the bonus structure did not significantly improve performance. This pattern may be influenced by how online platforms administer bonuses: requesters must review a submission before deciding whether to award a bonus at all. As a result, workers may behave more attentively not because of the bonus itself, but because of the implicit sense of being "evaluated".

An interesting alternative "incentive" strategy is introduced by C. Harris [34], who adapts the corporate practice of *pay-to-quit* to crowdsourcing. He shows that the effectiveness of this method depends heavily on how the offer is presented visually, as workers may easily ignore or dismiss the notification. Simply asking workers whether they wished to quit had little impact, as those performing poorly often chose to continue regardless. However, when the quit option was accompanied by feedback comparing their performance to that of other workers, low-performing participants were more likely to recognize their lack of abilities and opt out.

2.5. Guidance for Job Design

A crowdsourcing job consists of several interconnected components; among them are the content and presentation of the task, the workflow structure, and the compensation offered to workers [66]. Its success depends on achieving an effective balance across these elements. Consequently, substantial research has focused on developing tools and frameworks that support requesters in designing better tasks.

Tools for improving task design are very common. *CrowdForge* [51], for example, allows requesters to decompose complex tasks into smaller, manageable micro-tasks and later integrate their results into a coherent final output. This approach has made it possible to design jobs that go beyond simple labeling, such as creating reasoning datasets or generating paraphrases. *Crowdweaver* [52] extends this idea by offering a visual interface for organizing and orchestrating task decomposition pipelines. It allows requesters to specify which steps require human input and which are automated, and provides fine-grained control over parameters such as inter-worker agreement thresholds and acceptable amount of time since nobody has started the job.

AUTOMAN [8] represents a notable advancement in tooling for crowdsourcing: it integrates human compu-

tation directly into a programming language while guaranteeing a target level of answer quality. The system dynamically adjusts payment when current responses appear unreliable to improve the chances of attracting highly-skilled workers and continuously estimates answer correctness in real time by comparing contributions from multiple workers. A task remains active until *AUTOMAN* determines that the required confidence threshold has been met.

In contrast, resources that assist researchers in constructing payment formulas are scarce. Although many crowdsourcing tasks rely on monetary incentives, existing materials typically offer only high-level recommendations rather than concrete guidance on how to structure compensation. For example, Northwestern University's "Guidelines for Academic Requesters" [24] provide valuable advice on deploying tasks on Amazon Mechanical Turk, including several points directly or indirectly related to fair pay. However, these guidelines do not explain how to assemble a coherent payment schema. Similarly, the platform Prolific outlines its principles for fair compensation [71], focusing on minimum wage expectations, and provides a separate article detailing how to issue bonus payments [70]. Yet, these documents offer little context on how bonuses should be integrated into an overall incentive structure or how the different components of compensation interact to influence worker motivation and task outcomes.

There aren't many examples of tools developed to aid this step of the process. Among these, an example is *Fair Work* [86], which adjusts payments based on observed task duration to ensure that each worker earns at least a minimum-wage-equivalent rate. However, this approach relies heavily on accurately estimating task duration, an assumption that may not scale reliably across diverse tasks, requester practices, or worker populations.

3

Designing the Tool

This chapter presents a proposed solution to the previously identified gaps in crowdsourcing and describes its conceptual design. It begins by presenting the tool itself in section 3.1, followed by a discussion in section 3.2 on why a wizard-based structure is well suited for this context. section 3.3 then outlines the intended user flow of the tool, and section 3.4 explains how the language of the content is adapted to represent different levels of guidance. It ends with explaining in section 3.5 how the tool will provide support for reproducibility of the formula.

3.1. The Wizard of Incentive

One of the main challenges in designing monetary incentives for crowdsourcing tasks is the limited availability of shared knowledge and practical guidance from experienced practitioners. As a result, requesters are frequently forced to rely on their own judgment when deciding how to compensate workers and do not form a mental model of this process. This lack of support can lead to suboptimal incentive structures, which may negatively affect task outcomes, worker motivation, and the perceived fairness of compensation.

To address these challenges, a tool is proposed to support the incentive design process. *The Wizard of Incentive* [1] is specifically designed to guide requesters through this aspect of a crowdsourcing job. It can be used both as a preparatory aid, helping users explore available options before finalizing their incentive strategy, and as a reflective tool, allowing requesters to review and reassess an existing approach. By offering structured guidance without enforcing a single correct solution, the wizard aims to improve decision-making while preserving flexibility and autonomy. Visual examples of the tool's design can be found in Appendix A.

3.2. Why a Wizard?

Designing incentives for crowdsourcing is a comprehensive process that requires several interlinked decisions. A natural sequence exists to these decisions, which, if followed, can make the reasoning clearer and help requesters assemble a coherent incentive scheme.

A software wizard becomes particularly useful in this context. As described by the Nielsen Norman Group (NN Group) [31], wizards are an interaction design pattern that guide users through a process step by step, presenting only the information and choices relevant to each stage. By focusing user attention on one part of the process at a time, a wizard reduces cognitive load and prevents unnecessary distraction from details that are not yet relevant or no longer relevant.

Compared to a traditional form that presents all fields at once, a wizard helps users avoid feeling overwhelmed, especially those unfamiliar with incentive design. For inexperienced requesters, seeing the full complexity of the process upfront may lead them to fall back on minimal or default choices simply to complete the task. This could result in suboptimal incentive schemes or compensation practices that workers perceive as unfair. The NN Group highlights this use case as one where wizards can meaningfully improve user experience by breaking complex workflows into manageable steps.

[1] You can download or access the tool at: https://github.com/VIO47/TheWizardOfIncentive

Beyond simplifying the process, a wizard can support learning. Because it mirrors the underlying logic of incentive design, it can help requesters build a clearer mental model of how payment structures work and what considerations should guide each decision. In this sense, the wizard not only assists in producing a well-designed formula but also trains users to understand and replicate the reasoning behind it [64]. Thus, the wizard supports a longer-term goal: encouraging users to reflect more deeply and make more intentional decisions about their incentive choices, articulating the reasoning behind effective motivation strategies.

3.3. User Flow

The tool is designed to support users in constructing purposeful incentive schemes that are well aligned with the needs of their specific task. The core elements remain consistent across most circumstances: there is always a base pay that can be augmented by bonuses, all in pursuit of achieving a goal, and communicated to workers. It is ultimately the responsibility of the researcher to interpret these components and decide how they should be adapted to their particular context. Figure 3.1 describes the proposed flow in any scenario of using the tool.

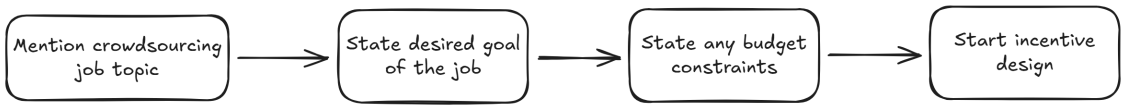

Figure 3.1: Steps taken during the usage of the tool

First, users are asked to summarize the intent of the crowdsourcing job. This step encourages consideration of the task as a whole, including the type of data to be collected, the specificity or niche nature of the topic, the expected difficulty of the task, and the time it may require. This initial reflection supports the subsequent stages of the process by helping users identify the most important aspects of the task—elements that will later inform and guide decisions in following steps.

Next, users are prompted to reflect on the underlying goals of compensating workers. This step encourages consideration of whether additional motivations are intended when offering payment, beyond simply publishing the job. It aims to raise awareness of whether a different motivational environment is needed: one that goes beyond attracting participation and instead influences how workers engage with the task. In some cases, this may lead to a more complex incentive structure, designed to better align workers with a specific context or to increase the perceived importance of the task.

Subsequently, users are prompted to consider whether constraints exist on the amount of monetary flexibility available. This includes reflecting on whether the budget is strict or whether it can be exceeded slightly if needed. This step is important not only for helping requesters remain aware of potential limitations during the design process, but also for providing context to others who may later read the job design description, making it easier to understand why certain compensation amounts or incentive structures were chosen.

After these three stages, the actual incentive design begins.

3.4. Language towards guidance

One of the uncertainties highlighted in section 1.3 concerns the appropriate amount of guidance to provide without overly constraining users toward a single line of thinking. Consequently, an important aspect to examine is how this factor influences the tool's effectiveness, ensuring that it supports decision-making while preserving flexibility in the design process.

The introduction of a wizard-based tool represents a shift from the standard workflow by offering structured support. Within this framework, the language and phrasing of the content constitute a key adjustable element. Prior work has demonstrated that different ways of framing the same information can influence preferences and decision-making, even among the same individuals [84]. If language has the capacity to shape perspectives, it may also affect how much contextual and domain-specific information about crowdsourcing users are exposed to during the design process.

3.4.1. Descriptive language

In linguistics, a descriptive approach to grammar refers to how language is used to express states, events, and relationships [37]. This perspective prioritizes explanation over evaluation and avoids making normative judgments about what defines a correct or desirable mode of expression. When used extensively, descriptive language can serve as a technique for providing impartial and objective support. Individuals can reason about a problem autonomously when the available solutions at a given stage are outlined without being labeled as right or wrong. Descriptive techniques, which strive for clarification rather than direction [56], are particularly effective for environments with numerous legitimate answers.

In this context, the initial level of guidance adopts a descriptive approach. Each question presents only the information necessary for users to reflect on the range of possibilities relevant to that stage of the design process, without providing additional explanations or justifications for why the step is included. An example can be seen in Figure 3.2. While minimal, this form of support still represents an improvement over having no guidance at all, as it introduces structure and direction that would otherwise be absent without the tool.

How will you determine the amount you will award as fixed base pay?
- Just use a flat amount of money
- Estimate pay based on expected completion time

Figure 3.2: Example of descriptive style question

3.4.2. Prescriptive language

In contrast to descriptive approaches, *prescriptive* grammar focuses on establishing norms that specify how actions should be performed [37]. Prescriptive language emphasizes correctness, standardization, and adherence to defined rules, often evaluating alternatives in terms of appropriateness or quality. As a result, it does not merely describe available options, but actively guides behavior by signaling preferred or acceptable choices.

How will you determine the amount you will award as fixed base pay?
Fixed payments are easy to implement, but you must ensure they accurately reflect the effort required to complete the task.
- Just use a flat amount of money
- Estimate pay based on expected completion time

Figure 3.3: Example of prescriptive style question

When applied within the wizard to shape how the content is presented, this approach provides users with more explicit guidance regarding the available options, while still allowing them to decide how to address their specific scenario. An excerpt from the content is shown in Figure 3.3. This form of support represents a more direct level of handholding, offering opportunities for learning through detailed explanations, such as outlining potential advantages and disadvantages or referencing practices supported by prior research and expert experience. However, this level of guidance may also introduce bias into the final outcomes. In some cases, general rules of thumb may not fully apply to a particular context, and users may be inclined to follow the suggested options rather than explore alternative solutions, potentially limiting creativity and diversity in the resulting designs [64].

3.5. Formula summary

Another important dimension of support within the design process is raising awareness of which information is essential to document and communicate in a final report. Reproducibility relies heavily on specificity [19], as all factors that may influence the outcome need to be clearly described in order to enable others to replicate the results under similar conditions.

Although the step-by-step structure of the tool already prompts users to reflect on individual elements as they progress through the design process, a final summary stage remains valuable. Presenting a final overview

of all questions and corresponding answers allows users to revisit decisions made earlier, which may otherwise be forgotten when the process takes a longer period of time. Seeing all questions and answers organized in a list also highlights the complexity of the incentive structure, as users can see how many different elements are involved. By showing these elements again, the summary stage helps users remember what they previously decided and increases the chance that important details of the incentive formula are reported.

4

Evaluation

This chapter presents the evaluation of the proposed incentive design wizard and outlines how the research questions are addressed. It begins with section 4.1 with a detailed description of the methodology, including the experimental setup, and the three task scenarios under which participants designed incentive schemes. Then, section 4.2 introduces the metrics and analytical approach used to assess the outcomes, explaining how qualitative analysis and structured indicators are used to translate data into answers. Finally, the characteristics of the participant sample are presented in section 4.3.

4.1. Methodology

To answer the research questions proposed in this study, a series of user experiments was conducted. These experiments were designed to observe how individuals approach the task of incentive design under different conditions of guidance, and how these differences manifest in the resulting payment structures.

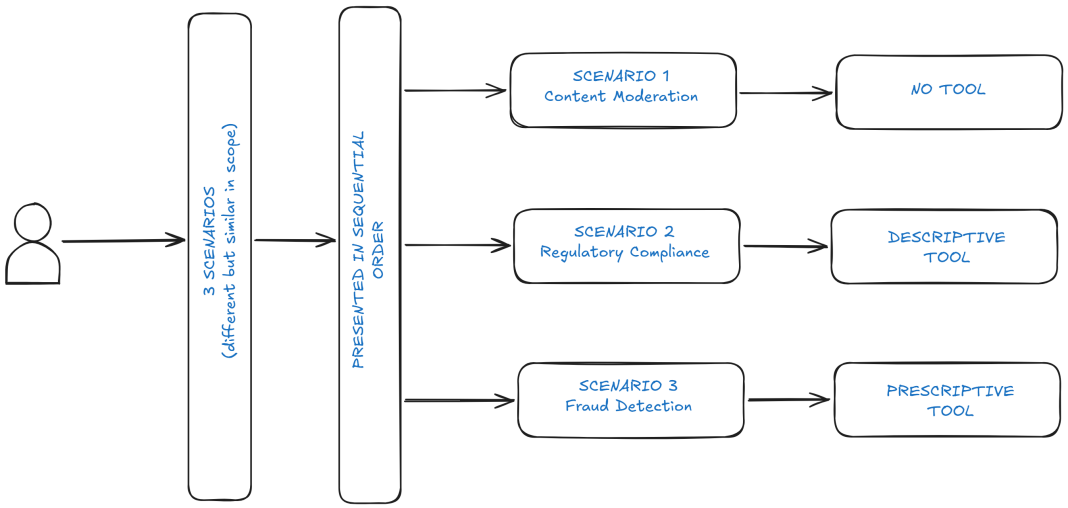

Figure 4.1: The flow of one participant study session

As presented in Figure 4.1, participants were asked to design incentive schemes for crowdsourcing jobs based on 3 provided scenarios. Each participant was exposed to all three guidance conditions (no tool, low-guidance tool, and directed-guidance tool) presented separately. To minimize learning effects and limit the likelihood that participants would carry over solutions or strategies from one condition to another, the task scenarios were intentionally varied across conditions while remaining equivalent in overall scope and complexity. However, given that the three conditions were conceptually related and incrementally built upon each other, they were presented in the same sequence: beginning with the condition without any tool support, followed by the descriptive tool condition, and concluding with the prescriptive tool condition.

Scenario 1 Recent online safety regulations have increased the requirements for moderating user-generated content. In response, an online forum platform plans to supplement its manual moderation process with an automated review step. A crowdsourcing job is set up for creating a dataset for this system. The goal is to create a dataset that reflects most user behaviors in which user-generated content, including both text and images, is reviewed against predefined platform policy rules. Workers are asked to determine whether each content item violates the policy and, when applicable, to select the appropriate violation category.

Scenario 2 Training data is prepared for a system used by an online marketplace to automatically review product documentation submitted by sellers across a wide range of product categories. The objective is to ensure that only products that meet regulatory requirements are added to seller profiles. The documentation includes materials such as certificates, declarations of conformity, and technical summaries required for selling products within the European Union. Workers are asked to examine short excerpts from these documents, together with basic metadata such as product category and submission date, to assess whether the submission complies with a predefined set of regulatory criteria. When non-compliance is identified, workers select the most appropriate reason from a set of predefined categories, including missing certificates, expired documentation, or inconsistencies between the documentation and the product information.

Scenario 3 Training data is prepared for a system intended to support the detection of potentially suspicious financial transactions for use by financial institutions in the Netherlands. The objective is to identify as many fraudulent or anomalous transactions as possible while minimizing the number of suspicious cases that go unflagged for manual review. Workers are presented with short transaction descriptions and relevant supporting data and are asked to assess whether a transaction appears unusual or potentially problematic. When a transaction is flagged, workers select the most appropriate explanation from a predefined set of categories, such as inconsistent merchant information, unexpected transaction patterns, or irregular transaction amounts.

RQ1 focuses on determining whether different levels of support during the incentive design process result in more complex payment structures. This research question calls for a within-subject analysis, in which each participant's incentive schemes are compared across the three guidance conditions to assess changes in structure, complexity, and use of incentive components.

RQ2 requires a between-subject analysis, as it aims to evaluate how consistent incentive designs are across different individuals when the same level of guidance is provided. By comparing multiple incentive schemes produced under identical guidance conditions, this analysis examines whether increased support leads to greater convergence in design choices and reduces variability between participants faced with similar scenarios.

Additionally, to gain a broader understanding of how the tool is perceived by participants, feedback was requested after each round regarding their experience (see Appendix C). Participants were also asked to indicate whether they would consider using the tool in a real-world setting.

4.2. Metrics

Given the research questions, the study primarily analyzes the tool's outcomes using a qualitative approach. At the same time, because one of the scenarios allows participants to freely construct their responses, it is inherently difficult to standardize those outputs or align them with a clearly defined and standard set of evaluation criteria.

To answer the first research question, the analysis focuses on the structural breadth of the incentive tree and the depth of the decision chain that leads to the final reward calculation. Simpler comparative indicators are also considered, such as the number of "complex" elements included in the schemes under each condition (especially what elements are omitted in the *no-tool* condition), the time required to complete each scenario, and, where applicable, the length of open-ended responses provided by participants.

To address the second research question, the metric examines whether incentive schemes created with tool support exhibit reduced variability and greater similarity across participants compared to those designed without guidance. In other words, for each condition, the resulting incentive formulas are aggregated and compared to assess structural similarity. However, the first scenario allows free-handed design without any guidance; therefore, the specificity and structure of participants' responses can vary considerably. To ensure

comparable data across conditions, the analysis focuses exclusively on the two tool-assisted scenarios, where standardized questions shape participant responses.

4.3. Participants

The study recruited 10 participants with different levels of experience in crowdsourcing and incentive design, who were directly contacted through the university mailing system. Sessions took place in person and online, with no incentives or rewards offered for participation.

The participant group consisted of computer science students with no prior experience in deploying crowdsourcing tasks, as well as individuals who had designed, managed, or published research involving crowdsourcing. Figures 4.2 and 4.3 provide an overview of the participant background survey results collected at the beginning of the session. 4 out of 10 participants had prior experience in designing crowdsourcing tasks, while the majority were classified as beginners, having never used specialized crowdsourcing platforms. When asked about their views on designing monetary incentives for crowdsourcing tasks, several experienced participants indicated that comparatively less effort is devoted to incentive design than to task design. This was reflected either in minimal time spent on incentive formulation (2 out of 10 participants) or the usage of quick estimates and defaults for determining participant compensation (2 out of 10 participants).

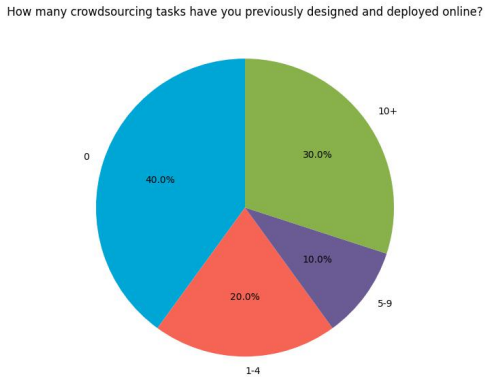

Figure 4.2: Distribution of experience with deploying online crowdsourcing jobs of the participants to the user study

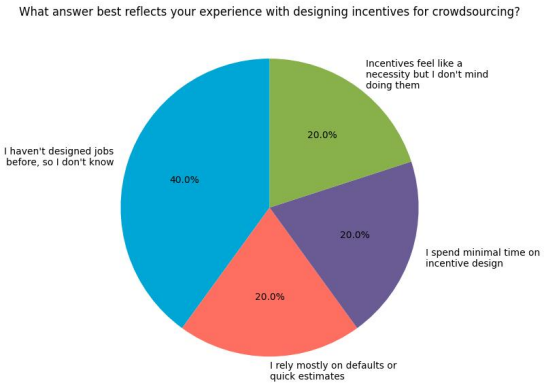

Figure 4.3: Distribution of opinions of participants about designing incentives for online crowdsourcing jobs

Given the qualitative nature of the study and the small sample size (n=10), formal statistical testing was not conducted. Instead, descriptive statistics including frequencies, percentages, and proportions were reported to characterize participant responses across conditions. This approach aligns with the study's focus on understanding patterns in user behavior and perceptions rather than testing statistical hypotheses.

5 Results

This chapter synthesizes findings from user study sessions, organizing the aggregated data into two complementary dimensions: the measurable outcomes users achieved with the tool (section 5.1) and their subjective perceptions of the tool's design and functionality (section 5.2). Together, these perspectives provide both quantitative and qualitative evidence of the tool's effectiveness and user experience.

5.1. What results were achieved?

Regarding the integration of budget constraints into incentive design, Figure 5.1 shows that 60% of participants (n=6) explicitly mentioned budget considerations when structuring their incentives within the first scenario, compared to 40% (n=4) who did not reference budget limitations.

Budget planning approaches showed identical distributions across both tool conditions, as visible in (Figure 5.2). 50% of the participants (n=5) opted for rough budget estimates in both the descriptive and prescriptive tool conditions. Fixed budget planning was selected by 20% (n=2) of participants in both conditions, while flexible budgeting was chosen by 30% (n=3) in each condition.

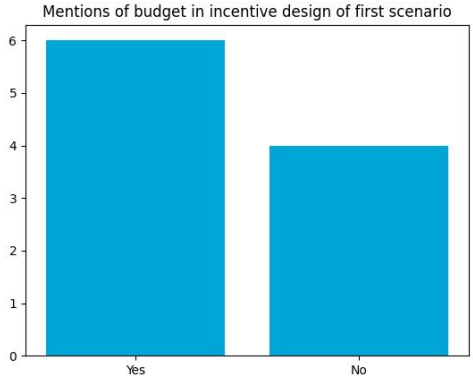

Figure 5.1: Number of participants who explicitly mentioned budget constraints in the first scenario

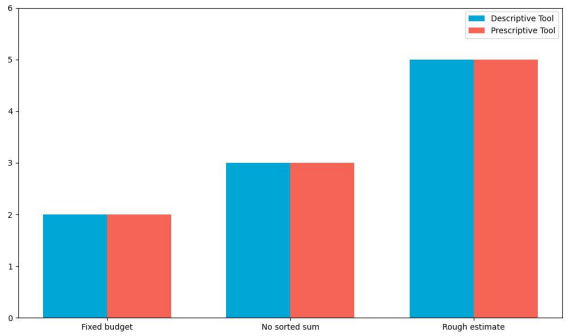

Figure 5.2: Budget planning approaches employed in the second and third scenarios

As illustrated in Figure 5.3, bonuses showed high adoption rates across scenarios. 60% of participants (n=6) used bonuses conditionally, specifically when the tool prompted them, whereas 40% (n=4) applied bonuses across all scenarios without prompting.

Penalty usage varied considerably across scenarios (see Figure 5.4). 40% of participants (n=4) used penalties only when explicitly available through the tool, representing the largest group. In contrast, 20% (n=2) applied penalties consistently across all scenarios, independent of tool availability. 30% of participants (n=3) abstained from penalties entirely, while 10% (n=1) used them exclusively in the first scenario.

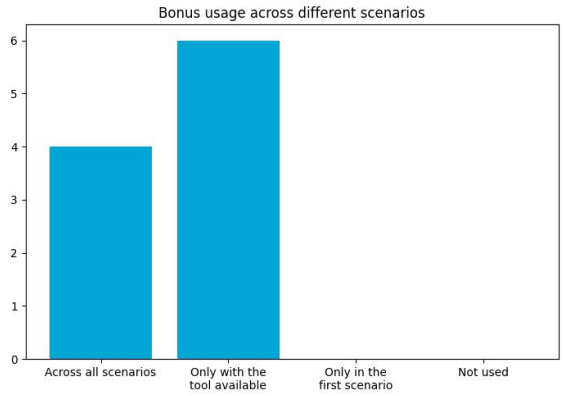

Figure 5.3: Usage of bonuses in incentive design across all three scenarios

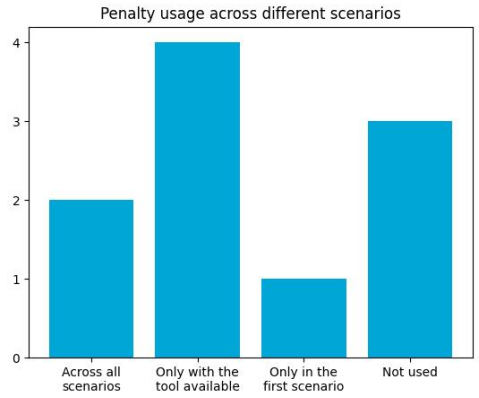

Figure 5.4: Usage of penalties in incentive design across all three scenarios

The logic behind awarding bonuses does not remain consistent between the two tool types (see Figure 5.5). In the descriptive tool, people strongly favored performance-based bonuses (70%, n=7), with the remaining participants selecting completion-based bonuses (30%, n=3). In contrast, the prescriptive tool produced more varied outcomes: 50% (n=5) chose performance-based bonuses, 20% (n=2) selected completion-based bonuses, and 30% (n=3) opted against bonuses.

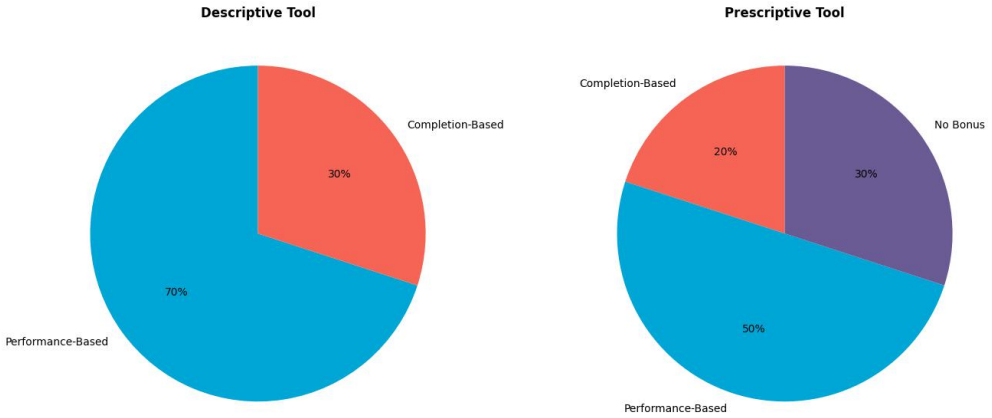

Figure 5.5: Choice of bonus award criteria in the second and third scenarios

Examining the incentive formulas from the free-handed design (scenario 1), 60% of participants (n=6) supplemented the financial element with additional incentivization strategies, while 40% (n=4) relied exclusively on a monetary approach, as presented in Figure 5.6.

Figure 5.1 shows how bonuses and penalties were used in both the prescriptive and descriptive tools, along with their specific values. Participants who employed both bonuses and penalties (n=11) set penalties at an average of 91% the bonus amount, suggesting that penalties and rewards were generally balanced. However, this average masks substantial variation in strategy: some participants used minimal penalties (as low as 0.1% of the bonus), while others employed penalties far exceeding the bonus value (up to 4x). The median ratio of 0.20 indicates that half of participants preferred conservative penalties, notably smaller than their bonuses. Experience level showed some differences: experienced users who used penalties (n=4) tended toward more balanced ratios (averaging 0.59), while beginners (n=7) showed greater variability, ranging from near-zero to 4x the bonus amount.

Across the three scenarios, participants' base payment structure choices varied by condition, as visible in Figure 5.7. In the no-tool baseline, 80% of participants (n=8) selected pay-rate models, while 20% (n=2)

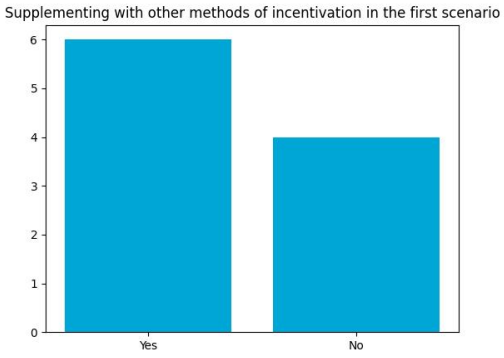

Figure 5.6: Frequency of external incentivization strategies applied by participants in the first scenario

Experience Level	Bonus	Penalty	Ratio
Experienced	3	—	—
Beginner	50	0.05	0.001
Experienced	1	—	—
Experienced	20	20.00	1.000
Beginner	10	2.00	0.200
Beginner	—	—	—
Beginner	—	—	—
Experienced	1	0.20	0.200
Beginner	2	6.00	3.000
Beginner	—	—	—
Experienced	100	100.00	1.000
Beginner	50	0.05	0.001
Experienced	1	—	—
Beginner	10	—	—
Beginner	2	8.00	4.000
Beginner	4	—	—
Beginner	20	5.00	0.250
Beginner	50	10.00	0.200
Experienced	3	—	—
Experienced	2	0.30	0.150
Average Penalty-to-Bonus Ratio (n=11): 0.91			
Total Penalty / Total Bonus: 0.461			

Table 5.1: Bonus and penalty values with penalty-to-bonus ratios. *Beginner* are users with 0-4 previously published studies; *Experienced* are users with 5+ previously published studies on online platforms for crowdsourcing

chose fixed-sum payments. When using the descriptive tool, preferences shifted to an equal split of selecting between fixed sum and pay-rate approaches (n=5,5). With the prescriptive tool, participants again favored fixed sum payments, with 60% of participants (n=6) choosing this option compared to 40% (n=4) who selected pay-rate models.

At the participant level, the trend mirrors the aggregate pattern (see Figure 5.2). 50% of participants (n=5) changed their payment structure between the no-tool baseline and descriptive tool. 80% of these changes (n=4; IDs=3, 4, 5, 8) were shifts from pay-rate to fixed sum, while one participant (20%, n=1; ID=6) switched from fixed sum to pay-rate. The remaining 50% (n=5; IDs=1, 2, 7, 10, 11) maintained their baseline choice when using the descriptive tool.

Between the descriptive and prescriptive tool conditions, choice stability increased markedly. Only 10% of participants (n=1, ID=10) changed their payment structure, shifting from pay-rate to fixed sum. 90% of participants (n=9, IDs=1, 2, 3, 4, 5, 6, 7, 8, 11) maintained identical selections across both tool-assisted conditions. When examining consistency across all three conditions, 40% (n=4, IDs=1, 2, 7, 11) selected the same payment structure in every condition. Conversely, 60% of participants (n=6, IDs=3, 4, 5, 6, 8, 10) modified their choice at least once, with the majority of changes occurring between the baseline and descriptive tool conditions rather than between the two tool variants.

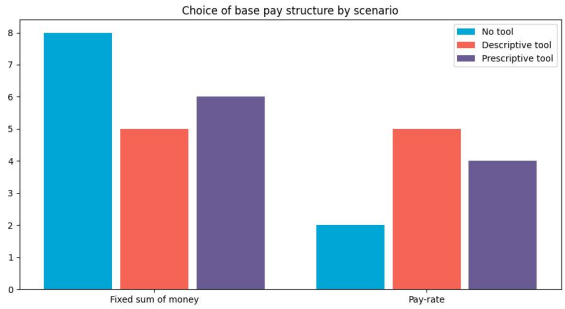

Figure 5.7: Type of base pay structure used across the three scenarios

#	No Tool	Descriptive	Prescriptive
1	Pay-rate	Pay-rate	Pay-rate
2	Pay-rate	Pay-rate	Pay-rate
3	Pay-rate	Fixed sum	Fixed sum
4	Pay-rate	Fixed sum	Fixed sum
5	Pay-rate	Fixed sum	Fixed sum
6	Fixed sum	Pay-rate	Pay-rate
7	Fixed sum	Fixed sum	Fixed sum
8	Pay-rate	Fixed sum	Fixed sum
10	Pay-rate	Pay-rate	Fixed sum
11	Pay-rate	Pay-rate	Pay-rate

Table 5.2: Base payment structure choices by tool

On a positive note, every participant's free-handed design incorporated a clear rationale explaining how their payment structure would enable workers to earn rewards.

5.2. User perception of the tool

While the designed incentive structures provide valuable insights, understanding participants' subjective responses to the tool is also relevant. This includes their overall satisfaction, perceptions of usability, and how the tool facilitated their design process.

Completion time varied significantly across tool conditions (Figure 5.8). The no-tool baseline required an average of 19 minutes. The descriptive tool reduced completion time to 7.5 minutes, representing a 60.5% decrease. The prescriptive tool achieved the lowest completion time at approximately 6.5 minutes, 65.8% shorter compared to the baseline, saving 12.5 minutes. Both tools substantially accelerated task completion, with the prescriptive tool showing a marginal 13% advantage over the descriptive tool (1 minute faster).

When comparing perceived ease of use between the two tool variants (Figure 5.9), the majority of participants (80%, n=8) reported finding both types of tools equally easy to use. One participant found the descriptive tool easier than the prescriptive tool, while one participant found the prescriptive tool harder. Overall, both tool designs achieved comparable usability ratings, with no significant perceived difference in ease of use.

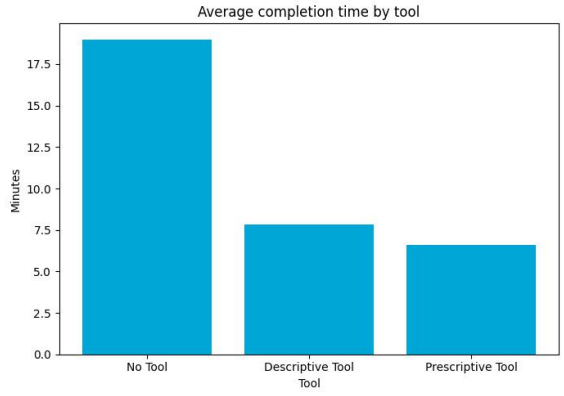

Figure 5.8: Average task completion time in minutes across the three experimental conditions: no tool, descriptive tool, and prescriptive tool.

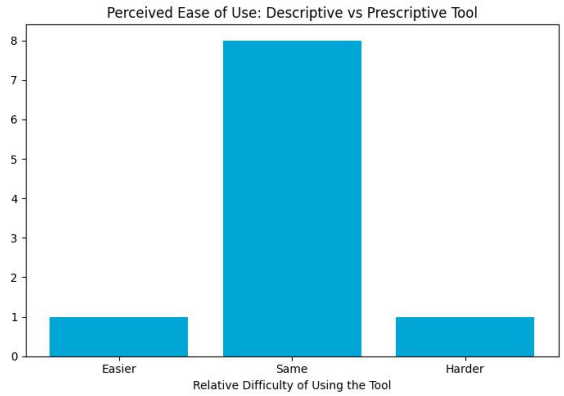

Figure 5.9: Participants' perceived ease of use when comparing the descriptive and prescriptive tools

Figure 5.10 shows that participants' perception of the tool's information usefulness differed between conditions. The prescriptive tool received higher ratings, with 90% (n=9) finding the presented information helpful, compared to 80% (n=8) in the descriptive tool condition. Conversely, 20% of participants (n=2) found the descriptive tool's information unhelpful, while only 10% (n=1) reported the prescriptive tool's information as unhelpful.

However, participants' willingness to use the tools in the future showed distinct patterns between conditions (see Figure 5.11). For the descriptive tool, responses were distributed across all options: 40% (n=4) indicated they would use it again, 40% (n=4) were uncertain about it, and 20% of participants (n=2) would not. The prescriptive tool showed stronger adoption intent, with 40% (n=4) willing to use it, 50% (n=5) uncertain, and only 10% explicitly declining. The prescriptive tool demonstrated notably lower rejection rates, with 90% of participants at least open to future use compared to 80% for the descriptive tool.

Figure 5.10 shows participants' perception of the tool's information usefulness across expertise levels and tool types. For the descriptive tool, beginners found the information more helpful (83%, n=5) compared to experienced users (75%, n=3). The prescriptive tool received higher ratings overall, with beginners unanimously finding the information useful (100%, n=6) and experienced users also rating it positively (75%, n=3). Across both tool types, experienced users showed slightly lower perceived usefulness (25% found it unhelpful, n=1) compared to beginners.

Figure 5.11 shows participants' willingness to adopt the tool in their future work, comparing expertise levels

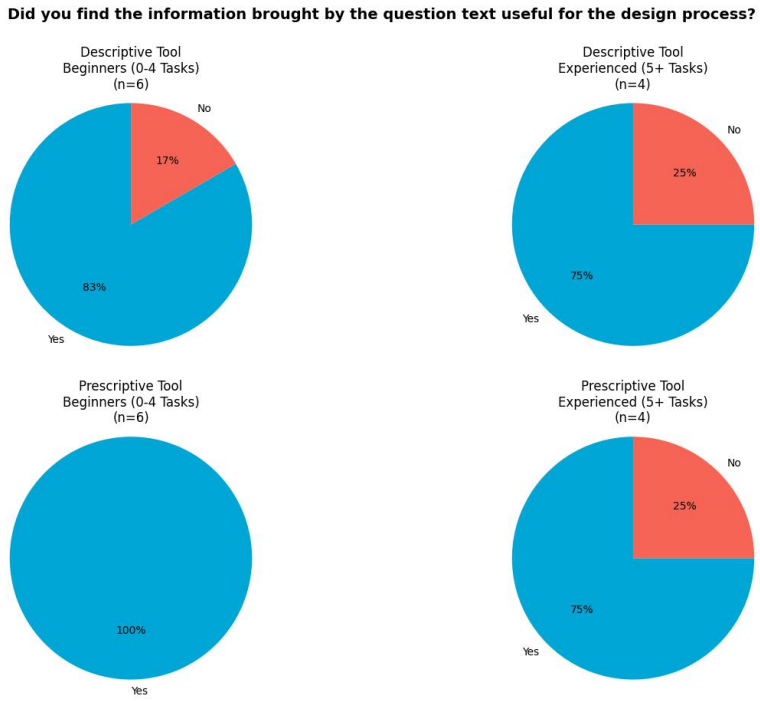

Figure 5.10: Participants' ratings of the usefulness of the information provided by the descriptive and prescriptive tools based on previous experience with crowdsourcing

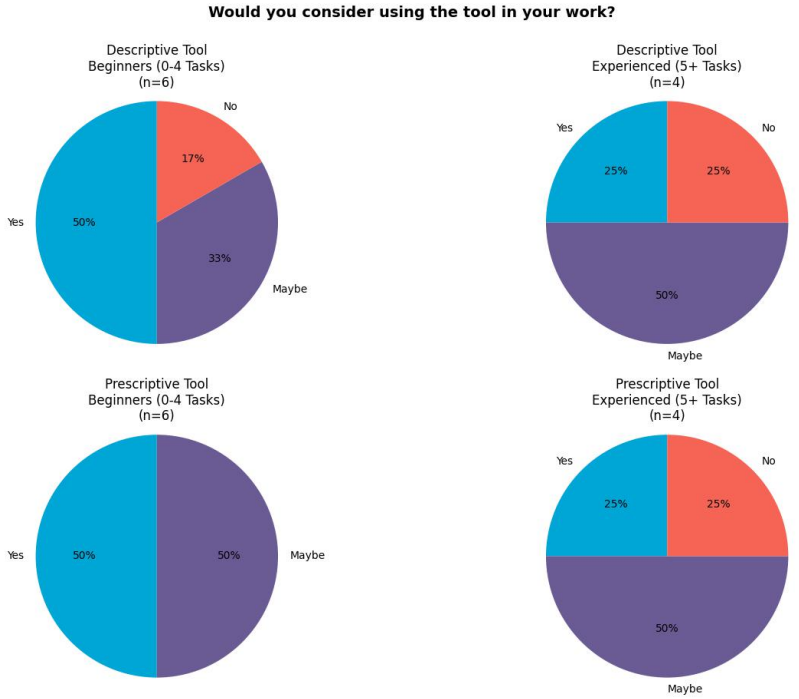

Figure 5.11: Participants' willingness to use the descriptive and prescriptive tools again in the future, based on previous experience with crowdsourcing

and tool types. For the descriptive tool, beginners showed stronger adoption intent than experienced users. Beginners were more willing to use the tool (50%, n=3), while experienced users were less willing (25%, n=1). For the prescriptive tool, beginners again showed stronger intent, with 50% (n=3) willing to use it and none declining. Experienced users of the prescriptive tool showed lower willingness (25%, n=1), similar to the descriptive tool condition.

6

Discussion

This chapter interprets the previous findings presented in chapter 5. In section 6.1, the data reveals how tools influence design choices and how people interact with them. In section 6.2, the study's limitations are addressed, contextualizing our findings and reminding us of potential interferences with the conclusion.

6.1. Data interpretation

6.1.1. The effect on incentive structure

Incentive structures are often underspecified in crowdsourcing literature and datasets [29], possibly suggesting that limited thought is invested in their design. The free-handed first scenario was created to frame and test this assumption. When comparing participants' free-style designs to the structured questions provided by the tool, a pattern emerges: some participants, particularly beginners, had never considered, or at least never articulated, certain design elements until explicitly prompted. For example, not all participants addressed budget constraints, and several failed to articulate their bonus award logic.

This pattern of omission can be explained through the lens of survey cognition theory. As Tourangeau et al. [81] describe, open-ended questions impose substantial cognitive demands: respondents must first retrieve relevant information from memory, then formulate and articulate a coherent answer. This two-phase process—retrieval followed by formulation—exhausts mental resources before the actual response is even produced. Consequently, even when participants possessed familiarity with various incentive mechanisms or had previously applied them, the act of designing under these demanding conditions meant that such knowledge remained inaccessible during the moment of decision-making.

In contrast, the structured scenarios in the tool employed predominantly closed-ended questions. Here, the cognitive task shifts from recall to recognition: participants needed only to identify and match their conceptual understanding against provided option labels. Since recognition relies on external cues that activate previously known information, this process is both faster and more accurate than unaided recall [82, 20]. This fundamental difference in cognitive effort helps explain the observable shifts in design choices between the free-style and structured scenarios.

Notable changes in design complexity emerged when participants transitioned from the free-style to the tool-assisted scenarios. Despite theoretically comparable stakes and requirements, participants frequently simplified their incentive structures: some shifted from more intricate base-pay models (i.e., pay-rate calculations towards a sum) to fixed lump-sum payments. This inverse trend, where structured guidance led to less complex designs, does not reflect the expectations of optimizing, but rather *satisficing* behavior — the tendency to select the first acceptable solution rather than continuing to search for the optimal one [80].

The open-ended design task required participants to invest substantial cognitive effort in constructing coherent incentive schemes. The tool's structure that was previously praised for helping may have unintentionally reduced how thoroughly participants thought through their designs. By offering predefined questions and a fixed set of answer choices, the tool provided clear stopping points that made it feel like each step was complete. This may have discouraged participants from exploring options outside of what the tool presented.

Instead of continuing to develop and improve their designs, participants may have simply picked an acceptable answer from the available options and moved on to the next question [54].

This pattern may also reflect a limitation inherent to the study design. Participants completed incentive designs in a hypothetical context rather than for actual deployment, which may have shifted their priorities toward task completion rather than design optimization. Without real-world consequences, the distinction between adequate and optimal designs becomes less salient, making satisficing a rational response to the experimental situation.

The free-style scenario consistently took longer to complete than the two tool-assisted scenarios, revealing a fundamental difference in cognitive demand. However, this temporal commitment did not always result in more elaborate designs. Participants' baseline-pay choices were significantly more consistent between the tool-assisted and repeated designs than between the free-style and tool-assisted scenarios. This inconsistency may not necessarily arise from changing preferences, but from the varying cognitive accessibility of design options across different elicitation methods. A design element could effectively be forgotten during the design process when it has not been the subject of prolonged prior deliberation because the participant's working memory cannot easily "hold" or retrieve that information on demand [83].

In this way, by lowering the cognitive barriers to identifying available options, the tool can help users become aware of possibilities. However, awareness by itself does not ensure adoption. The ultimate choice of whether to integrate these cues into their design, or to merely acknowledge them and continue with their established process, remains with the designer. The tool succeeds in making the design space visible, but it cannot compel designers to venture into unfamiliar territory. For experienced designers with already-formed mental models, this distinction matters: the tool may highlight options they already knew about but had not recalled, yet they retain full agency in deciding whether such options genuinely improve their specific design.

However, in its current form, the tool constrained some participants' creative decision-making by focusing exclusively on financial incentives. When switching to the tool, a user might feel pressured to "compensate" by allocating more money because there were no other ways to encourage high-quality contributions. In practice, this is unrealistic: from a deployer's perspective, the objective is to achieve high-quality outcomes without disproportionately increasing costs. This limitation is reflected in Figure 5.6, where more than 50% of participants supplemented monetary rewards with other forms of motivation, such as non-monetary extrinsic rewards (e.g., digital goods) or intrinsic incentives.

The central question is how structured prompts shape incentive designs differently than traditional free-style design. The data reveal that the tools' aid produced more detailed and complete designs than designing freely from memory alone, though the improvements depended on how much experience participants had with incentive design.

For the majority of participants with minimal prior experience, the tool functioned less as a complexity enhancer and more as a discovery mechanism. The wizard helped them recognize what options and design elements were available to them—essentially serving as an interactive tutorial. For these novice designers, the guided structure did not push them toward greater complexity per se, but rather made visible a design space they had previously been unaware of or unable to articulate. In this sense, the tool succeeded in surfacing considerations (such as bonus structures, payment models, budgets, and communication with participants) that would have otherwise remained unknown or inaccessible due to the cognitive barriers of the open-ended scenario.

For participants with existing experience in incentive design, the effect was more ambiguous. While some leveraged the tool's structure to refine their thinking, others found it constraining, particularly when it limited them to financial mechanisms alone. In these cases, the tool did not necessarily increase design complexity through this awareness; rather, it introduced a process-level friction. Experienced designers had to construct a mapping between their conceptual models and the tool's required input categories, translating their ideas into the wizard's predetermined structure to proceed. This translation burden sometimes obscured rather than clarified their underlying design reasoning.

Despite this friction, the research revealed an encouraging finding: the baseline already demonstrated a higher level of complexity than might be expected considering how undocumented this process is. The way most participants reasoned about payments reflected genuine logic and justification, not arbitrary choices. When working in the free-style format, participants explained their payment choices with some degree of

detail. Rather than just picking a number or vaguely mentioning a topic-related element, they described why they chose their base pay, bonuses, and payment structure, connecting these decisions to what they thought would motivate workers and fit the task requirements. This pattern aligns with findings presented by Desai [77], which similarly observed that people naturally gravitate toward substantive responses when given open-ended questions (in this case vaguely asking "*How should incentive design look like?*"). Instead of minimal or superficial answers, participants constructed reasoned responses that reflected their own standards of adequacy.

6.1.2. The effect on design uniformity

Despite the tool's structured approach, participants' design decisions diverged primarily at the level of implementation details, rather than in high-level design choices. At a structural level, there was relatively consistent adoption of key components: bonuses were widely incorporated regardless of condition, and the ratio of choices between the two types of base-pay were almost equally split. This suggests that designers converged on the necessity of multi-component compensation structures, recognizing that base pay alone is insufficient for effective task incentivization. In contrast, penalty usage was notably less consistent, with some participants including penalties while others omitted them entirely (see Figure 5.4).

The divergence became more pronounced when examining how these components were configured. Although bonuses were commonly used, the specific amounts varied substantially between participants, indicating differing judgments about task difficulty and appropriate worker compensation. Similar variation appeared in other parameter choices, such as whether to rely on pilot studies as benchmarks and how to set compensation targets. When penalties were included, their magnitude also differed considerably, further highlighting the lack of consensus at the parameter level.

These differences in implementation reflect deeper design philosophies regarding worker motivation and fairness. Some participants adopted a reward-focused strategy, believing that bonuses alone were sufficient to elicit desired performance. Others prioritized quality assurance and compliance, incorporating penalties to discourage poor work or non-compliance, sometimes with disproportionate ratios between bonuses and penalties (see Figure 5.1).

These differences in implementation reflect deeper design philosophies regarding worker motivation. Some participants adopted a more lenient strategy, where only bonuses on performance were enough for obtaining the desired results. However, some thought of ensuring that there is less cheating or ensure more full diligency in the participant's work, sometimes using disprortionate scales between bonuses and penalties, as displayed in Figure 5.1. However, very restrictive designs are visible even when the tool did not allow for the introduction of penalties without accompanying bonuses.

Interestingly, despite containing questions with limited answer options, the wizard failed to produce convergence. Rather than pushing participants toward a consensus [54], the constrained choices sometimes resulted in nearly equal splits across available options. This pattern reflects a deeper reality: because the tool still required open-ended input in several steps by asking participants to specify amounts, thresholds, and task-specific details from scratch, perfect uniformity was never achievable.The tool provided guided structure, but it did not eliminate the need for individual judgment. This outcome, however, is not a limitation but a strength. The persistent diversity suggests that the tool does not suppress individual thinking and minimizes the chance of satisficing considerably influencing the designs, as previously hypothesized in section 6.1.1. While defaulting is an inevitable phenomenon in any structured decision-making process, the varied outcomes indicate that designers engaged actively with choices rather than passively accepting predefined options. Designers retained agency in reflecting their own views about what constitutes fair compensation, what makes a task difficult, and what level of reward is appropriate. The tool provided a framework, but it did not override the fundamental role of human judgment and creativity in incentive design.

The persistent diversity in designs can be understood through contextual decision-making. Individual choices about compensation are shaped by personal context and values, in this case including domain knowledge, estimates of worker effort, and beliefs about fair pay. Additionally, as Dietz and Stern [22] found, these decisions are driven more by self-interest than concern for others; this might remain the only consistent pattern: all designers have an unconscious tendency to minimize costs. However, beyond that shared cost-cutting impulse, individual priorities diverge [12]. Some prioritize fairness to workers, others focus on task completion, and still others on budget constraints. When using the tool, each designer encounters the same

options but weighs them through their own priorities, leading to different results. The tool cannot override this fundamental diversity of values. Uniformity in incentive design might therefore be limited by the simple fact that different people care about different things.

Although structured formats could encourage cheating by blind selection of answers, participants' diverse outcomes indicate they exercised genuine choice. Prescriptive guidance failed to produce convergence, suggesting the tool enhanced rather than constrained reasoning. By making the design process more explicit and systematic, it enabled individual decision-making rather than suppress it.

The findings reveal that structured guidance provides some form of uniformity to design, but it operates selectively: guidance constrains high-level choices (overall incentive structures) while preserving autonomy at the operational level (specific implementation details such as pilot iteration counts or bonus amounts). This selective constraint of narrowing the macro but not the micro reflects the inherently ill-structured nature of incentive design, where no single "correct" solution exists. Designers converged on similar elements but diverged in their concrete implementation, suggesting that even systematic support cannot eliminate individual creativity and judgment in design work.

However, these are the decisions that primarily shape workers' final compensation. Yet in an oversupplied labor market where requesters have access to a constant stream of workers, they face limited incentive to design payments that reward exceptional performance or diligent effort. As Luft [59] suggests, attractive compensation structures draw more applicants, but this mechanism may fail when labor supply already exceeds demand [61, 11]. Consequently, even with guidance tools, workers' experience in the crowdsourcing marketplace may not substantially improve if requesters lack motivation to offer fair compensation. As a result, even with guidance tools, workers' experiences in the crowdsourcing marketplace may not improve significantly if requesters lack motivation to provide fair compensation. This erodes the effort towards the perception of a just and equitable labor environment in which employees can expect consistent and reasonable compensation for their efforts.

6.1.3. Perception on the tool

Beyond answering our research questions, it is valuable to assess whether this tool would be adopted in practice. The results are mixed. Participants found the information provided by the wizard useful and it substantially reduced design time, as participants completed their designs in roughly half the time when using the tool. However, this efficiency gain was not apparent to them in the moment, and more importantly, most participants remained unconvinced about actually switching to the tool for their own work.

A main barrier might have been that some incentive designs couldn't be inputted into the tool's fixed structure of base pay, bonuses, and penalties. During the sessions, participants often explained what they wanted to do and asked how it could be implemented in the system, which suggests they needed help to proceed. This indicates that the task was not immediately straightforward and required additional effort to understand how to work within the tool's constraints. As described in Fred Davis' *Technology Acceptance Model* [21], this kind of difficulty reduces perceived ease of use, which in turn lowers the likelihood that users will accept and continue using a system. In addition, when a system is harder to use, people are also less likely to see it as helpful for their work. As a result, having to find workarounds can increase resistance to adopting the wizard in regular workflows, especially when users have little prior experience with similar systems, as was the case when tools like email were first introduced [28].

On the other side, expanding the tool to accommodate more design variations faces its own constraints. Iyengar and Lepper [45] discovered that offering more options does not necessarily improve decision-making or satisfaction, even though people initially perceive more choice as beneficial. Simply adding more answer options to accommodate diverse incentive structures could overwhelm users rather than help them [79]. Similarly, expanding the tool with additional questions about supplementary payment mechanisms could lengthen the wizard to the point where users abandon the process entirely. The design challenge is to balance expressiveness with usability without compromising either.

Regarding prototype enhancement, the choice between the two approaches is not obvious based solely on user perception. Both the descriptive and prescriptive tools had users who expressed satisfaction. However, other benchmarks point to a clearer direction. The prescriptive tool demonstrated slightly faster completion times, though this metric warrants careful interpretation. More significantly, the prescriptive structure provided substantial value for novice designers by making the design space visible and guiding them through

considerations they might otherwise have overlooked. For experienced designers, the tool's expanded and instructional structure becomes less critical; they can simply skip the detailed guidance and focus on high-level questions, as users naturally do when encountering text they do not deem important [63]. This pattern suggests that the prescriptive approach provides the stronger foundation, as it can effectively serve a broader range of user expertise levels.

6.2. Limitations

While this study provides some base insights into incentive design in online crowdsourcing, several methodological constraints limit the generalizability and ecological validity of the findings.

First, the participant pool was skewed toward early beginners, resulting in limited variability in prior experience with incentive design and crowdsourcing platforms. As a consequence, the findings primarily reflect the tool's utility for novice users and do not allow for robust conclusions regarding its effectiveness across different expertise levels. It remains unclear whether experienced users would benefit to the same extent. According to the expertise reversal effect, instructional guidance that aids beginners may become redundant or counterproductive for experts [48]. Furthermore, research on help-seeking behavior shows that individuals with higher domain competence are less likely to request or rely on external support. Thus, experienced users may perceive the tool as unnecessary or interact with it in qualitatively different ways.

Second, the study relied on a limited sample size. A small sample reduces statistical power, constrains the detection of subtle effects, and increases sensitivity to individual-level variability. It also restricts the representativeness of the participant pool, limiting the extent to which conclusions can be generalized to the broader population of crowd workers or requesters.

Third, the experimental procedure required participants to complete three scenarios from scratch. This repeated construction process may have induced cognitive fatigue. Because incentive design requires reflective reasoning and trade-off evaluation, fatigue may systematically affect subsequent replies, potentially diminishing elaboration, inventiveness, or consistency in the final scenario when compared to previous ones.

Related to the previous point, the design introduces a potential learning effect. While the tool preserves a largely consistent structural framework across the descriptive and prescriptive scenarios, the content becomes progressively more elaborate and detailed. This incremental expansion may have enabled participants to internalize the question sequence or refine response strategies after initial exposure. In small samples, such practice effects can disproportionately influence the findings, as changes observed in later responses may reflect procedural familiarity rather than substantive shifts in motivational reasoning.

Although the three scenarios were designed to be as similar as possible, each participant has a subjective assessment of their importance. Consequently, two participants might assign substantially different monetary incentives to the same scenario based on their individual judgment of its significance. For example, some participants may have assigned larger monetary incentives to scenarios they deemed more important, while others did not view the same scenarios as equally critical.

Another important limitation concerns ecological validity. The scenarios did not involve real monetary stakes. Incentive structures were evaluated or constructed in a hypothetical setting, meaning that participants were not exposed to actual financial consequences. Existing literature suggests that hypothetical experimental settings can yield decision patterns that differ from those observed under real financial stakes [57, 35]. Accordingly, incentive structures evaluated without real budgetary consequences may not fully reflect behavior during an actual online deployment.

6.3. Future work

The results of the user study reveal several promising directions for enhancing the tool and deepening our understanding of how designers engage with incentive specification. A key finding is that participants were hesitant to fully embrace the tool, possibly due to the perceived constraint of usage. This suggests that expanding the tool's scope beyond financial mechanisms is important. Future iterations should allow for greater customization and support for pairing monetary incentives with complementary motivational elements, such as recognition systems, gamification, or social incentives, without forcing designers into a constrained set of options. By accommodating a richer variety of input types and incentive combinations, the tool could better reflect the full landscape of motivational strategies that designers actually employ.

Beyond the content design, there is significant potential in leveraging the structured data collected through the wizard to provide immediate value to users. One promising direction is to automatically generate human-readable summaries of incentive designs in formats suitable for research papers, technical reports, or dataset documentation that can be directly copied to speed up this process. Alternatively, the collected design specifications could be standardized into database formats and embedded as metadata within crowdsourcing datasets, creating a richer record of the incentive contexts under which data was collected. Both approaches would transform the tool from a design aid into an end-to-end solution that simplifies documentation and enables better reproducibility in crowdsourcing research.

Finally, this study compared two groups with very different levels of experience: beginners and experts. While these groups clearly responded differently to the tool, we only had two endpoints. Future research should study people across the full range of experience levels. This would help us find the critical points where designers stop wanting to use the tool and where it remains useful. At what experience level does someone feel confident enough in their own approach that they reject external guidance? Where exactly does the tool become more helpful versus more frustrating? By testing with many participants at each experience level, we could identify these transition points and develop better strategies for introducing the tool to different types of users. These insights would be crucial for deciding how and when to deploy the tool in practice.

7

Conclusion

As artificial intelligence continues to advance and become increasingly integrated into various domains, making crowdsourcing a critical infrastructure for data collection. The design of incentive structures in these tasks directly affects data quality and worker welfare, yet guidance on compensation design remains scarce. This work addresses that gap by investigating how structured guidance can improve incentive design decisions, contributing to fairer and more effective crowdsourcing practices.

This study established foundational insights into how structured guidance affects incentive design in crowdsourcing. By conducting a user study based on the development *The Wizard of Incentive*, a tool that guides requesters through compensation schema construction, two core questions were investigated: **(1)** whether descriptive and prescriptive guidance formats influence the complexity of designed incentive schemes, and **(2)** whether such tools reduce variability in compensation decisions across requesters. These questions address a critical gap in crowdsourcing practice: while incentive design directly influences data quality and worker welfare, few resources exist to support informed decision-making, leaving requesters to rely on trial and error.

The tool's effectiveness in improving incentive design varied based on requesters' prior experience. Beginners showed the greatest benefit, gaining structured context on compensation considerations that would otherwise require extensive independent research. However, the tool had notable limitations: it excluded non-monetary and intrinsic incentives despite their prevalence in participant designs, and experienced designers found the structured input fields restrictive when translating complex ideas. However, even without guidance, requesters did not design arbitrarily; their decisions reflected deliberate reasoning, suggesting a reasonably strong baseline for incentive design intuition.

When it comes to how designs from different people on the same task looked like, the main observation is that the tool restricted the high-level answer set to a couple of similar variations of designs, with the items such as base payment and whether to use bonuses or not, differed. The way these were used was still the varying factor of the designs. Just because there visually are options, it is still up to the designr's diligence to use this information well to create the tool. This was very much observed into how different the perspectives of implementing bonuses and penalties were for the same scenario, meaning that workers might still be treated unequally for doing the same task.

The tool would require significant refinement before requesters would willingly adopt it into their workflow. Despite recognizing the value of its guidance, participants expressed hesitation, which could be linked to the tool's inflexibility: certain design decisions could not be translated into the required input format, creating a barrier to adoption. Nevertheless, the tool's primary strength lies in its ability to structure the decision-making process and provide relevant context. With modest improvements to accommodate diverse design approaches, it could serve as an effective starting point for beginners seeking to construct more thoughtful incentive schemes, introducing much-needed structure into a currently unguided process.

The prescriptive version shows particular promise as a foundation for future development. It maintains designer autonomy while facilitating informed decision-making by providing extensive context without pre-

scribing specific options. However, a critical limitation is the exclusion of non-monetary and intrinsic incentives. Future iterations should integrate these elements, allowing designers to construct comprehensive reward structures within a single coherent workflow rather than fragmenting compensation into separate processes. Testing such improvements with larger and more diverse participant groups would reveal how these additions affect design patterns across varying experience levels and whether the tool can accommodate more creative, customized approaches without requiring extensive translation of ideas.

Precisely because crowdsourcing relies on undefined groups of workers without formal employment protections, it has become increasingly important to technological advancement at unprecedented scales. Yet fair worker compensation remains largely unaddressed, perpetuating a cycle where workers resort to shortcuts to sustain income while requesters implement increasingly restrictive controls. Though standardizing incentive design may appear incremental, it represents a critical intervention toward fostering more educated and ethical requesters. By supporting deliberate, informed decision-making around compensation, we move beyond transactional labor markets and move toward systems that recognize worker dignity and data quality as fundamentally interconnected. Scaling this approach could lay the groundwork for rapid technological development while ensuring that workers enabling this progress are fairly valued and genuinely respected.

References

[1] L. von Ahn. “Games with a purpose”. In: *Computer* 39.6 (2006), pp. 92–94. DOI: 10.1109/MC.2006.196.

[2] Luis von Ahn. “Human computation”. In: *Proceedings of the 46th Annual Design Automation Conference*. DAC ’09. San Francisco, California: Association for Computing Machinery, 2009, pp. 418–419. ISBN: 9781605584973. DOI: 10.1145/1629911.1630023. URL: https://doi.org/10.1145/1629911.1630023.

[3] Luis von Ahn and Laura Dabbish. “Designing games with a purpose”. In: *Commun. ACM* 51.8 (Aug. 2008), pp. 58–67. ISSN: 0001-0782. DOI: 10.1145/1378704.1378719. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/1378704.1378719.

[4] Amazon Web Services. *Mechanical Turk Code Samples*. https://github.com/aws-samples/mturk-code-samples. Accessed: 2025-02-18. 2025.

[5] Judd Antin and Aaron Shaw. “Social desirability bias and self-reports of motivation: a study of amazon mechanical turk in the US and India”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’12. Austin, Texas, USA: Association for Computing Machinery, 2012, pp. 2925–2934. ISBN: 9781450310154. DOI: 10.1145/2207676.2208699. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/2207676.2208699.

[6] Appen. *AI Data Quality*. https://www.appen.com/ai-data-quality. Accessed: 2025-02-18. 2025.

[7] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. MIT Press, 2019. URL: http://fairmlbook.org/.

[8] Daniel W. Barowy et al. “AutoMan: a platform for integrating human-based and digital computation”. In: *Proceedings of the ACM International Conference on Object Oriented Programming Systems Languages and Applications*. OOPSLA ’12. Tucson, Arizona, USA: Association for Computing Machinery, 2012, pp. 639–654. ISBN: 9781450315616. DOI: 10.1145/2384616.2384663. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/2384616.2384663.

[9] Benjamin B. Bederson and Alexander J. Quinn. “Web workers unite! addressing challenges of online laborers”. In: *CHI ’11 Extended Abstracts on Human Factors in Computing Systems*. CHI EA ’11. Vancouver, BC, Canada: Association for Computing Machinery, 2011, pp. 97–106. ISBN: 9781450302685. DOI: 10.1145/1979742.1979606. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/1979742.1979606.

[10] Janine Berg. *Income Security in the On-Demand Economy: Findings and Policy Lessons from a Survey of Crowdworkers*. Tech. rep. 74. Geneva: International Labour Office, Inclusive Labour Markets, Labour Relations and Working Conditions Branch, 2016.

[11] Janine Berg et al. *Digital Labour Platforms and the Future of Work: Towards Decent Work in the Online World*. International Labour Organization, 2018. URL: https://www.ilo.org/publications/digital-labour-platforms-and-future-work-towards-decent-work-online-world.

[12] Dimitris Bertsimas, Vivek F. Farias, and Nikolaos Trichakis. “On the Efficiency-Fairness Trade-off”. In: *Management Science* 58.12 (2012), pp. 2234–2250. ISSN: 0025-1909. DOI: 10.1287/mnsc.1120.1549. URL: http://dx.doi.org/10.1287/mnsc.1120.1549.

[13] Irma Borst. “Understanding Crowdsourcing: Effects of motivation and rewards on participation and performance in voluntary online activities”. PhD thesis. Erasmus University Rotterdam, 2010.

[14] Kevin J. Boudreau and Karim R. Lakhani. “Using the Crowd as an Innovation Partner”. In: *Harvard Business Review* 91.4 (Apr. 2013), pp. 60–69, 140. ISSN: 0017-8012.

[15] Alessandro Bozzon, Marco Brambilla, and Stefano Ceri. “Answering search queries with CrowdSearcher”. In: *Proceedings of the 21st International Conference on World Wide Web*. WWW ’12. Lyon, France: Association for Computing Machinery, 2012, pp. 1009–1018. ISBN: 9781450312295. DOI: 10.1145/2187836.2187971. URL: https://doi.org/10.1145/2187836.2187971.

[16] Zana Buçinca et al. “Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems”. In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI ’20. Cagliari, Italy: Association for Computing Machinery, 2020, pp. 454–464. ISBN: 9781450371186. DOI: 10.1145/3377325.3377498. URL: https://doi.org/10.1145/3377325.3377498.

[17] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. “Amazon’s Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?” In: *Perspectives on Psychological Science* 6.1 (2011). Original work published 2011, pp. 3–5. DOI: 10.1177/1745691610393980. URL: https://doi.org/10.1177/1745691610393980.

[18] Dana Chandler and Adam Kapelner. “Breaking Monotony with Meaning: Motivation in Crowdsourcing Markets”. In: *Journal of Economic Behavior & Organization* 90 (2013), pp. 123–133. DOI: 10.1016/j.jebo.2013.03.003. URL: https://www.sciencedirect.com/science/article/pii/S016726811300036X.

[19] Jon F. Claerbout and Martin Karrenbach. “Electronic Documents Give Reproducible Research a New Meaning”. In: *Proceedings of the 62nd Annual International Meeting of the Society of Exploration Geophysicists*. Society of Exploration Geophysicists, 1992, pp. 601–604.

[20] Fergus I. M. Craik and Endel Tulving. “Depth of processing and the retention of words in episodic memory”. In: *Journal of Experimental Psychology: General* 104.3 (1975), pp. 268–294.

[21] Fred D Davis. “Perceived usefulness, perceived ease of use, and user acceptance of information technology”. In: *MIS Quarterly* 13.3 (1989), pp. 319–340.

[22] Thomas Dietz and Paul C. Stern. “Toward a theory of choice: Socially embedded preference construction”. In: *The Journal of Socio-Economics* 24.2 (1995), pp. 261–279. ISSN: 1053-5357. DOI: https://doi.org/10.1016/1053-5357(95)90022-5. URL: https://www.sciencedirect.com/science/article/pii/1053535795900225.

[23] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. “Pick-a-crowd: tell me what you like, and i’ll tell you what to do”. In: *Proceedings of the 22nd International Conference on World Wide Web*. WWW ’13. Rio de Janeiro, Brazil: Association for Computing Machinery, 2013, pp. 367–374. ISBN: 9781450320351. DOI: 10.1145/2488388.2488421. URL: https://doi.org/10.1145/2488388.2488421.

[24] Dynamo Contributors. *Guidelines for Academic Requesters*. Version 1.1. 25 pages. 2014. URL: https://irb.northwestern.edu/docs/guidelinesforacademicrequesters-1.pdf.

[25] Carsten Eickhoff et al. “Quality through flow and immersion: gamifying crowdsourced relevance assessments”. In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’12. Portland, Oregon, USA: Association for Computing Machinery, 2012, pp. 871–880. ISBN: 9781450314725. DOI: 10.1145/2348283.2348400. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/2348283.2348400.

[26] Chiara Franzoni, Marion Poetz, and Henry Sauermann. “Crowds, Citizens, and Science: A Multi-Dimensional Framework and Agenda for Future Research”. In: *Industry and Innovation* 29.2 (2021), pp. 251–284. DOI: 10.1080/13662716.2021.1976627. URL: https://doi.org/10.1080/13662716.2021.1976627.

[27] Martha Garcia-Murillo and Ian MacInnes. “The Impact of AI on Employment: A Historical Account of Its Evolution”. In: *30th European Regional ITS Conference, Helsinki 2019*. 205178. International Telecommunications Society (ITS), 2019. URL: https://www.econstor.eu/bitstream/10419/205178/1/Garcia-Murillo-MacInnes.pdf.

[28] Laura Garton and Barry Wellman. “Social Impacts of Electronic Mail in Organizations: A Review of the Research Literature”. In: *Communication Yearbook* 18.1 (Jan. 1995), pp. 434–453. ISSN: 0147-4642. DOI: 10.1080/23808985.1995.11678923. eprint: https://academic.oup.com/anncom/article-pdf/18/1/434/59958245/anncom_18_1_434.pdf. URL: https://doi.org/10.1080/23808985.1995.11678923.

[29] Timnit Gebru et al. “Datasheets for datasets”. In: *Commun. ACM* 64.12 (Nov. 2021), pp. 86–92. ISSN: 0001-0782. DOI: 10.1145/3458723. URL: https://doi.org/10.1145/3458723.

[30] Mary L. Gray and Siddharth Suri. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. New York, NY: Houghton Mifflin Harcourt, 2019. ISBN: 9781328566249.

[31] Nielsen Norman Group. *Wizards: Definition and Best Practices*. Accessed: 2025-02-11. 2020. URL: https://www.nngroup.com/articles/wizards/.

[32] S. Gupta et al. “Patterns in the Growth and Thematic Evolution of Artificial Intelligence Research: A Study Using Bradford Distribution of Productivity and Path Analysis”. In: *International Journal of Intelligent Systems* 2024.1 (2023), p. 5511224. DOI: 10.1155/2024/5511224. URL: https://doi.org/10.1155/2024/5511224.

[33] Kotaro Hara et al. “A Data-Driven Analysis of Workers’ Earnings on Amazon Mechanical Turk”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI ’18. Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–14. ISBN: 9781450356206. DOI: 10.1145/3173574.3174023. URL: https://doi.org/10.1145/3173574.3174023.

[34] Christopher G. Harris. “The Effects of Pay-to-Quit Incentives on Crowdworker Task Quality”. In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW ’15. Vancouver, BC, Canada: Association for Computing Machinery, 2015, pp. 1801–1812. ISBN: 9781450329224. DOI: 10.1145/2675133.2675185. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/2675133.2675185.

[35] Glenn W. Harrison and E. Elisabet RutstrÃ¶m. *Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods*. None 2008. DOI: None. URL: https://ideas.repec.org/h/eee/expchp/5-81.html.

[36] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. “How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System”. In: *Proc. ACM Hum.-Comput. Interact.* 7.CSCW2 (Oct. 2023). DOI: 10.1145/3610067. URL: https://doi.org/10.1145/3610067.

[37] Eli Hinkel. *Descriptive Versus Prescriptive Grammar*. In: *The TESOL Encyclopedia of English Language Teaching*. Wiley, 2018, pp. 1–6. DOI: 10.1002/9781118784235.eelt0053.

[38] Chien-Ju Ho, Shahin Jabbari, and Jennifer Wortman Vaughan. “Adaptive task assignment for crowdsourced classification”. In: *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. ICML’13. Atlanta, GA, USA: JMLR.org, 2013, pp. I–534–I–542.

[39] Chien-Ju Ho et al. “Incentivizing High Quality Crowdwork”. In: *Proceedings of the 24th International Conference on World Wide Web*. WWW ’15. Florence, Italy: International World Wide Web Conferences Steering Committee, 2015, pp. 419–429. ISBN: 9781450334693. DOI: 10.1145/2736277.2741102. URL: https://doi.org/10.1145/2736277.2741102.

[40] John J. Horton and L. R. Chilton. “The Labor Economics of Paid Crowdsourcing”. In: *Proceedings of the 11th ACM Conference on Electronic Commerce (EC ’10)*. New York, NY, USA: ACM, 2010, pp. 209–218. DOI: 10.1145/1807342.1807376.

[41] Jeff Howe. “The Rise of Crowdsourcing”. In: *Wired Magazine* 14.6 (2006). Accessed: 2025-11-13. URL: https://www.wired.com/2006/06/crowds/.

[42] Kazushi Ikeda and Michael S. Bernstein. “Pay It Backward: Per-Task Payments on Crowdsourcing Platforms Reduce Productivity”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI ’16. San Jose, California, USA: Association for Computing Machinery, 2016, pp. 4111–4121. ISBN: 9781450333627. DOI: 10.1145/2858036.2858327. URL: https://doi.org/10.1145/2858036.2858327.

[43] Panagiotis G. Ipeirotis. “Analyzing the Amazon Mechanical Turk Marketplace”. In: *XRDS: Crossroads, The ACM Magazine for Students* 17.2 (2010), pp. 16–21. DOI: 10.1145/1869086.1869094. URL: https://doi.org/10.1145/1869086.1869094.

[44] Panagiotis G. Ipeirotis. “Analyzing the Amazon Mechanical Turk marketplace”. In: *XRDS* 17.2 (Dec. 2010), pp. 16–21. ISSN: 1528-4972. DOI: 10.1145/1869086.1869094. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/1869086.1869094.

[45] Sheena S. Iyengar and Mark R. Lepper. “When Choice is Demotivating: Can One Desire Too Much of a Good Thing?” In: *Journal of Personality and Social Psychology* 79.6 (2000), pp. 995–1006. DOI: 10.1037/0022-3514.79.6.995.

[46] David H. Jonassen. *Learning to Solve Problems: A Handbook for Designing Problem-Solving Learning Environments*. 1st. New York, NY, USA: Routledge, 2010, p. 472. ISBN: 9780203847527. DOI: 10.4324/9780203847527. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pfi.4140440909.

[47] Hernisa Kacorri, Kaoru Shinkawa, and Shin Saito. “Introducing game elements in crowdsourced video captioning by non-experts”. In: *Proceedings of the 11th Web for All Conference*. W4A ’14. Seoul, Korea: Association for Computing Machinery, 2014. ISBN: 9781450326513. DOI: 10.1145/2596695.2596713. URL: https://doi.org/10.1145/2596695.2596713.

[48] Slava Kalyuga et al. “The Expertise Reversal Effect”. In: *Educational Psychologist* 38.1 (2003), pp. 23–31. DOI: 10.1207/S15326985EP3801_4.

[49] N. Kaufmann, T. Schulze, and Daniel Veit. “More than fun and money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk”. In: *Proceedings of the 17th Americas Conference on Information Systems (AMCIS 2011), 4-8 August 2011, Detroit, Michigan, USA*. 2018, p. 340. URL: http://aisel.aisnet.org/amcis2011_submissions/340.

[50] S. Kaur. “Incentive-Tuning: Understanding and Designing Incentives for Empirical Human-AI Decision-Making Studies”. Master’s thesis. Delft, The Netherlands: Delft University of Technology, 2024. URL: https://resolver.tudelft.nl/uuid:08882824-d5ff-451f-8d07-1a2bd5ad1554.

[51] Aniket Kittur et al. “CrowdForge: crowdsourcing complex work”. In: *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. UIST ’11. Santa Barbara, California, USA: Association for Computing Machinery, 2011, pp. 43–52. ISBN: 9781450307161. DOI: 10.1145/2047196.2047202. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/2047196.2047202.

[52] Aniket Kittur et al. “CrowdWeaver: visually managing complex crowd work”. In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. CSCW ’12. Seattle, Washington, USA: Association for Computing Machinery, 2012, pp. 1033–1036. ISBN: 9781450310864. DOI: 10.1145/2145204.2145357. URL: https://doi.org/10.1145/2145204.2145357.

[53] Aniket Kittur et al. “The Future of Crowd Work”. In: *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW)*. 2013, pp. 1301–1318. DOI: 10.1145/2441776.2441923. URL: https://hci.stanford.edu/publications/2013/CrowdWork/futureofcrowdwork-cscw2013.pdf.

[54] Jon A. Krosnick. “Response strategies for coping with the cognitive demands of attitude measures in surveys”. In: *Applied Cognitive Psychology* 5.3 (1991), pp. 213–236. DOI: https://doi.org/10.1002/acp.2350050305. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.2350050305. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.2350050305.

[55] Joey J. Lee et al. “GREENIFY: A Real-World Action Game for Climate Change Education”. In: *Simulation & Gaming* 44.2-3 (2013), pp. 349–365. DOI: 10.1177/1046878112470539. eprint: https://doi.org/10.1177/1046878112470539. URL: https://doi.org/10.1177/1046878112470539.

[56] Geoffrey Leech. “Descriptive grammar”. In: *The Cambridge Handbook of English Corpus Linguistics*. Ed. by Douglas Biber and RandiEditors Reppen. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, 2015, pp. 146–160.

[57] Steven D. Levitt and John A. List. “What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?” In: *Journal of Economic Perspectives* 21.2 (Spring 2007), pp. 153–174. DOI: None. URL: https://ideas.repec.org/a/aea/jecper/v21y2007i2p153-174.html.

[58] Leib Litman, Jonathan Robinson, and Charlotte Rosenzweig. “The Relationship Between Motivation, Monetary Compensation, and Data Quality Among US- and India-Based Workers on Mechanical Turk”. In: *Behavior Research Methods* 47.2 (2015), pp. 519–528. DOI: 10.3758/s13428-014-0483-x. URL: https://doi.org/10.3758/s13428-014-0483-x.

[59] Joan Luft. “Bonus and penalty incentives contract choice by employees”. In: *Journal of Accounting and Economics* 18.2 (1994), pp. 181–206. ISSN: 0165-4101. DOI: https://doi.org/10.1016/0165-4101(94)00361-0. URL: https://www.sciencedirect.com/science/article/pii/0165410194003610.

[60] Tie Luo et al. “Incentive Mechanism Design for Crowdsourcing: An All-Pay Auction Approach”. In: *ACM Trans. Intell. Syst. Technol.* 7.3 (Feb. 2016). ISSN: 2157-6904. DOI: 10.1145/2837029. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/2837029.

[61] David Martin et al. “Being a turker”. In: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW ’14. Baltimore, Maryland, USA: Association for Computing Machinery, 2014, pp. 224–235. ISBN: 9781450325400. DOI: 10.1145/2531602.2531663. URL: https://doi.org/10.1145/2531602.2531663.

[62] Winter Mason and Duncan J. Watts. “Financial incentives and the ”performance of crowds””. In: *SIGKDD Explor. Newsl.* 11.2 (May 2010), pp. 100–108. ISSN: 1931-0145. DOI: 10.1145/1809400.1809422. URL: https://doi.org/10.1145/1809400.1809422.

[63] Meike Morren and Leonard J. Paas. “Short and Long Instructional Manipulation Checks: What Do They Measure?” In: *International Journal of Public Opinion Research* 32.4 (2020), pp. 790–800. ISSN: 1471-6909. DOI: 10.1093/ijpor/edz046. URL: https://doi.org/10.1093/ijpor/edz046.

[64] Donald A. Norman. *The Design of Everyday Things*. Revised and Expanded. New York: Basic Books, 2013.

[65] Jonas Oppenlaender, Tahir Abbas, and Ujwal Gadiraju. “The State of Pilot Study Reporting in Crowdsourcing: A Reflection on Best Practices and Guidelines”. In: *arXiv preprint* arXiv:2312.08090 (2023). URL: https://arxiv.org/abs/2312.08090.

[66] Jonas Oppenlaender, Tahir Abbas, and Ujwal Gadiraju. “The State of Pilot Study Reporting in Crowdsourcing: A Reflection on Best Practices and Guidelines”. In: *Proc. ACM Hum.-Comput. Interact.* 8.CSCW1 (Apr. 2024). DOI: 10.1145/3641023. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3641023.

[67] Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. “Running Experiments on Amazon Mechanical Turk”. In: *Judgment and Decision Making* 5.5 (June 2010), pp. 411–419. URL: https://ssrn.com/abstract=1626226.

[68] Chirag Patel et al. “Monetary rewards and self-selection in design crowdsourcing contests: Managing participation, contribution appropriateness, and winning trade-offs”. In: *Technological Forecasting and Social Change* 191 (2023), p. 122447. ISSN: 0040-1625. DOI: https://doi.org/10.1016/j.techfore.2023.122447. URL: https://www.sciencedirect.com/science/article/pii/S0040162523001324.

[69] Amandalynne Paullada et al. “Data and its (dis)contents: A survey of dataset development and use in machine learning research”. In: *Patterns* 2.11 (2021), p. 100336. ISSN: 2666-3899. DOI: https://doi.org/10.1016/j.patter.2021.100336. URL: https://www.sciencedirect.com/science/article/pii/S2666389921001847.

[70] Prolific Researcher Help Center. *How do I send bonus payments?* https://researcher-help.prolific.com/en/articles/445233-how-do-i-send-bonus-payments. Accessed: 2025-11-16. 2025.

[71] Prolific Researcher Help Center. *Prolific’s Payment Principles*. https://researcher-help.prolific.com/en/articles/445230-prolific-s-payment-principles. Accessed: 2025-11-16. 2025.

[72] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. “Improving Worker Engagement Through Conversational Microtask Crowdsourcing”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–12. ISBN: 9781450367080. DOI: 10.1145/3313831.3376403. URL: https://doi.org/10.1145/3313831.3376403.

[73] Alexander J. Quinn and Benjamin B. Bederson. “Human computation: a survey and taxonomy of a growing field”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’11. Vancouver, BC, Canada: Association for Computing Machinery, 2011, pp. 1403–1412. ISBN: 9781450302289. DOI: 10.1145/1978942.1979148. URL: https://doi.org/10.1145/1978942.1979148.

[74] M. J. Raddick et al. “Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers”. In: *Astronomy Education Review* 9.1 (2010). Retrieved November 16, 2025. URL: https://www.learntechlib.org/p/106580/.

[75] Jakob Rogstadius et al. “An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets”. In: *Proceedings of the International AAAI Conference on Web and Social Media* 5.1 (2011), pp. 321–328.

[76] John P. Rula et al. “No ”one-size fits all”: towards a principled approach for incentives in mobile crowdsourcing”. In: *Proceedings of the 15th Workshop on Mobile Computing Systems and Applications*. HotMobile ’14. Santa Barbara, California: Association for Computing Machinery, 2014. ISBN: 9781450327428. DOI: 10.1145/2565585.2565603. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/2565585.2565603.

[77] Desai Connor S. and Reimers Stian. “Comparing the use of open and closed questions for web-based measures of the continued-influence effect”. In: *Behavior Research Methods* 51 (2019), pp. 1426–1440. DOI: 10.3758/s13428-018-1066-z. URL: https://doi.org/10.3758/s13428-018-1066-z.

[78] Mizuki Sakamoto and Tatsuo Nakajima. “Gamifying Social Media to Encourage Social Activities with Digital-Physical Hybrid Role-Playing”. In: *Social Computing and Social Media*. Ed. by Gabriele Meiselwitz. Cham: Springer International Publishing, 2014, pp. 581–591. ISBN: 978-3-319-07632-4.

[79] Barry Schwartz. *The paradox of choice: Why more is less*. Ecco, 2004.

[80] Herbert A. Simon. “Rational choice and the structure of the environment”. In: *Psychological Review* 63.2 (1956), pp. 129–138.

[81] Roger Tourangeau, Lance J. Rips, and Kenneth Rasinski. *The Psychology of Survey Response*. Cambridge University Press, 2000.

[82] Endel Tulving and Zena Pearlstone. “Availability versus accessibility of information in memory”. In: *Journal of Verbal Learning and Verbal Behavior* 5.4 (1966), pp. 381–391.

[83] Amos Tversky and Daniel Kahneman. “Availability: A heuristic for judging frequency and probability”. In: *Cognitive Psychology* 5.2 (1973), pp. 207–232. DOI: 10.1016/0010-0285(73)90033-9. URL: https://doi.org/10.1016/0010-0285(73)90033-9.

[84] Amos Tversky and Daniel Kahneman. “The Framing of Decisions and the Psychology of Choice”. In: *Science* 211.4481 (1981), pp. 453–458. DOI: 10.1126/science.7455683. URL: https://psycnet.apa.org/record/1981-31998-001.

[85] Jennifer Wortman Vaughan. “Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research”. In: *Journal of Machine Learning Research* 18.193 (2018), pp. 1–46. URL: http://jmlr.org/papers/v18/17-234.html.

[86] Mark E. Whiting, Grant Hugh, and Michael S. Bernstein. “Fair Work: Crowd Work Minimum Wage with One Line of Code”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7.1 (Oct. 2019), pp. 197–206. DOI: 10.1609/hcomp.v7i1.5283. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/5283.

[87] Ming Yin, Yiling Chen, and Yu-An Sun. “Monetary Interventions in Crowdsourcing Task Switching”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 2.1 (Sept. 2014), pp. 234–241. DOI: 10.1609/hcomp.v2i1.13160. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/13160.

[88] Haichao Zheng, Dahui Li, and Wenhua Hou. “Task Design, Motivation, and Participation in Crowdsourcing Contests”. In: *Int. J. Electron. Commerce* 15.4 (July 2011), pp. 57–88. ISSN: 1086-4415. DOI: 10.2753/JEC1086-4415150402. URL: https://doi.org/10.2753/JEC1086-4415150402.

A

The Wizard of Incentive: Tool Overview

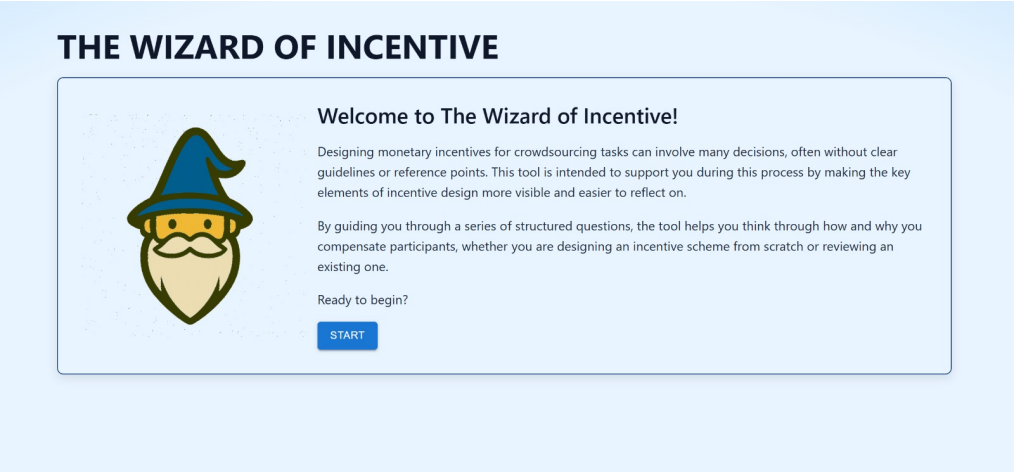

Figure A.1: The welcome screen of "The Wizard of Incentive".

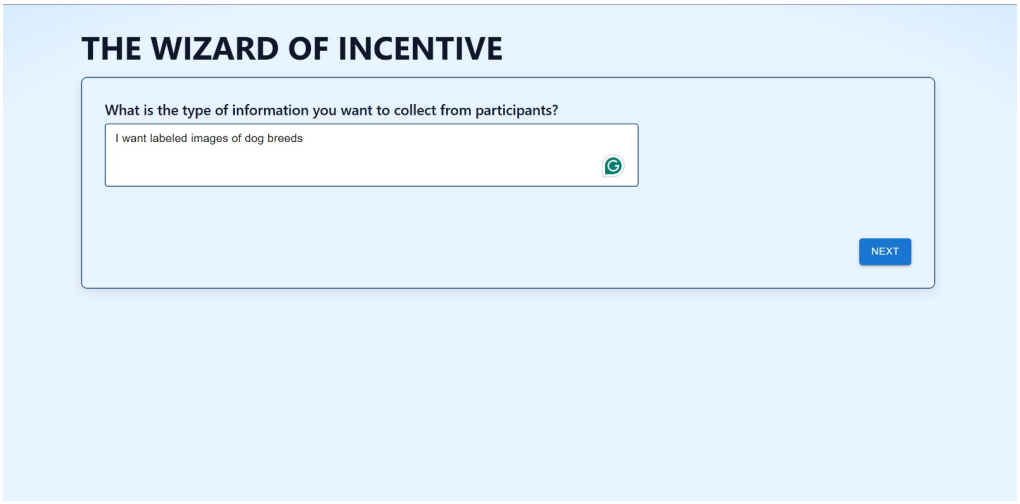

Figure A.2: An example of an open-ended question prompting the user to describe the type of information they wish to collect from participants.

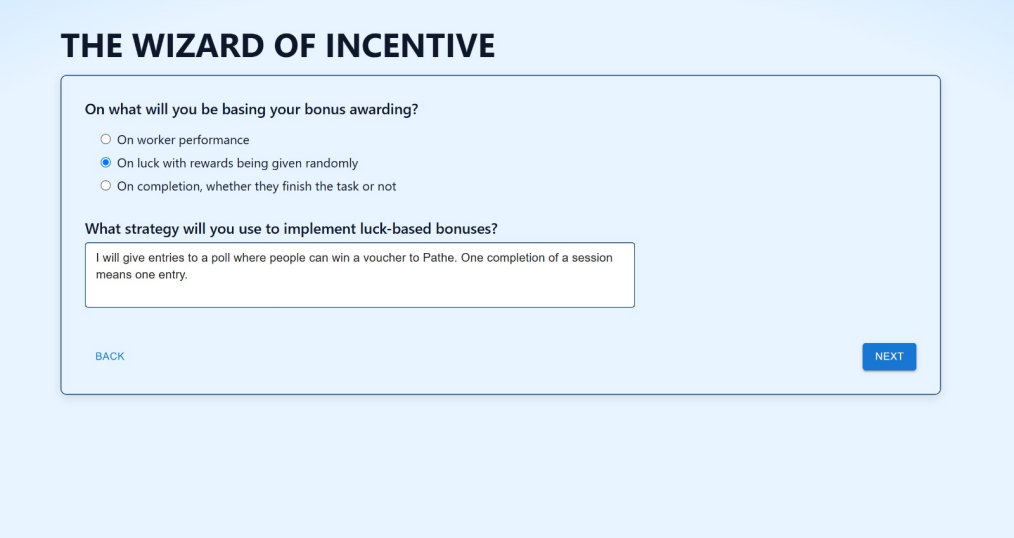

Figure A.3: An example of a closed-ended question combined with an open-ended follow-up, guiding the user through bonus awarding logic.

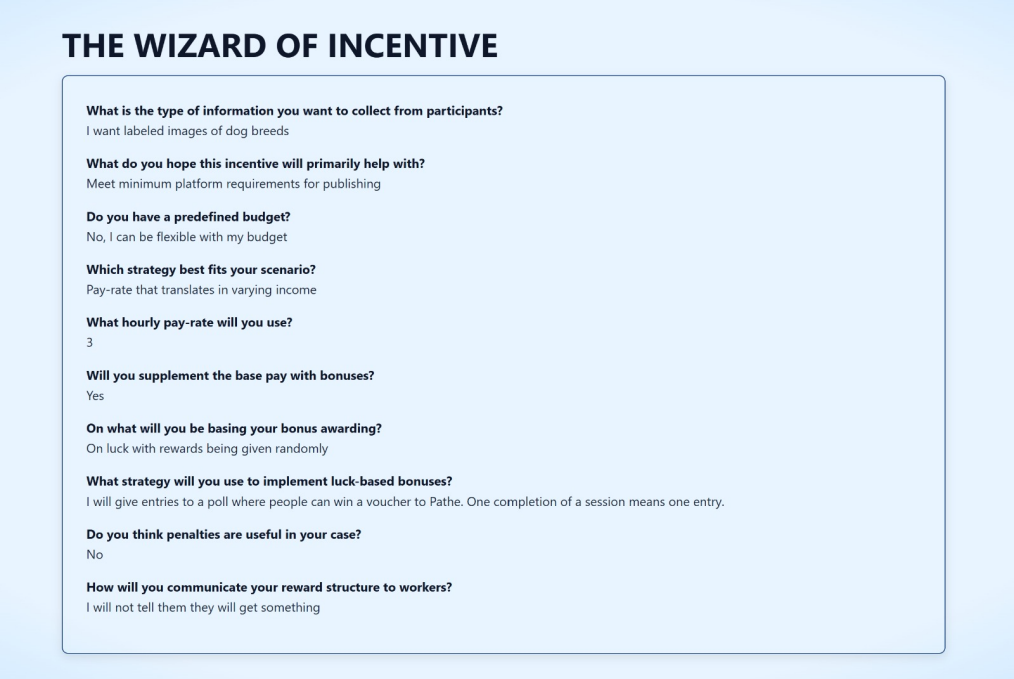

Figure A.4: The incentive formula summary screen, presenting a complete overview of all design decisions made throughout the wizard in a structured, readable format.

B

Appendix A: The Wizard of Incentive – Introductory Questionnaire

This short questionnaire collects background information about participants' experience with crowdsourcing. When submitting the form, no identifiable information (e.g., name or email address) is collected unless explicitly provided by the participant.

1. **To remove any identifiable data from your answers, please type the number provided by the researcher in the field below.** *(Required)*

2. **How many crowdsourcing jobs have you previously published?** *(Required)*

3. **What answer best reflects your experience with designing incentives for crowdsourcing?** *(Required)*

   - I haven't published before, so I don't know.
   - I think about incentives late in the process.
   - Incentives feel like a necessary task, but I do not mind doing them.
   - I rely mostly on defaults or quick estimates.
   - I spend minimal time on incentive design.

4. **Are you currently designing a crowdsourcing job?** *(Required)*

   - Yes
   - No

5. **What is the topic of the crowdsourcing job?** *(Required)*

C

Appendix B: Tool Feedback Questionnaire

1. **To remove any identifiable data from your answers, please type the number provided by the researcher in the field below.** *(Required)*

2. **Answer ID** *(Required)*

3. **Answer the following questions based on your experience using the tool.** *(Required)*

	Very difficult	Somewhat difficult	Neither difficult nor easy	Somewhat easy	Very easy
How easy or difficult was it to overall use the tool?	○	○	○	○	○
How easy or difficult was it to navigate through the steps?	○	○	○	○	○

4. **Did you find the information provided in the question text useful for the design process?** *(Required)*
   - ○ Yes
   - ○ No

5. **Would you consider using the tool in your work?** *(Required)*
   - ○ Yes
   - ○ No
   - ○ Maybe

D

Appendix C: Usage of Artificial Intelligence

QuillBot

QuillBot[1] was used to rephrase ideas. The tool was configured with the "Fluency" tone setting and minimal synonyms to maintain the original voice and conceptual integrity while improving clarity and readability. This was primarily applied to ensure consistent writing quality and professional presentation throughout the thesis.

ChatGPT 5

The *ChatGPT Free version*[2] was used for technical guidance in developing and deploying the research tool due to the limited knowledge of product development from scratch. This included architecture recommendations, implementation strategies, debugging assistance, and deployment procedures. During the development phase, the AI tool served as a technical resource and advisor; the author was responsible for all design decisions and final implementation.

ChatGPT 5 was also used to help create scenarios for the user study. The AI tool assisted in verifying that the scenarios were equally difficult and similar in scope, keeping the user study design consistent and fair.

Claude

Claude Haiku 4.5 and Sonnet 4.5[3] was used to provide technical assistance during the data analysis phase. It specifically helped in the development of a data extraction script for retrieving the collected user study data from Firebase in a structured format. It also assisted in writing scripts for Pyplot visualizations to represent the analysis results in the project document. The analytical approach, data interpretation, and conclusions drawn from the visualized data remain the sole responsibility of the author.

[1] Quillbot: https://quillbot.com/paraphrasing-tool
[2] ChatGPT: https://chatgpt.com/
[3] Claude: https://claude.ai/