Crowdsourcing as part of producing content for a critical reading comprehension game

University of Oulu

Information Processing Science

Master's Thesis

Stefan Grundström

2023

Abstract

This thesis aimed to examine how crowdsourcing can be used as a part of content creation for a critical reading comprehension game on a topic, misleading graphs, that are difficult for people to interpret. In crowdsourcing tasks, the worker is shown a graph that is intentionally designed to be misleading, from which the worker is supposed to create four headline options that are used as content of a critical reading comprehension game. To ensure the quality of the headlines, they are validated using crowdsourcing and two expert evaluators. As a result of the thesis, a graphical user interface was created to manage crowdsourcing projects.

The major challenge of crowdsourcing is quality control when unknown people from different backgrounds perform tasks on a different basis. The tasks were formed around a tricky topic, in which case it is difficult to keep the amount of usable data high in relation to the total amount of gathered data. The topics of the graphs and the task interface were intentionally designed to be simple so as not to take too much focus from the context of the misleading graph.

The results show that there is a lot of variation in the quality of the responses although an effort was made to select the best among the workers. It was noticeable that misleading graphs or assignments were often misinterpreted in the task of creating headlines. A small part of the responses was completely in accordance with the assignment. In the task of validating headlines, the worker's task was to determine how well the headline formed in the previous task corresponded to the assignment. The results show that it was too easy for the worker to click and move on to the next task without proper consideration.

Keywords
Crowdsourcing, misleading graphs, critical reading, information systems

Supervisor
Professor, University of Oulu, Mika Mäntylä

Postdoctoral researcher, University of Oulu, Miikka Kuutila

Abbreviations

API Application programming interface

CDN Content Delivery Network

CS Computer science

DS Design science

DSR Design science research

IS Information Systems

IT Information Technology

QC Quality control

QCC Quality Control of Crowdsourcing

SDK Software development kit

UI User interface

URL Uniform Resource Locator

WoC the Wisdom of the Crowd

Contents

1 Introduction

The popularity of crowdsourcing has grown strongly in industry and academia. The work done by crowdsourcing is not dependent on time and place, which has brought a change in people's work-life balance. (Deng, X., 2016.) Providing opinions or ideas, labeling images, or transcribing text are all examples of tasks that computers are not particularly good at and are well suited for crowdsourcing. (Daniel, F., 2018.) A typical crowdsourcing task is called a micro task because it is simple and can be usually completed in minutes. Completing the task will be rewarded with a small monetary prize. (Deng, X., 2016.)

Although crowdsourcing could make it possible for a large group of workers to perform tasks cost-effectively within a short timeframe, it also has its own challenges. Many of the challenges are related to data quality and its control. (Hettiachchi Mudiyanselage, D. E., 2021; Kittur, A. 2008). The fact that workers are unknown, and the requester receives only limited information about them through the crowdsourcing platform creates its own challenges. The suitability and motivation of the workers are tried to be ascertained using different quality control methods because there are also a lot of poorly performing workers, such as spammers, whose aim is only to make money. (Venetis, P., 2012; Jin, Y., 2020). Quality control has been studied in many research and various methods have been developed for quality improvement, but it is still one of the most significant research targets in crowdsourcing. (Hettiachchi Mudiyanselage, D. E., 2021; Kittur, A., 2008). Insufficient quality control has often been seen to have an impact on the quality of the responses. The importance of quality control is also emphasized by the fact that even an ethical worker can produce poor-quality responses when he or she misinterprets the purpose of the task, which may be the result of poor worker introduction or selection for the task. (Le, J., 2010).

Graphs are an effective way to present various statistics of everyday life. (Lauer, C., 2020.) Compared to textual data, graphs are a quick way to get a general view of statistics. (Garzón-Guerrero, J. A., 2020.) Since graphs inspire readers' trust, making graphs an easy target for misinformation. (Lauer, C., 2020.) A misleading graph is based on valid data, but it is somehow manipulated so that it gives a misleading impression. (Kiili, K., 2021.) Misleading graphs may be misinterpreted even after their misleading methods have been taught. (Yang, B. W., 2021) Then the reader lacks graph literacy, which refers to the ability to read and understand graphs. (Kiili, K., 2021). According to Kiili (2021), there is little previous research on misleading graphs compared to well-formed graphs, therefore, more research is required in its field.

This thesis combines crowdsourcing and misleading graphs. Previous research has shown that both have their own challenges. Neither of them is unambiguous and may produce, among other things, misunderstandings, and low-quality responses. There are several different platforms for crowdsourcing, of which Toloka was selected for this thesis. One of the requirements of the crowdsourcing platform was an API, which, in addition to Toloka, was offered by, for example, Amazon Mechanical Turk (MTurk) and Microworkers. The final choice was most influenced by the documentation, which Toloka had clearer than the other alternatives. Crowdsourcing project management requires familiarization, and many manual steps and sufficient quality control is needed for proper data quality. This thesis aims to find out if it is possible to produce usable content for a critical reading comprehension game by crowdsourcing on a topic, that is a misleading graph, that is also difficult for humans to interpret. Along with that, this thesis tried to find ways to bring crowdsourcing project management closer to the content creation processes of the critical reading comprehension game and to automate crowdsourcing projects' manual steps as much as possible. During the research, a prototype of the tool was implemented, which makes it possible to create content, that is headlines, for the game.

1.1 Research Problem and Method

Crowdsourcing makes it possible to perform tasks that are difficult or impossible for computers to perform. (Daniel, F., 2018). This thesis aims to find out if is it possible to generate usable content for a critical reading comprehension game using crowdsourcing from a subject that is also difficult for humans to understand. Crowdsourcing tasks involve misleading graphs and news headlines.

The thesis has three different research questions:

R1: Can crowd workers create a variety of news headlines based on misleading graphs?

R2: How to determine that the material produced is usable and correct for the critical reading comprehension game?

R3: How to integrate crowdsourcing project management into a part of a critical reading comprehension game's content creation by utilizing a crowdsourcing platform's API.

The aim is to answer these questions using the Design Science Research method. The literature review is an essential part of the research, as it provides information on the problems and solutions of previous research. Previous research is also used in the implementation of this research.

1.2 Structure

This thesis begins with an introduction that provides background on the starting points and motivation of the research. The research questions are also presented in this section.

Chapter 2 called Background reviews previous research about crowdsourcing and misleading graphs. The main focus is on crowdsourcing, which is discussed in several subsections.

Chapter 3 called Research Methodology discusses the thesis research method.

This follows the Prototype development chapter where the Prototype-related topics include requirements, design, and implementation. The creation and validation of the headlines are discussed in their own subsections, as they have also been implemented as separate phases.

The Data Gathering chapter describes how data has been collected from the crowdsourcing platform and its workers.

The Data Analysis chapter describes how the gathered data has been analyzed.

The Results chapter presents the research findings.

The Discussion chapter goes through the facts that appeared in the research. The chapter answers the research questions and discusses the limitations and future work. The results of the research are also compared to previous studies.

Finally, there is the Conclusion chapter which summarizes the main findings of the research.

2 Background

This section reviews previous research that is relevant to the thesis. Topics include crowdsourcing, data quality, and misleading graphs. The broad concept of crowdsourcing is further divided into different subsections.

2.1 Crowdsourcing

It has been a while since Howe (2008) introduced the term crowdsourcing in 2006. The basic idea of crowdsourcing was to involve volunteers in completing tasks via the network that would otherwise be done by traditional employees. This opens the possibility of using labor around the world and at the same time with a lot lower costs. (Howe, 2008). The term crowdsourcing has since evolved while retaining its basic idea. In the same way as the employer, which in this case is called the job requester, the various employment opportunities for the employment of the employee i.e., a crowd worker, also increase. However, these terms cannot be directly compared because in crowdsourcing there is no employment relationship and no obligations between the jog requester and the crowd worker. Because the crowdsourcing tasks are voluntary, the crowd worker is free to choose which tasks she or he wants to perform, when, and from where. Crowdsourcing tasks are called micro tasks which are "an open source form of micro work" and likewise, micropayments are paid for completing them. The tasks that job requesters post are available on crowdsourcing platforms. Once the tasks are available, they are expected to be completed by a previously unknown group of individuals. (Deng, 2016).

Lorenz et al. (2011) describe the data collected through crowdsourcing as the "wisdom of the crowd". Crowdsourcing is popular in the scientific community and industry because of its nature. A large amount of data can be collected efficiently regardless of time and place.

Generally, crowdsourcing is used to perform simple tasks for a small payment. However, crowdsourcing is not limited to tagging images or other simple tasks, it can also be used to perform complex tasks such as article writing. Other examples of crowdsourcing tasks could be testing new technology, designing, and analyzing data. (Yung, 2014.) Crowdsourcing is also good for tasks where that computers are not particularly good at and that require human knowledge and contribution. Examples of those tasks are product recommendations and image labeling (Nguyen, 2021).

2.1.1 Crowdsourcing Platforms

According to Geiger (2014), a common factor for crowdsourcing platforms is how they have approached workers. A crowdsourcing platform can have a large group of workers around the world that decide freely which tasks to take and contribute. Activities are entirely voluntary which applies to all crowdsourcing platforms. Organizations can get workers with various backgrounds outside the organization if they are interested in the task given and feel that they can complete it. (Geiger, 2014.)

On crowdsourcing platforms, the roles can be divided into two parts: a requester and a worker. A requester is a person responsible for publishing the tasks, accepting the responses, and rewarding workers. A worker is a person who performs tasks on crowdsourcing platforms and expects payment in return. (Hettiachchi Mudiyanselage, 2021.)

Crowdsourcing platforms provide a way to engage a large group of workers within a short timeframe with low costs. Even so, a requester may require changing the original plan of the task if the quality does not meet expectations. That can lead to a larger sample size, an extension of the deadline, or an increase in the reward. In addition, for the sample result quality, it is worth focusing on the design of the task even more when workers perform tasks that seek subjective or qualitative responses. (Kittur, 2008.)

Crowdsourcing platforms such as Amazon Mechanical Turk or Toloka provide a wide range of features for typical crowdsourcing processes. These include, for example, creating tasks from different templates, publishing created tasks, defining quality control, and monitoring the progress. (Ramírez, 2020.) In this thesis, Toloka's platform has been used.

2.1.2 Crowdsourcing and Organizations

From an organizational point of view, crowdsourcing is a way to gain expertise called crowd capital from outside own organizations. Crowd capital could be external experts, laborers, or some specialized skills. The worker is motivated by the opportunity to, for example, work as a freelancer, develop his or her own skills or gain some sort of recognition. (Feng, 2018.)

According to Tran-Thanh (2014), the importance of crowdsourcing in the operations of organizations is growing rapidly. At its best, it can reach millions of workers who are ready to perform assigned tasks. Although that number sounds promising, the requester must be able to find among them workers whose skills match the requirements of the task. Tran-Thanh sees dividing complex tasks into smaller tasks to be one of the biggest crowdsourcing challenges. This is related to the requirement to ensure that quality remains consistent for each individual task while keeping costs as low as possible. In addition, from an organizational perspective, this all should work in a complex workflow.

2.1.3 Challenges

In crowdsourcing a wide variety of individuals with different backgrounds and skills are involved to perform tasks, which creates challenges for project management. For this reason, for example, it is particularly challenging when aggregating the responses. Aggregation techniques have been developed but they have their own shortcomings as they consider different factors such as worker expertise and task difficulty. This makes it difficult to find the right aggregation technique for own usage. (Nguyen, 2021.)

Lack of workers' background information also causes difficulties in job performance as traditional work is more predictable and consistent than it is in crowdsourcing. This is largely due to the fact that workplaces have the same employees whose characteristics are known, and they have a certain type of expertise. In crowdsourcing tasks, worker turnover can be high, and little is known of them. Workers are not obligated to perform tasks, in which case tasks are performed for their own reasons. Therefore, successful data collection in crowdsourcing is usually the result of good quality control. (Lease, 2011.)

According to Venetis (2012), crowdsourcing platforms have one common problem with workers' performance. Among the workers are those who are seriously trying to do tasks the right way and produce good material. On the other hand, there are those workers who perform poorly based on their skills or those who just want to earn money from non-existent work input. The latter can be called a spammer.

Different types of workers also bring their own challenges to the quality of responses. Vuurens (2011) introduces the types of workers present on the crowdsourcing platforms: ethical workers and spammers. The most desirable type of these is called ethical workers. Ethical workers follow instructions and produce appropriate data. This type of group can be divided into two subtypes: proper workers and sloppy workers. Proper workers are those who respond as desired and are accurate. Unfortunately, sometimes even an ethical worker can choose a task that is not suitable for him or her, in which case it will appear as poor data quality. These workers who intend to produce relevant data but still produce poor quality are called sloppy workers. Next are the spammers, who can also be divided into two subtypes: random spammers and uniform spammers. Uniform spammers are easier to detect as they are constantly responding in the same way. Random spammers do not want to get caught spamming, so instead of schematic responses, they produce randomized responses. (Vuurens, 2011.)

Although crowdsourcing platforms provide the worker the freedom to choose any tasks that may be suitable for him or her based on his or her preferences, an excessive number of opportunities can result in “an information overload situation”. Such a situation may result in a worker performing tasks that may not be his or her strength. If the worker cannot find completely suitable tasks, he or she satisfies less suitable alternatives. This is reflected in the worker’s motivation and the quality of the result. This is also problematic in the sense that real potential workers for a particular task may miss the opportunity, which may result in the loss of several good contributions. (Geiger, 2014.)

Crowdsourcing is efficient and cost-effective at best, but it has its own challenges that do not necessarily occur in traditional work. The biggest challenges are ensuring the quality of responses and detecting and excluding bad workers from the crowd. As stated in previous research, quality control will also be an essential part of this thesis, and its implementation is described in section 4.4.

2.2 Data Quality in Crowdsourcing

Data quality, which can also be called the accuracy of output data, is one of the key topics in the research of crowdsourcing. The terms goodness and correctness can also be used in this context. Since crowdsourcing has a wide range of workers with different skills and motives, data consistency can be considered an essential part of determining data quality. This means, in short, that different workers produce similar answers to a task. (Daniel, 2018.)

According to Hettiachchi Mudiyanselage (2021), crowdsourcing has its challenges related to the quality of the collected data. A lot of research has been done on it and it is still a major factor worth exploring to improve the data quality. Based on previous research, various methods have been created to improve quality. Examples of viable solutions found are task design, feedback to workers, and the usage of different aggregation methods. Not all quality improvement methods are suitable for every task, but they are developed in the sense that they fit as many tasks as possible. One important point is that the workers are unknown, and the requester receives only a little information about them through the crowdsourcing platform. This highlights the importance of data quality improvement measures. (Hettiachchi Mudiyanselage, 2021.)

There may be additional costs associated with crowdsourcing in ensuring the best possible quality of data. In addition to micropayment for completing a task, the requester could reward workers who produced good-quality responses with an additional bonus. These kinds of explicit methods are moderately predictable. However, methods, such as analyzing worker behavior, may require some indirect costs that are harder to predict. (Hettiachchi Mudiyanselage, 2021.)

It is typical for crowdsourcing to offer the same task to several workers after which those responses are aggregated into one most accurate response. For this process, Hettiachchi Mudiyanselage uses the term “Truth Inference”. (Hettiachchi Mudiyanselage, 2021.) Such a redundancy-based strategy is used because the crowd may contain workers who produce poor-quality responses. The correct response is obtained by using some aggregation method such as majority vote, where the majority determines the truth. (Zheng, 2017.)

Malicious workers have a wide variety of means of abuse. They may even use bots to automate their responses. It may also be problematic that workers may cooperate, that is, for example, share information forward, in their responses, thus reducing the quality of the data and the effectiveness of the aggregation method used. (Hettiachchi Mudiyanselage, 2021.) Other growing concerns worth mentioning as a result of poor-quality control include risks to financial, intellectual property, and privacy, malicious attacks, and project failure. (Daniel, 2018.) So, in crowdsourcing, it would be important to be able to detect both intentional and unintentional poorly done responses among good-quality ones.

In the following subsections, quality control is discussed in four different aspects: workers’ motivation and abilities, platform-specific data quality, aggregation, and task design. In each subsection, issues that must be considered in quality control for the respective aspect are discussed.

2.2.1 Workers’ Motivation and Abilities

Worker filtering can maintain better quality even in situations where the number of responses or the reward for each answer has had to be limited. This has worked, at least when it comes to a worker’s cognitive skills. However, testing of cognitive skills must be done separately, in which case it incurs costs. (Hettiachchi et al., 2019.) Quality is also affected by other differences between workers. These factors are, for example, skills, motivation, and background. There is a large group of workers registered on various crowdsourcing platforms to whom it is possible to assign tasks. The problem is to single out unsuitable workers from suitable ones. There are those workers that try to earn with the least effort possible, give incorrect responses, or are otherwise dishonest. Indeed, one of the most significant problems with crowdsourcing has been quality control. (Qiu, 2016.)

One way to select the right workers for the task is to review workers’ experiences through different tasks on the crowdsourcing platform. Earning badges or similar platform-related certificates reflects a worker’s performance in tasks. Badges also have another meaning, for which earning them can also motivate workers. (Daniel, 2018.) In terms of the weight of the responses, how to compare an expert and a non-expert, to aggregate as high-quality data as possible. By default, responses from the expert are of better quality, but this is not always the case. (Tran-Thanh, 2014.)

While workers’ expertise is an important factor in generating good quality data, the importance of motivation should not be overlooked. Through motivation, a worker strives to make a certain kind of result. Workers with the same characteristics, the one with better motivation is more likely to perform better in the task and produce better quality data. An extreme example of poor motivation is a worker who performs tasks only for money, regardless of the result. (Jin, 2020). Motivation can be divided into two parts: extrinsic and intrinsic. In crowdsourcing, incentives can be used to influence workers' motivation. The type of incentives used affects the outcome. "Reward-driven", associated with extrinsic motivation, incentives are for faster task completion, and "interest-driven", associated with intrinsic motivation, incentives are for higher quality. (Daniel, 2018.) It is also good to note that although the reward is one of the ways to motivate, workers react differently to the amount of the payment, making budgeting difficult to ensure good responses (Tran-Thanh, 2014).

According to Le (2010), training workers before the actual tasks could improve the quality of responses. In training, the requester knew the correct answers, and the workers were given immediate feedback on their answers. In this way, misunderstanding among ethical workers can be reduced and unethical workers can be removed from the crowd. This also had a unifying effect on responses.

The selection of workers has great importance for the quality of the responses. In this regard, filtering is one efficient way to eliminate unsuitable workers. However, expertise is not necessarily enough for a good result, so it is important to pay attention to the worker's motivation as well. Different means of motivation are, for example, bonuses, feedback, and various certificates or other recognitions. These means aim to encourage the workers to perform the tasks as well as possible.

2.2.2 Platform-specific Data Quality

Concerns about data quality have led the research community and crowdsourcing platform providers to take action to ensure better quality. Workers can be tested before they are allowed to perform tasks. Their performance in other tasks can be monitored. Various techniques can be used to find lazy and poor-quality workers based on statistics. (Qiu, 2016). Toloka is one of the crowdsourcing platforms that provides a way to track and identify workers alongside the tasks they perform. Workers can be assigned skills that can be updated according to how well they perform. The correctness of the answers or a behavioral approach can be used to determine skills. Skills can also be used to filter workers, for example by preventing workers with a too-low skill level from entering a task. All this can be done automatically. (Hettiachchi Mudiyanselage, 2021.)

Both the requester and the platform provider share the same problem of quality control. Redundancy on the same tasks is commonly used for the problem of the mixed skill level of workers. This means assigning the same task to several workers. Most of the platform providers also offer the opportunity to review responses before accepting them. In this case, bad-quality answers can be rejected and left unrewarded. (Baba, 2013.)

Filtering is also one of the common quality control methods used on crowdsourcing platforms to reduce unsuitable workers. Some of the unsuitable workers can be filtered before the actual tasks by a quiz. Workers who pass the quiz will be able to do the actual tasks. At this point, it is possible to filter poorly performing workers by including control questions among the real questions. (Jin, 2020)

Golden standard questions are the tasks for which there is a correct answer that the requester knows. The purpose of these tasks is to obtain information about the quality of the worker's responses. It provides little contrast for evaluating other task responses of the worker. With this technique, it is possible to find and filter some poorly responding workers. More complex techniques consider the actions of the worker on the crowdsourcing platform. Some crowdsourcing platforms, for instance, MTurk provide the ability to select workers according to their backgrounds like gender and age. Previous research has shown that this also has an impact on the quality of workers. (Hettiachchi et al., 2019.)

Crowdsourcing platforms offer different ways to improve data quality. The suitability of the workers and the quality of the answers can be monitored, measured, and tested before, during, or after the task starts. This thesis utilizes Toloka's existing quality control solutions to improve data quality. Not every quality control method is suitable for every task, in which case it is necessary to find the most suitable methods for this research.

2.2.3 Aggregation

If it is not possible to identify and exclude malicious workers before the tasks are performed, filtering can also be done afterward during the aggregation process. (Hettiachchi Mudiyanselage, 2021.)

In terms of quality control, a way must be found to assess the output of workers who performed the same task. How these can be judged against each other and how much weight each one gets. The golden standard evaluates only a worker's performance on itself in which case it cannot be used directly to evaluate other tasks. In addition to this limitation, it is laborious and expensive to maintain. Instead of the golden standard, Zhu proposed iterative voting. This technique also adds work, but it can be used to evaluate each task, and in this technique, the workers do the extra work rather than the requester. The assessment of the quality of the specific task is based on the general opinion of the workers which Zhu uses the term 'input agreement'. (Zhu, 2012.)

The wisdom of the crowd (WoC) is a quality control of crowdsourcing (QCC) method that is used during or after task completion. Its main target is to unify the responses by giving weight to the responses that have been generally considered to be correct and again deducting weight from the responses that have been generally considered to be incorrect. Workers or their contributions are not excluded from others in this method. Thus, the worker does not have to perform every task with 100 percent accuracy. This method has its requirements to work properly: redundancy and aggregation. There should be enough workers per task and most of the workers should be reliable. The majority vote is an example of the WoC approach, where each answer is equal, and the most common answer is considered the correct one. (Jin, Y., 2020). Vuurens (2011) also considers the number of workers to be an important factor. There should be more workers per task to make the outcome as reliable as possible. The answers can then be aggregated into one that best suits the task. Vuurens mentions that one of the most common techniques is to use majority voting. However, the more sloppy workers or spammers appear among the crowd, the more majority voting loses its effectiveness. (Vuurens, J., 2011.)

The weaknesses of basic quality control methods such as golden standard questions and qualifications tests, which are often used in crowdsourcing, are their limitation to a specific task. In addition to being laborious, these methods are sometimes tricky to implement, for example in a situation where the requester does not know the correct answer, or it does not exist. Instead of these basic methods, Hettiachchi, D. et.al. (2020) suggests quality control methods based on workers' past performance on the crowdsourcing platform. Tasks can only be performed by suitable workers based on their history. These methods are dependent on the history of the worker, so it does not work for a worker with little or no activity on the crowdsourcing platform. A combination of quality control methods, such as removing erroneous workers, and quality metrics, such as task completion time, has been proposed as a solution to this problem. (Hettiachchi et al., 2020.)

2.2.4 Task Design

According to Ramirez (2020), task design is an important factor in data quality in such an environment with a wide range of people from different backgrounds. Task design is not just about the task interface, it should include all the necessary guidance so that contributions could reach the goals set. The task design usually takes its latest form through iterative steps where the aim is to learn from mistakes. The quality of the instructions is important as it can reflect the quality of the responses. The task interface could speed up the execution of tasks and it also has an impact on quality. In addition, the reward should be proportionate to the complexity of the task.

Daniel (2018) finds that the task description quality has had a direct impact on workers' performance and motivation. The worker must understand the task description correctly to give the right kind of answer. In addition, the worker may omit the task due to a difficult task description. Like the task description quality, the quality of the user interface has been also found to have an impact on workers. A user interface that emphasized issues such as user-friendliness and understandability, attracts workers and makes them perform better. A high-quality user interface can also be used to better control spammers.

2.3 Misleading Graphs

Graphs allow the statistics to be presented in an illustrative way to the public. This way could help to understand statistics, but in certain situations where the graph does not follow general design practices, it can also have a misleading effect. Misleading graphs are made either by the author's incompetence or intentionally. Either way, misleading graphs have become a growing problem because of graph design tools that let everyone, even inexperienced people, create graphs. Misleading graphs are also problematic in that people have difficulty interpreting them critically. This problem occurs even if they are aware that the graph is misleading. (Ramly, 2021.)

Misleading graphs are incorrectly formed graphs that can distort the reader's perception of the data. Often a distorted perception is also due to the reader's lack of critical reading of the graph which Garzón-Guerrero (2020) called statistical sense. Statistical sense can be divided into three factors: statistical and graphical literacy, statistical reasoning, and attitudes toward statistics. Statistical and graphical literacy includes sufficient knowledge of the necessary terms and concepts about graphs and statistics to critically interpret graphs. Statistical reasoning is problem-solving in which, as the name implies, reasoning seeks to examine the data given and the way it is presented. (Garzón-Guerrero, J. A., 2020.)

Yang (2021) addresses her empirical research that graphs can mislead people even if they have been taught the misleading method used in the graph. Teaching methods reduced misunderstandings but were not completely eliminated and occurred throughout the research. The y-axis truncation method, "truncation effect", was used in the research. This method aims to exaggerate the differences so that even a small deviation can appear large.

Graphs also have the special feature that they can be based on truthful data but can still be misleading. There is also some evidence that graphs have a more favorable effect on humans than other ways of presenting statistics, even when they do not provide additional information. (Yang, 2021.)

There is mainly research on graph literacy that focus on well-formed graphs rather than misleading graphs. According to Kiili (2021), "Graph literacy refers to the ability to read and understand graphs". Successful reading of a graph requires three different areas: Reading the data, reading between the data, and reading beyond the data. As misinformation and disinformation continue to increase, critical reading skills are seen as very valuable. (Kiili, 2021.)

According to Frees (1998) graphs are more prone to abuse than text-based statistics, which is largely due to their flexible nature, but also to their complexity of interpretation. Frees presents a list of eight items for graph design that the creator should consider to make the graphs easier to interpret:

(1) Avoid chartjunk

(2) Use small multiples to promote comparisons and assess change (3) Use complex graphs to portray complex patterns

(4) Relate graph size to information content

(5) Use graphical forms that promote comparisons

(6) Integrate graphs and text

(7) Demonstrate an important message

(8) Know the audience. (Frees, 1998.)

Davis (1999) mentioned two particularly harmful ways to mislead the reader: non-zero baselines and distorted scales. Non-zero baselines distort the proportions that give the impression of large differences. The non-zero baseline can also be done in such a way that the baseline is marked as zero but in reality, this is not the case. This makes the interpretation of the graph even more difficult. Davis underlines the fact that a graph should be styled in such a way that each feature in it has a purpose and fits the data. However, if there are features in the graph that have no actual meaning, the reader must be able to easily identify them.

Lauer (2020) introduces some deceptive tactics with different graphs. For bar and line graphs, the common way is to truncate the y-axis. The parts of the pie chart can be distorted by the 3-D effect. The posterior portions are made to appear smaller by placing the pie chart at a certain angle or increasing the height of the anterior portion. Changing the size of the bubbles in the bubble chart can exaggerate the differences between them. In addition, titles and other textual data may affect the reader's conception, but this has less of an effect than the graph itself. This fact highlights the effect of the graph on perception formation.

There is a lot of research on effective visualization to describe data and it receives a considerable amount of attention. On the other hand, its negative side, misleading, has received less attention, even though it has become one of the biggest concerns of information dissemination. Whether the graphs were well-formed or misleading, the reader also has a major role in achieving the right result and that is forming the right kind of perception. (Nguyen, 2021.)

3 Research Methodology

In this section, the thesis research method, design science, is discussed. Then, the project in general will be reviewed.

3.1 Design Science

Design science's main contribution is to create and evaluate IT artifacts that could solve identified problems. It differs from natural sciences and social sciences whose main purpose is to try to understand reality. Peffers et al. (2007) defined six steps of design science research: problem identification and motivation, the definition of the objectives for a solution, design and development, demonstration, evaluation, and communication. The artifact evolves with the research as it starts with the functional-level problems and gradually moves towards more detailed solutions using the listed steps. (Peffers, 2007.)

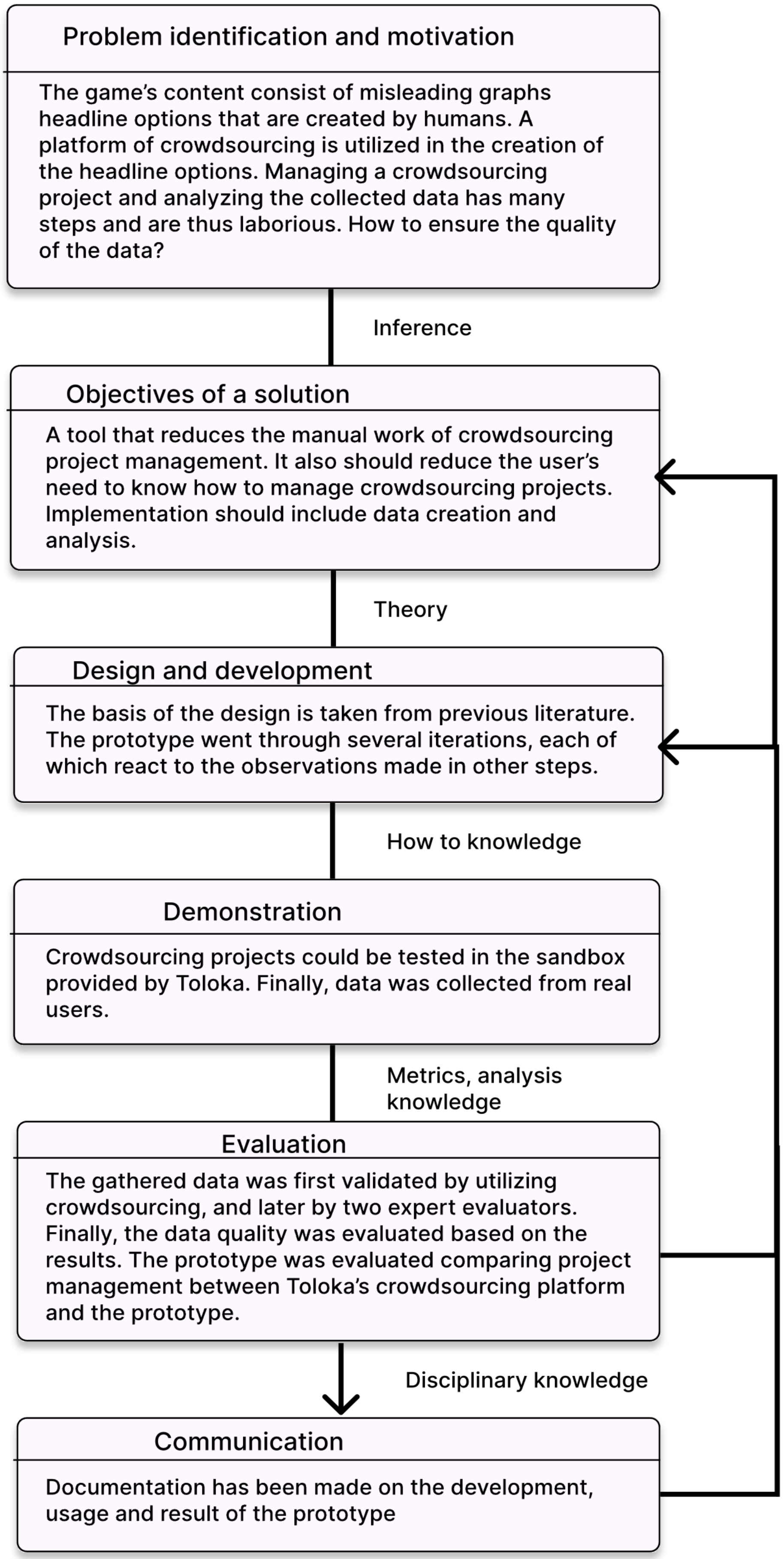
Information systems (IS) seek to solve organizations' problems using information technology. It is a discipline of applied research that makes extensive use of the practices of other disciplines. (Peffers, 2007.)

Peffers (2007) emphasizes that the research processes should be rigorous from development to evaluation. The artifact produced in the research should meet certain requirements. The artifact has been developed for a specific problem, so it should be relevant to it and no such solution to the problem has been made before. The development of the artifact utilizes previous theories and knowledge. Utility, quality, and efficacy are important aspects of evaluation.

Offermann et al. (2009) compare different approaches to design science research. The processes are divided into three different phases: problem identification, solution design, and evaluation. This thesis mainly follows Peffers' approach but has also been influenced by Offermann's approach where literature research is in a significant part in different phases. Peffers' design science research first phase called problem identification includes as the name implies problem identification and motivation, and objectives for a solution. Design and development are done in the solution design phase. The evaluation phase is where the demonstration and evaluation are performed.

3.2 Project

In this thesis, design science was used as the research method. The research proceeded according to Peffers' (2007) definition, which has six different steps introduced in the design science chapter. Figure 1. shows the six steps which have been followed in this research. The research started with the identification of the research problem. Although some requirements and objectives had already been set for the project, they were far from complete and required further specification.



Problem identification and motivation

The game’s content consist of misleading graphs headline options that are created by humans. A platform of crowdsourcing is utilized in the creation of the headline options. Managing a crowdsourcing project and analyzing the collected data has many steps and are thus laborious. How to ensure the quality of the data?

Objectives of a solution

A tool that reduces the manual work of crowdsourcing project management. It also should reduce the user’s need to know how to manage crowdsourcing projects. Implementation should include data creation and analysis.

Design and development

The basis of the design is taken from previous literature. The prototype went through several iterations, each of which react to the observations made in other steps.

Demonstration

Crowdsourcing projects could be tested in the sandbox provided by Toloka. Finally, data was collected from real users.

Evaluation

The gathered data was first validated by utilizing crowdsourcing, and later by two expert evaluators. Finally, the data quality was evaluated based on the results. The prototype was evaluated comparing project management between Toloka’s crowdsourcing platform and the prototype.

Communication

Documentation has been made on the development, usage and result of the prototype

Figure 1. The design science steps.

The project in this thesis is part of a larger project (Figure 2.). This larger project aims to implement a critical reading comprehension game themed around misleading graphs. During the thesis, the focus is on one part of the game project which is to implement a tool that can be used to produce content for this game by utilizing crowdsourcing. The game's content consists of misleading graphs and headlines. The misleading graphs are to be implemented as part of, yet another project related to the larger project, but it is not within the scope of this thesis. These graphs are only intended to be used in the tool implemented in this thesis. The tool is used to generate headlines based on graphs and descriptions that are entered into it. That is, the responsibility of this thesis project is only to create and validate the headlines for the larger project. However, it was necessary to make some graphs within this project for scheduling reasons.

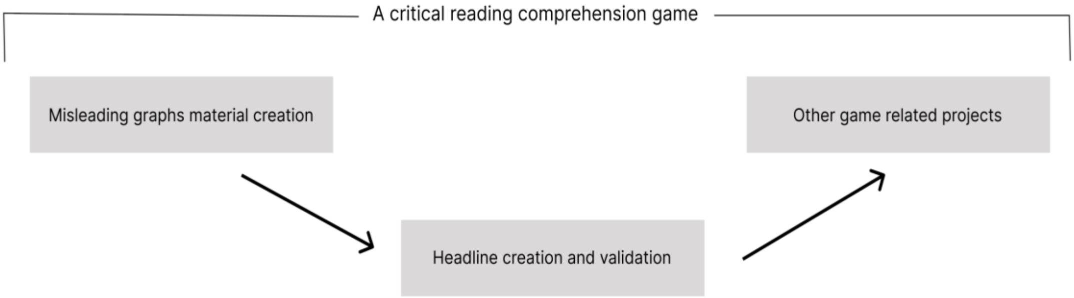

A critical reading comprehension game
Misleading graphs material creation
Other game related projects
Headline creation and validation

Figure 2. The division of the game project.

The thesis project started with the initial requirements which were very general. At this point, the idea was to use crowdsourcing to create and validate content based on misleading graphs for a critical reading comprehension game. Crowdsourcing project management would utilize the API of a crowdsourcing platform. The subject of the data could be related to the credibility of social media or news. The requirements evolved as the project progressed, but the main idea remained the same. The reasons for this were, for example, the progress of other projects and the increase in information gained during implementation.

Figure 1. summarizes the main points and steps of how design science research has been used. This thesis's starting point was the need for the game project. From that onwards had to look at previous research on these two topics, misleading graphs, and crowdsourcing. In addition to this different crowdsourcing platforms had to be compared, from which the most suitable one was chosen for this research. Toloka was selected based on the developer-friendly API and its documentation, which was clearer than the other alternatives that were considered. Toloka is a cloud-based crowdsourcing platform that helps you to collect and analyze data. Next, had to find out how the platform works from the requester's point of view. In terms of planning and implementation, it was necessary to know the different stages, possibilities, and possible shortcomings. These processes before design and implementation were related to the design science research problem-identification and the definition of the objectives for solution steps. Based on the acquired knowledge and the given requirements, the final research questions were formed. However, the definition of the objectives for the solution step did not end here, they developed and increased through the different iterations of the project.

The research and the implementation of the prototype progressed one iteration at a time meaning that different steps were repeated several times whenever necessary. Previous research was the basis when different areas were planned and implemented. Especially the research related to quality control and the design of the crowdsourcing task was useful.

The developed prototype was first demonstrated through the sandbox provided by Toloka and later in a real environment with real workers. The demonstrations could occasionally reveal some deficiencies, leading to a return to previous steps. The demonstration step was moved to the evaluation step when sufficient data had been collected that is, one crowdsourcing project related to the headline creation and two related to the headline validation had been completed. In the evaluation of the data, workers' evaluations of the headlines were compared with the evaluations of the expert evaluators. The prototype was evaluated by comparing crowdsourcing project management between Toloka's crowdsourcing platform and the prototype. The last step was the communication step, which is based on this entire research and its documentation.

4 Prototype Development

In this section, the different stages of prototype development will be reviewed in more detail. For data gathering, three crowdsourcing projects were formed, and the related planning and implementation were done in separate phases, and the same also applied to the development of the prototype. In the following subsections, a similar division has been made, which starts with the headline creation theme, followed by the headline validation theme. Requirements and Quality Control also have their own chapters.

4.1 Requirements

During the project, a tool prototype should be developed, which can be used to create and validate headlines according to the assignment. The assignment includes a misleading diagram, a description of how the diagram is misleading, and instructions for creating the right kind of headlines or how validation should be done. The tool itself does not create headlines, but it creates tasks for the crowdsourcing platform, where volunteer workers perform them. Validation is also done in the same way.

Phase 1 initial requirements were to collect two headlines per task related to the given graph. One headline should be credible and the other misleading. These two headlines could be then compared using, for example, a pairwise comparison in phase 2. The goal is to find out how credible the headlines are. On this basis, the design and implementation of the tool began.

The concept of the critical reading comprehension game developed in the meantime, when the assignments changed in terms of the number of headlines. Now players were supposed to get four headlines per task. One headline is supposed to match the diagram, one should be in line with the misleading means, and the last two should be invalid in some way. In addition to misleading diagrams, assignments may include well-formed diagrams. Each diagram has an additional information field that indicates whether the diagram is misleading or not. The means of manipulating the diagram are also explained. The material is meant to be as simple as possible.

4.2 Phase 1 - Headline Creation

This chapter presents the planning and implementation of phase 1, called headline creation. Headline creation refers to an act where headlines are made by volunteer workers.

4.2.1 Planning

The idea of the project was to utilize one of the crowdsourcing platforms, so it was natural to start getting to know the subject area and the chosen platform. In terms of planning, it was good to gain experience with how projects were managed in Toloka's user interface. In which order things were done and what was possible to do. Going through Toloka's project life cycle was important for the tool so that it was possible to pay attention to the most laborious aspects. Next, a brief description of the different phases of Toloka's project managed via the user interface.

In the Toloka platform, it was possible to test projects in a testing environment, which is called a sandbox, before they are taken to real users for production. With the help of the sandbox, it was possible to simulate every phase of the project, right up to receiving and aggregating answers. The sandbox was also helpful for testing Toloka API requests before implementing them into the GUI. Testing API requests was done using an API platform called Postman. The sandbox could also be used to design the layout of the tasks.

In Toloka's user interface, the initialization of the project starts with selecting or creating the layout of the task. If you can find an option from the ready-made templates that are just right for your use case, this step is quick and easy. Designing and creating a task layout from scratch is a slightly more laborious alternative although the user interface has its editor. The editor can be Toloka’s template builder with its predefined elements or a basic code editor which supports HTML, JavaScript, and CSS.

The next step is to define the project’s general information, including the project’s name, description, and instructions. After that pools can be created for the new project. A pool has general information like name, description, and price. A pool can also have its audience and quality control. The audience consists of filters and skills that can be defined to select workers with certain characteristics. For example, the required educational background or worker’s age can be determined. Quality control is rules set for the pool, which should reduce low-quality answers and spammers. For example, a worker can be banned from the project if she or he responds to tasks too quickly or skips too many tasks.

Once the pool has been created, tasks can be added to it. The user interface accepts three kinds of files, which are XLSX, TSV, and JSON, where the tasks are defined according to the template. A task must contain all the information that is defined for the layout of the project’s task. For example, if an image is presented within the task, its Uniform Resource Locator (URL) must be specified in the file. If desired, the correct answer or a hint can be included in the task definition.

After uploading the tasks, the pool can be opened. After this, waiting until getting enough answers from the workers. Submitted answers can either be accepted or rejected. Finally, the answers can be downloaded as a file for personal use.

As can be concluded from the above flow of events, there are many steps in defining the project and some of them can be done in many ways, such as quality control or task design. During the design phase of the tool, had to consider the things that are done automatically and those that the user must do. These actions that the user takes need to be enhanced compared to Toloka’s user interface. In this project, the enhancements practically meant automating as many activities as possible. Adding tasks to Toloka’s user interface seemed tedious, especially when the tasks contained images. The images first had to be uploaded and store somewhere where they could be retrieved for Toloka’s task. After uploading, the image URL had to be specified separately for each task.

4.2.2 Implementation

The tool made in this thesis is a simple GUI with which Toloka projects are created and managed. Default values have been set for Toloka’s projects, according to which they are formed. In this first phase, the user's main task is to define the general information of the project and the pool, such as the project’s name and description, and enter tasks. The tool itself has specifications for everything else, such as quality control, which are determined according to the template chosen by the user. Limiting user actions in the tool is based on the fact that offering options on a wide scale would not have been significant for the thesis itself but would have significantly increased the development work. The aim of the tool prototype was to enable predefined use cases to create headlines for the critical reading comprehension game, but it was not intended to be the all-inclusive or final version of the tool.

The tool's functionalities are divided into smaller reusable components which have been made into three different sets. One is responsible for the functionality of the Toloka, the other for the GUI and the third contains otherwise important general functions such as image management. The idea of such division was that the development project is thus easier to manage and in further development as easy as possible to expand. It would be possible to use some components in a completely different implementation than in the GUI which is created using a python framework called Tkinter. Python was selected as the programming language because various libraries and software development kits, such as Toloka-Kit and Crowd-Kit, had been made for it, which facilitated and accelerated the development of the tool's prototype. In addition, Tkinter is built into its standard library which can be used to create simple cross-platform GUIs which means that the same code works on multiple operating systems, for example, Windows and macOS. In this way, there was no need to take over more technology and there was no need to separately consider the requirements of different operating systems.

Determining tasks was considered the biggest bottleneck in the management of Toloka's projects. The definition of tasks consists of several steps: user interface design and implementation, storing and using images, and task output field definitions. The configurations related to the user interface were implemented with the ready-made templates already mentioned. As for the images and output field definitions, a user-friendly solution was made for the tool, which combined both. The user does not need to leave the task creation view to upload images and does not need to know the URL of the images. ImageKit.io was used to store and retrieve images. ImageKit.io offers an image content delivery network (CDN), a media library, and a software development kit (SDK) for python. The media library takes care of the storage of the images, the image CDN delivers the images and with the python SDK, the necessary image management functionality can be integrated into the tool faster and easier. In the tool, the images uploaded to the media library were retrieved for the view of the new task definition and displayed on the table. From this table, the user clicked on a suitable image, after which its URL was automatically updated to the task definition.

In headline creation, the subject areas from which the graphs were created were tried to be kept simple. The purpose was that the task has no prior knowledge requirements, and anyone could create headlines according to the topic. This was done largely because misleading graphs are a difficult subject in themselves. Workers should be able to interpret the misleading graph and be able to perform the task based on it. There is no need for complicated background stories that take attention away from the main thing itself, that is, the interpretation of the graphs.

In addition to the subject, the user interface of the tasks was kept as simple as possible. Only one task was shown at a time to keep the worker's focus on the ongoing task. All the necessary information was visible at the same time. The input fields had placeholders and additional information bars that described what kind of headlines were wanted for it. The following subsections go through the user interfaces in more detail.

4.2.3 The Design of Tool

In Toloka, project phases are created in a certain order, which is why the tool also follows the same order in its user interface. The design of the tool is very minimalistic, and simplicity has been the starting point in the design of the user interface. The elements themselves are meant to describe their purpose.

The project is started by selecting the type (Figure 6). After this, a view of the project creation form opens to the user (Figure 7). The public instructions field, that is the task main instructions, is automatically filled with the template's default instructions. It is the user's task to fill in the remaining fields.

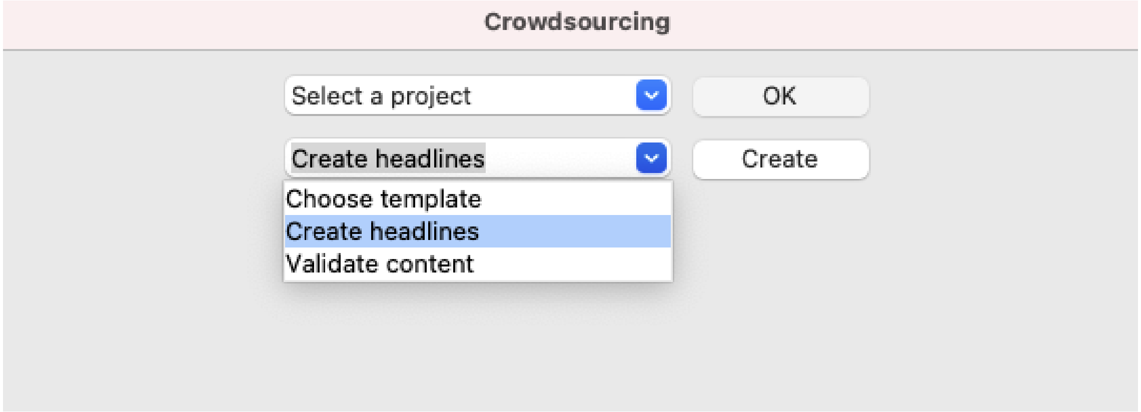
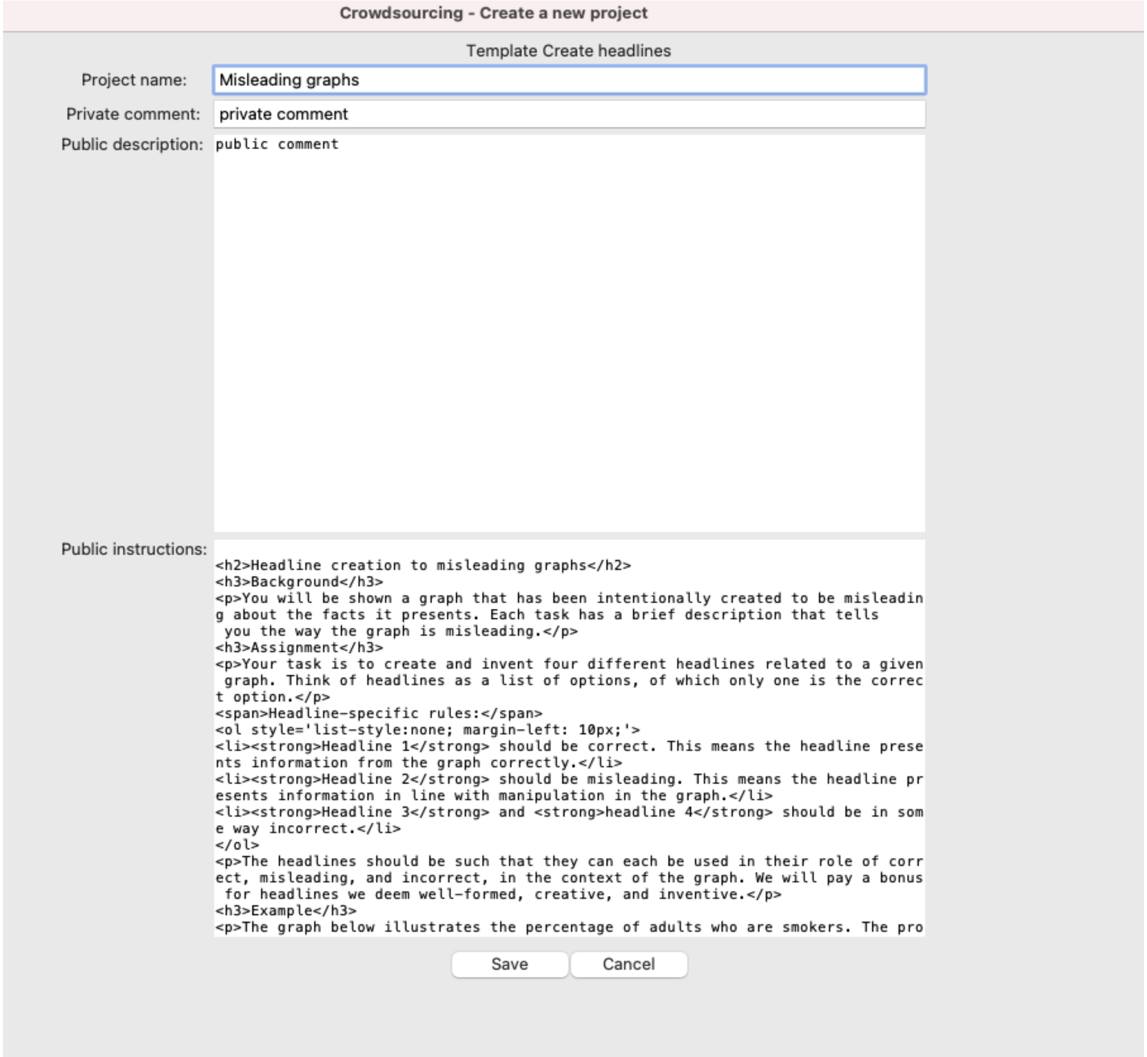


Figure 6. Tool home view.



Crowdsourcing - Create a new project

Template Create headlines

Project name: Misleading graphs

Private comment: private comment

Public description: public comment

Public instructions:

```
<h2>Headline creation to misleading graphs</h2>
<h3>Background</h3>
<p>You will be shown a graph that has been intentionally created to be misleadin
g about the facts it presents. Each task has a brief description that tells
 you the way the graph is misleading.</p>
<h3>Assignment</h3>
<p>Your task is to create and invent four different headlines related to a given
 graph. Think of headlines as a list of options, of which only one is the correc
t option.</p>
<span>Headline-specific rules:</span>
<ol style='list-style:none; margin-left: 10px;'>
<li><strong>Headline 1</strong> should be correct. This means the headline prese
nts information from the graph correctly.</li>
<li><strong>Headline 2</strong> should be misleading. This means the headline pr
esents information in line with manipulation in the graph.</li>
<li><strong>Headline 3</strong> and <strong>headline 4</strong> should be in som
e way incorrect.</li>
</ol>
<p>The headlines should be such that they can each be used in their role of corr
ect, misleading, and incorrect, in the context of the graph. We will pay a bonus
 for headlines we deem well-formed, creative, and inventive.</p>
<h3>Example</h3>
<p>The graph below illustrates the percentage of adults who are smokers. The pro
```

Figure 7. A New project creation view.

By saving the project, the view switches back to the first view (Figure 6). From this view, the user can select a previously formed project by name and switch to the project management view (Figure 8). This view offers paths through which images and task templates can be added to the project. These will then be available to each project pool. The hide project button hides the project from the tool, but it has no effect on the project in Toloka's user interface. The pools button takes you to a view where the managed pools are created and selected. When the user creates a new pool, he or she can specify a name and a task-specific reward for it.

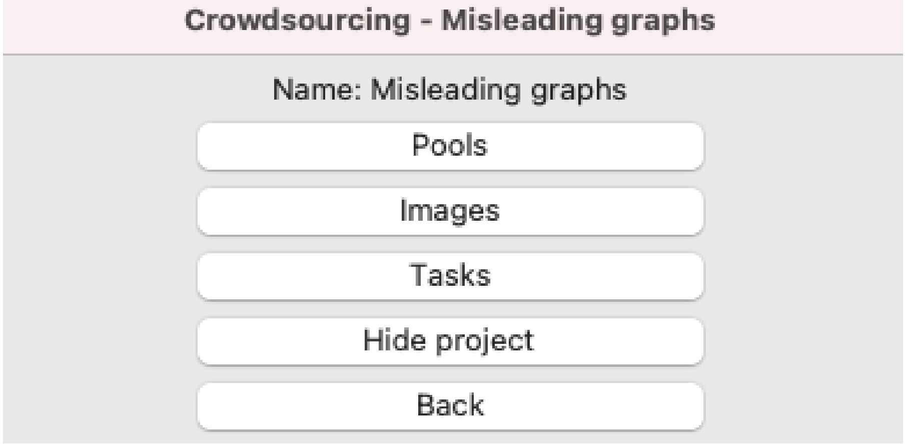
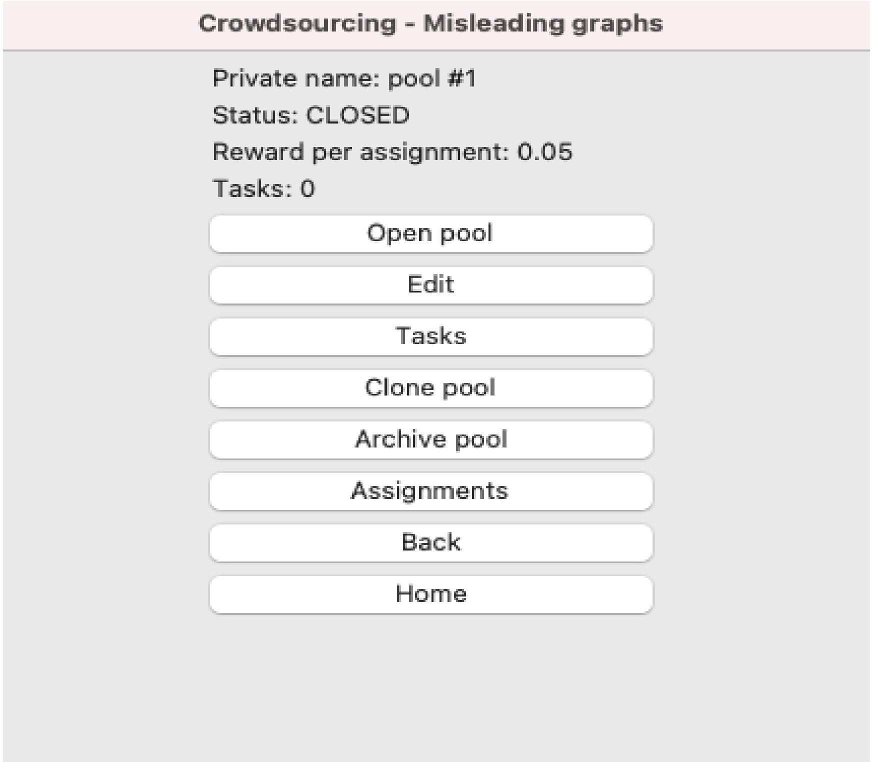


Figure 8. Project main menu view.

After selecting the pool, the user will be taken to a view where it can be managed (Figure 9). Through this view, the user can open, close, clone, and archive the pool. The pool name and task-specific reward can be changed. Tasks can be added to the pool from ready-made task templates or by making a completely new one. In addition, the completed assignments can be downloaded as a CSV file from this view. However, this function of creating a CSV file was implemented in phase 2.



Crowdsourcing - Misleading graphs

Private name: pool #1

Status: CLOSED

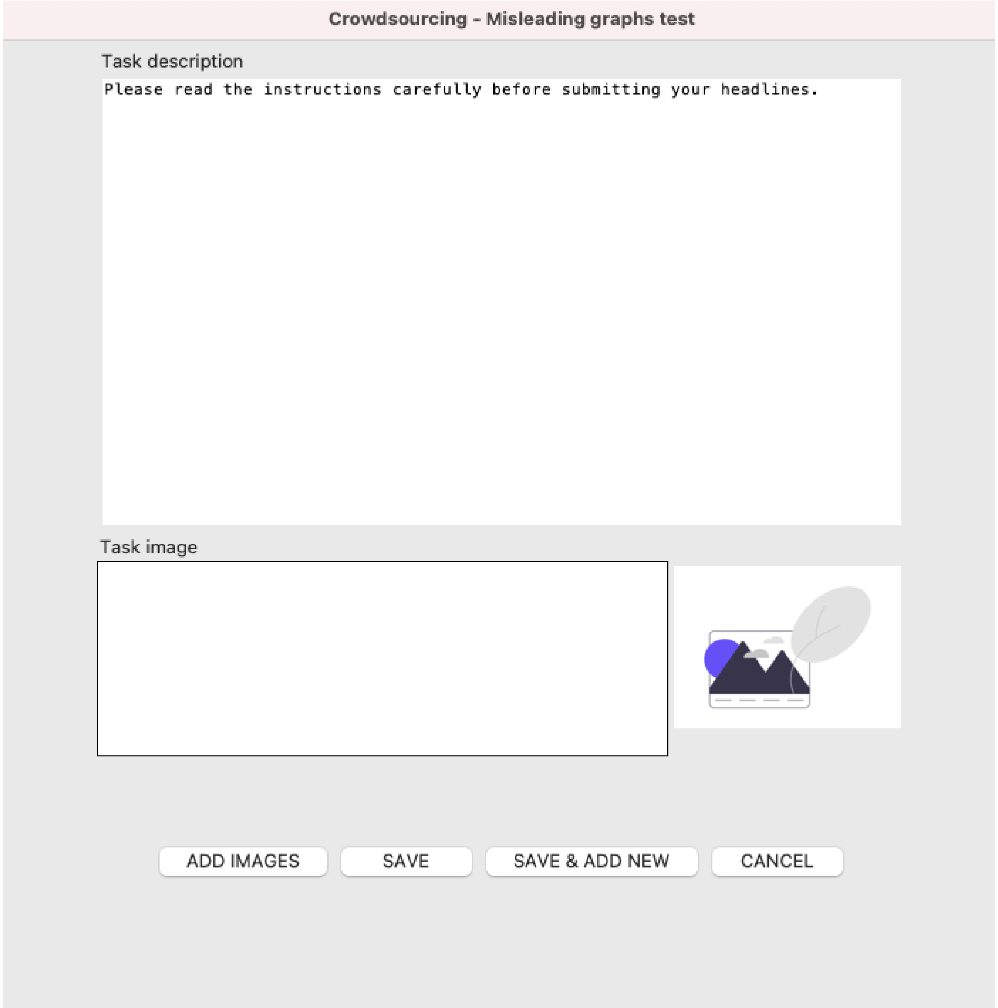
Reward per assignment: 0.05

Tasks: 0

Figure 9. Pool main menu view.

Task templates can be added in the tasks view. There is a table from the created task templates, from which the user selects the templates from which he or she wants to create the actual tasks of the pool. The actual tasks also have their own table.

In the new task creation view, the user can define the task description and select a suitable image from the list (Figure 10). If the required image is not in the list, the user can also upload it here. Each description begins with the same sentence that aims to attract the worker’s attention. This sentence is automatically filled in the description of each task.



```
Please read the instructions carefully before submitting your headlines.
```

Figure 10. Task creation view.

4.2.4 Task Design of the Toloka User Interface

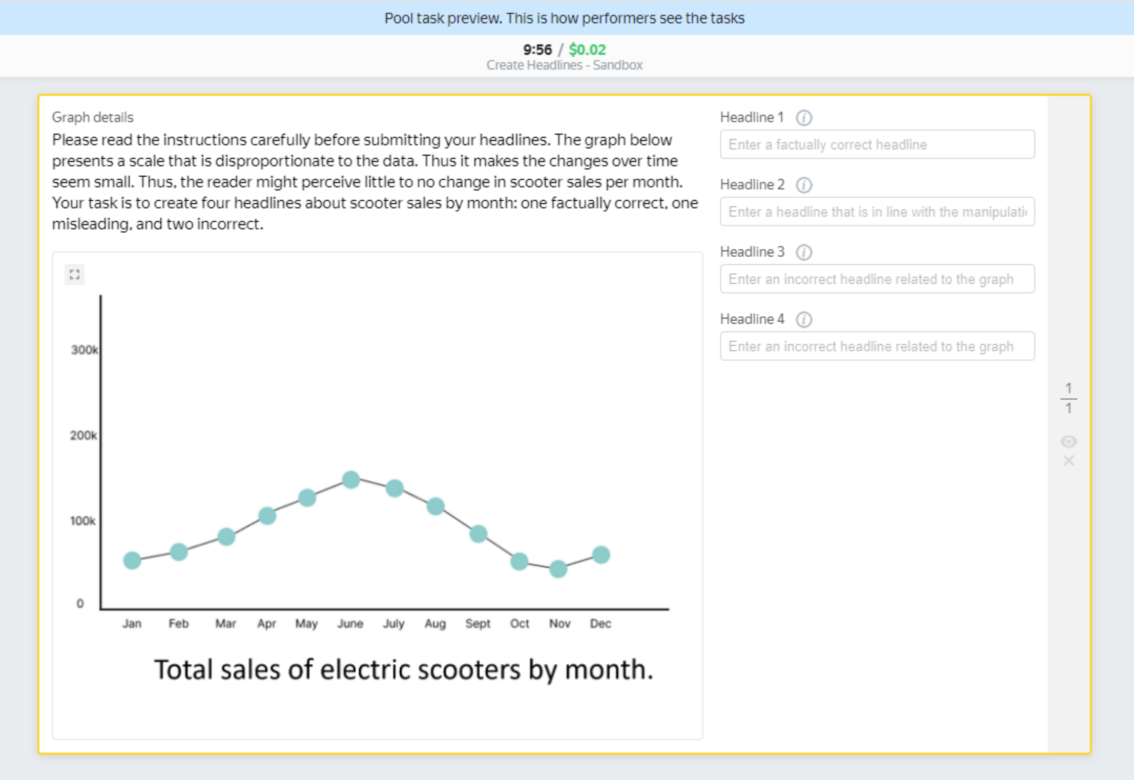
Previous research considered the design of the task as one significant factor in terms of the quality of the answers (Ramírez, 2020). The design of the tasks in this project went through a few iterations. The elements were almost the same in each iteration, but their placement and size varied. The goal was to make a view as simple as possible, where all the necessary information would be easily accessible.

The graph was one of the most central elements of the task, as the worker must interpret it so he or she could form headlines according to the assignment. Based on this, the graph was immediately placed first on the top and the description and inputs were placed below it. At this point, the graph was in its original size. With this layout, it was difficult to quickly form an overall picture, because not all elements were visible at the same time.

In the next iteration, the size of the graph was reduced so that all elements could be visible at the same time even on a small computer screen. In this iteration the graph was on the top left, the description was bottom left, and the inputs were on the top right. This layout was otherwise functional, but from the point of view of completing the task in the right order, it was not fully functional, because the graph is noted first after which the task description would have been read. The final layout was obtained, by exchanging the position of the graph and the description.

The iterations also involved some small fine-tuning. The task was to create four different headlines. For workers to create the correct type of headlines in each input, a placeholder and an info bar have been added and developed during different iterations. Also, reducing the number of tasks on a page from four to just one, to maintain the worker's focus on a specific task.

In addition to the task description, the task also had broader instructions. It gave the assignment as usual, but also an example of the same type of assignment that has been solved. This example was used to reduce uncertainty about the objectives of the task and was a bit like a practice for the actual task. Below are presented the four headline creation tasks that were used in this thesis.



Pool task preview. This is how performers see the tasks

9:56 / $0.02

Create Headlines - Sandbox

Graph details

Please read the instructions carefully before submitting your headlines. The graph below presents a scale that is disproportionate to the data. Thus it makes the changes over time seem small. Thus, the reader might perceive little to no change in scooter sales per month. Your task is to create four headlines about scooter sales by month: one factually correct, one misleading, and two incorrect.

Total sales of electric scooters by month.

Headline 1

Enter a factually correct headline

Headline 2

Enter a headline that is in line with the manipulati

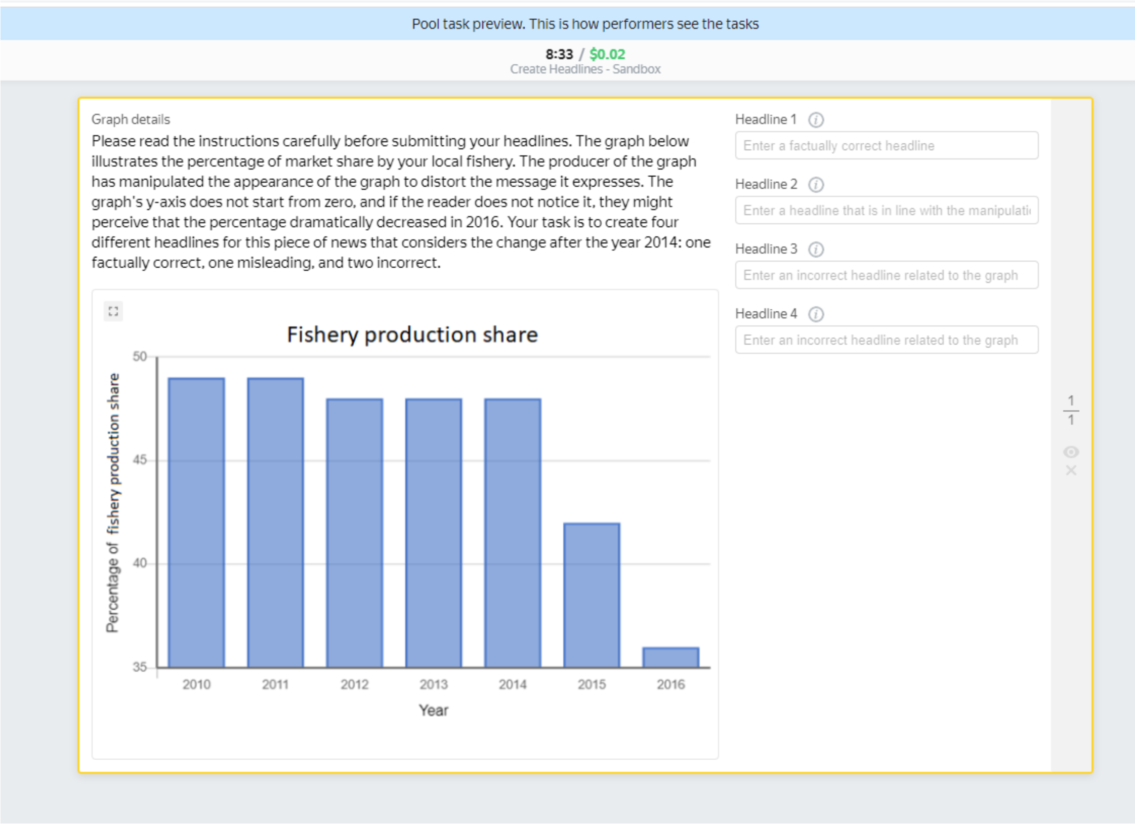
Headline 3

Enter an incorrect headline related to the graph

Headline 4

Enter an incorrect headline related to the graph

Figure 11. Headline creation task 1.



Pool task preview. This is how performers see the tasks

8:33 / $0.02

Create Headlines - Sandbox

Graph details

Please read the instructions carefully before submitting your headlines. The graph below illustrates the percentage of market share by your local fishery. The producer of the graph has manipulated the appearance of the graph to distort the message it expresses. The graph's y-axis does not start from zero, and if the reader does not notice it, they might perceive that the percentage dramatically decreased in 2016. Your task is to create four different headlines for this piece of news that considers the change after the year 2014: one factually correct, one misleading, and two incorrect.

Headline 1

Enter a factually correct headline

Headline 2

Enter a headline that is in line with the manipulati

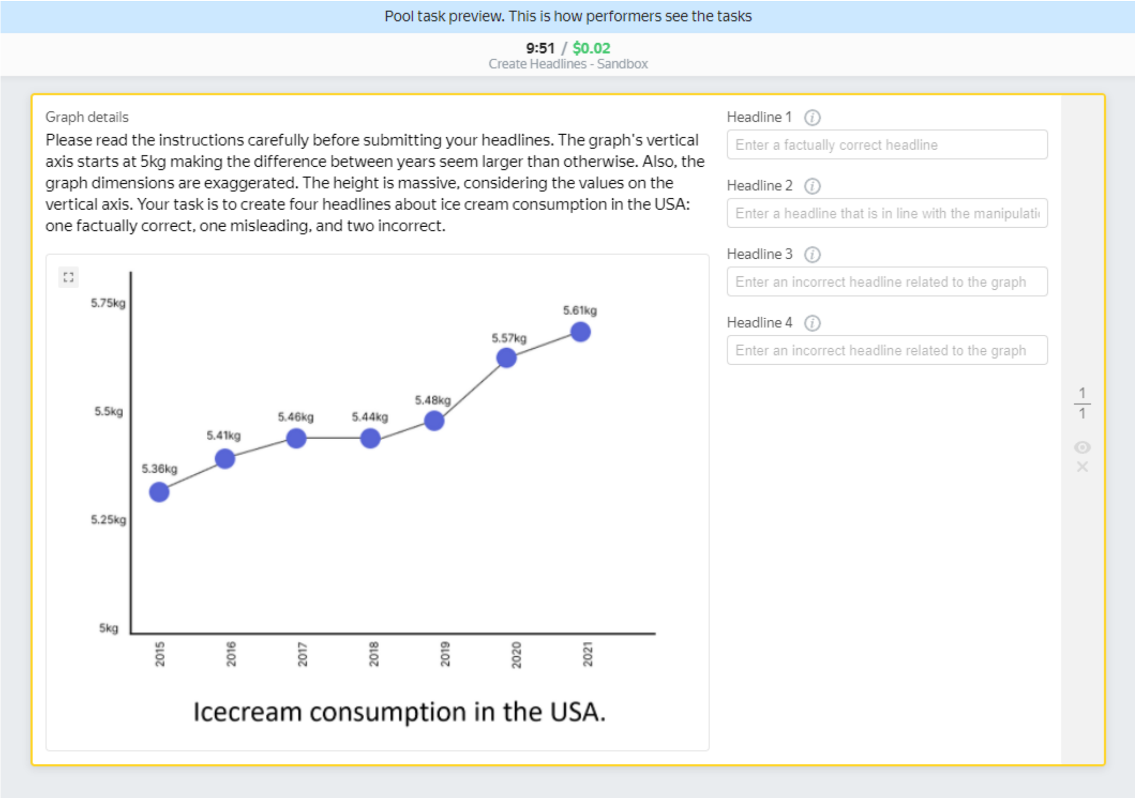
Headline 3

Enter an incorrect headline related to the graph

Headline 4

Enter an incorrect headline related to the graph

Figure 12. Headline creation task 2.



Pool task preview. This is how performers see the tasks

9:51 / $0.02

Create Headlines - Sandbox

Graph details

Please read the instructions carefully before submitting your headlines. The graph's vertical axis starts at 5kg making the difference between years seem larger than otherwise. Also, the graph dimensions are exaggerated. The height is massive, considering the values on the vertical axis. Your task is to create four headlines about ice cream consumption in the USA: one factually correct, one misleading, and two incorrect.

Headline 1

Enter a factually correct headline

Headline 2

Enter a headline that is in line with the manipulati

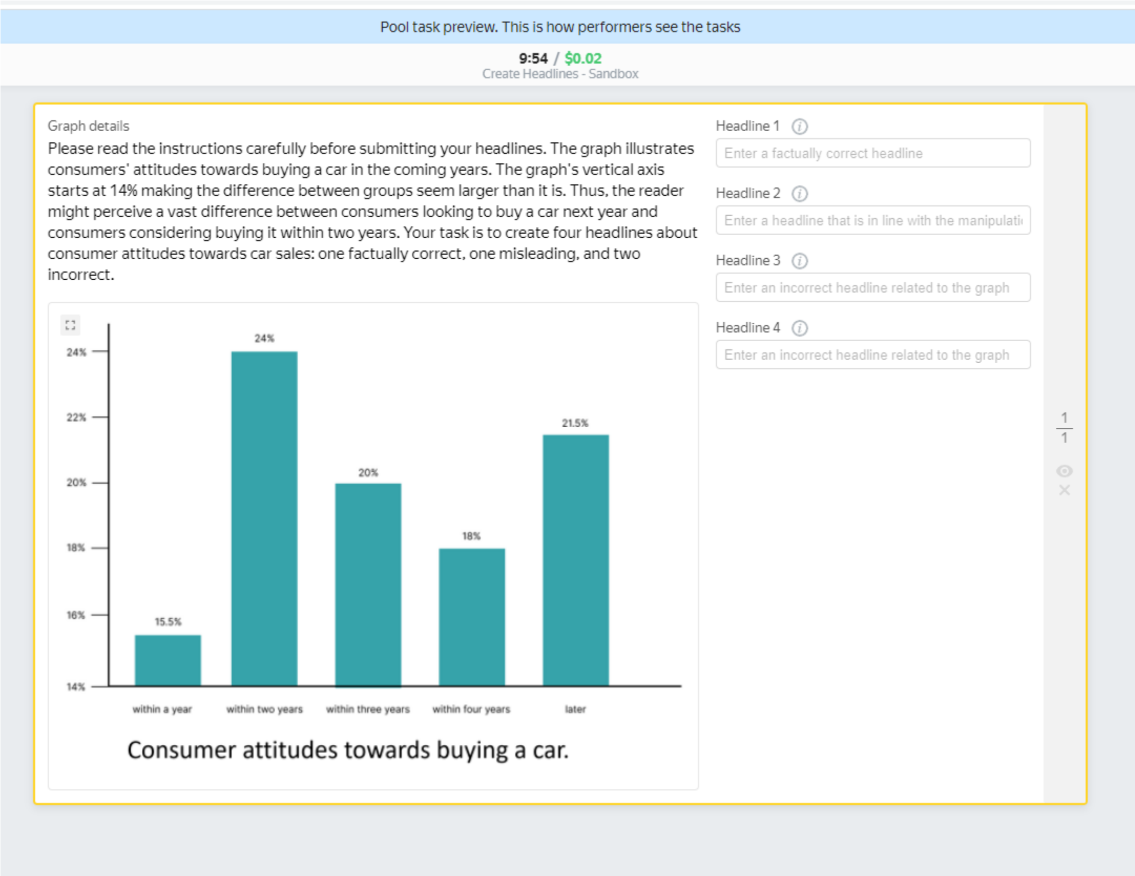
Headline 3

Enter an incorrect headline related to the graph

Headline 4

Enter an incorrect headline related to the graph

Figure 13. Headline creation task 3.



Pool task preview. This is how performers see the tasks

9:54 / $0.02

Create Headlines - Sandbox

Graph details

Please read the instructions carefully before submitting your headlines. The graph illustrates consumers' attitudes towards buying a car in the coming years. The graph's vertical axis starts at 14% making the difference between groups seem larger than it is. Thus, the reader might perceive a vast difference between consumers looking to buy a car next year and consumers considering buying it within two years. Your task is to create four headlines about consumer attitudes towards car sales: one factually correct, one misleading, and two incorrect.

Headline 1

Enter a factually correct headline

Headline 2

Enter a headline that is in line with the manipulati

Headline 3

Enter an incorrect headline related to the graph

Headline 4

Enter an incorrect headline related to the graph

Figure 14. Headline creation task 4.

4.3 Phase 2 - Headline Validation

This section presents the planning and implementation of phase 2, called headline validation. Headline validation refers to an act where ready-made headlines, which voluntary workers made in the previous phase, are reviewed, and evaluated if they correspond to the given graph and assignment. This phase is also done by voluntary workers.

4.3.1 Planning

Phase 1 was planned and implemented as an independent section before Phase two. Thus, the planning of phase 2 did not actually begin until the data from the first phase had been gathered. At this stage, problems common to both phases are considered to be solved. The layout of Toloka's user interface was also suitable for this phase and did not need major changes. The graphical user interface of the tool was wanted to be kept as similar as possible in the parts where it was possible. The means of quality control had been studied in the previous phase and they were considered suitable for this phase as well. So, planning focused on things that did not have to be dealt with in Phase 1.

According to Vuurens' (2011), data evaluation can also be done with crowdsourcing. In this way, a large number of judgments can be obtained quickly and inexpensively. Majority voting has often been used to ensure better quality judgments. In this way, it is possible to achieve a quality level that is comparable to the level of experts. With these findings in mind, validation started to be implemented.

Phase 2 was to validate the previously generated headlines, which is why it was necessary to find a way to import them. The validation task would also include the same graph and its description, on the basis on which it was initially formed. Defining validation tasks to Toloka should be as effortless and automated as far as possible.

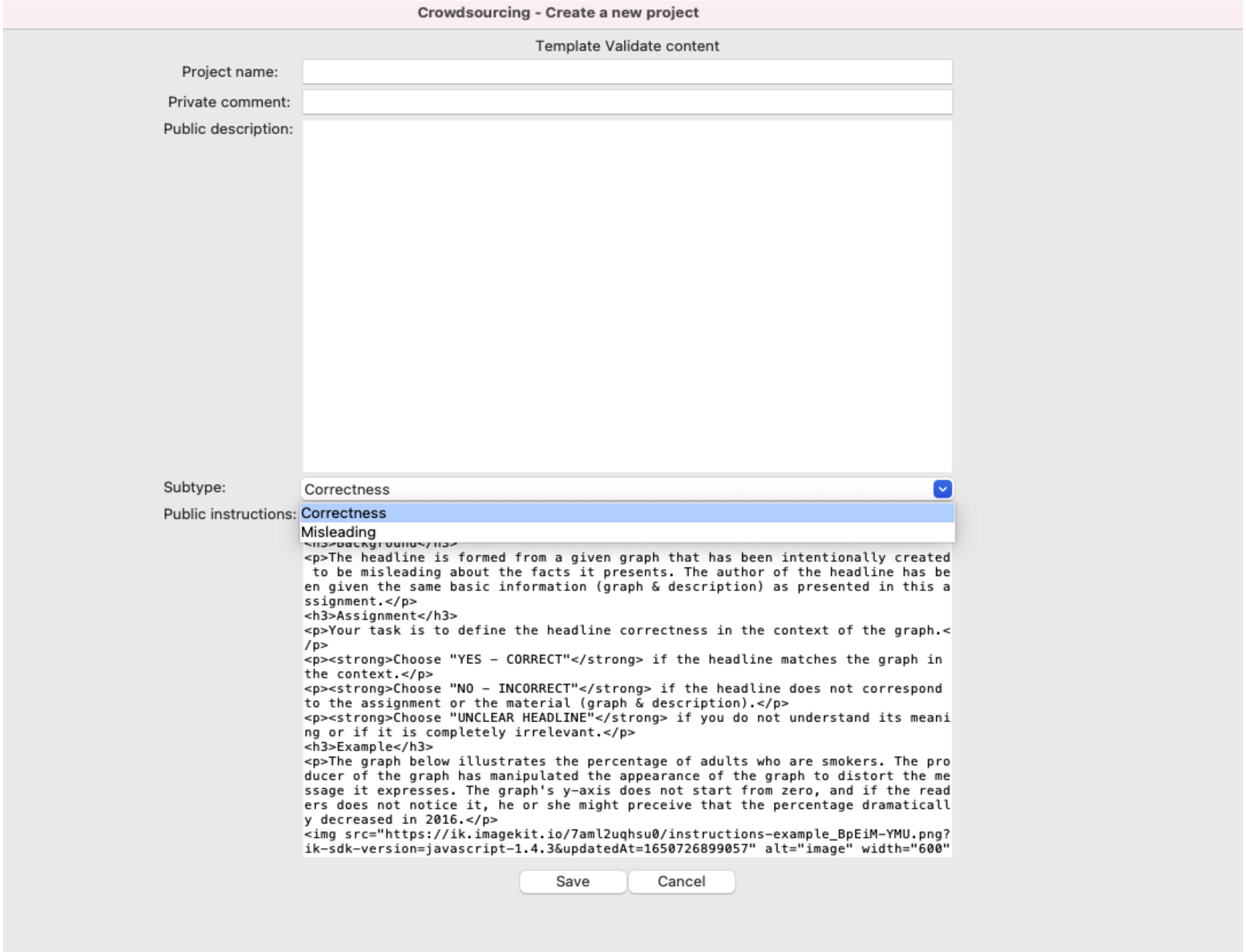
4.3.2 Implementation

The implementation of this phase was focused on the graphical user interface and functionalities of the tool. To go back a bit to the previous phase, the option to create a CSV file was added to the interface of the headline creation phase. This file contained everything needed for the task output fields to create the validation tasks. There was also a CSV generation feature in the headline validation phase, but it had a different purpose. A majority vote was to be aggregated from the responses. Aggregation was done using Toloka's Crowd-Kit Python library, where a certain kind of CSV data was wanted. Thus, a CSV file suitable for Crowd-kit was created from the responses.

The changes affecting the user interfaces were largely related to a new template that was made for the headline validation phase. Initially, there was only one model for all validation tasks, but in the end, it was wanted to distinguish correctness and misleading headlines, which is why subtypes were added to the template. More about the graphical user interface and Toloka's task user interface in the following subsections.

4.3.3 The Design of Tool

In phase 2, the intention was to use as many elements as possible from the implementation of phase 1. The creation of the new project happened almost the same way as the headline creation template. Since each template had its guidelines, and there were two of them in headline validation, subtypes were needed to define the default instructions (Figure 15). The subtypes were correctness and misleading.



```
<h3>Background</h3>
<p>The headline is formed from a given graph that has been intentionally created
 to be misleading about the facts it presents. The author of the headline has be
en given the same basic information (graph & description) as presented in this a
ssignment.</p>
<h3>Assignment</h3>
<p>Your task is to define the headline correctness in the context of the graph.<
/p>
<p><strong>Choose "YES - CORRECT"</strong> if the headline matches the graph in
the context.</p>
<p><strong>Choose "NO - INCORRECT"</strong> if the headline does not correspond
to the assignment or the material (graph & description).</p>
<p><strong>Choose "UNCLEAR HEADLINE"</strong> if you do not understand its meani
ng or if it is completely irrelevant.</p>
<h3>Example</h3>
<p>The graph below illustrates the percentage of adults who are smokers. The pro
ducer of the graph has manipulated the appearance of the graph to distort the me
ssage it expresses. The graph's y-axis does not start from zero, and if the read
ers does not notice it, he or she might preceive that the percentage dramaticall
y decreased in 2016.</p>
<img src="https://ik.imagekit.io/7aml2uqhsu0/instructions-example_BpEiM-YMU.png?
ik-sdk-version=javascript-1.4.3&updatedAt=1650726899057" alt="image" width="600"
```

Figure 15. A new project creation view.

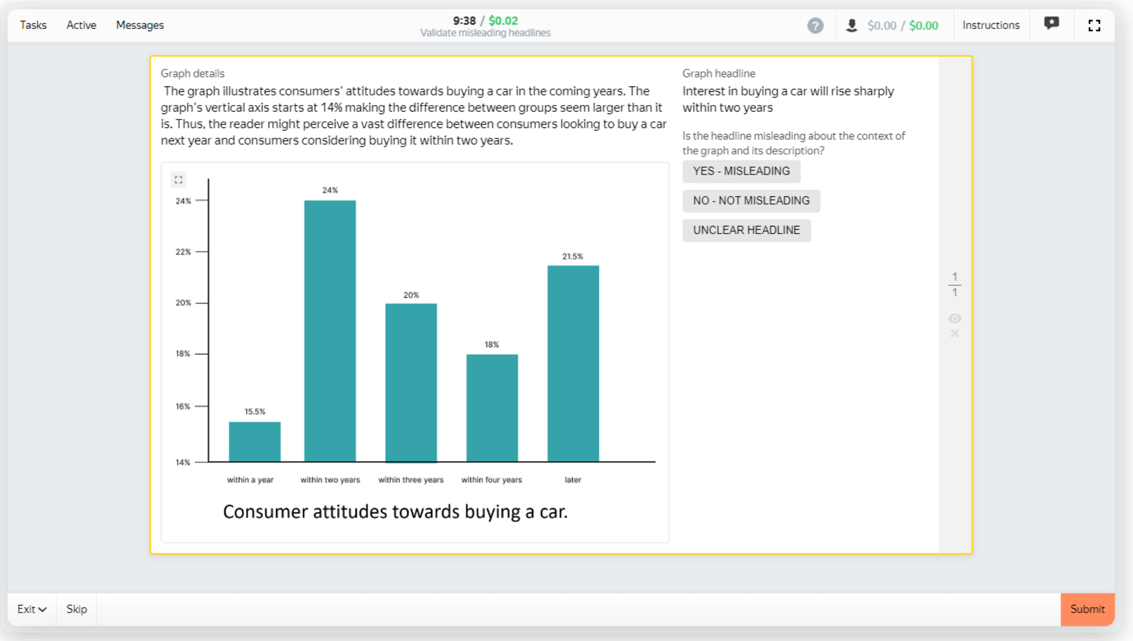
The headline validation tasks are based on the responses of the headline creation phase. The subtype influences which answers are used to form a validation task of a certain type. The subtype also changes the question-and-answer options presented in the task.

For the headline validation template, one more table was added to the tasks view, where the user could search for existing headline creation template projects. From the table, the user first selects a project and then a pool. If there are ready-made assignments in the pool, validation tasks can be created automatically from the answers to these tasks. In addition, a new feature was the creation of validation tasks from a CSV file, which can be uploaded in this view. However, this CSV file needs to be created using this tool for it to be the right type. Otherwise, the user did not see any other differences between the templates in the GUI.

4.3.4 Task Design of the Toloka User Interface

The layout of the phase 2 task corresponds largely to the phase 1 task's layout. The input fields have been replaced with three option buttons. The planning and implementation of the design for this phase task were more straightforward than the previous task design because it was possible to use the old layout so much. The description is placed at the top left and the phase 1 graph is below it. One of the headlines is placed at the top right and the option buttons are below it.

Headline validation was initially intended to be done with two alternative options: yes and no. Headline validation was divided into two types of questions: Whether the headline is misleading or whether the headline is correct. From the options "yes" could be answered if the headline was misleading or correct. However, this thesis aimed to find out whether it is possible to obtain high-quality data through crowdsourcing. In this case, the answer "no" would not be enough to adequately describe the quality of the answer. Therefore, an "unclear" option was added to distinguish low-quality answers. For the options to be sufficiently descriptive, each one of them was given an additional term as shown in Figure 16. The types of validation tasks differed from each other mainly in that they had slightly different options, titles, and instructions.



Graph details

The graph illustrates consumers' attitudes towards buying a car in the coming years. The graph's vertical axis starts at 14% making the difference between groups seem larger than it is. Thus, the reader might perceive a vast difference between consumers looking to buy a car next year and consumers considering buying it within two years.

Consumer attitudes towards buying a car.

Graph headline

Interest in buying a car will rise sharply within two years

Is the headline misleading about the context of the graph and its description?

Figure 16. Validate content task misleading layout.

4.4 Quality Control

Toloka offers a wide range of different quality control methods. From the point of view of this prototype, the focus was mainly on filtering the workers who do tasks mainly for money. Responses that are too fast indicate poor-quality responses. The worker has not been able to read the instructions, familiarize himself or herself with the graph and its misleading methods, and produce thoughtful content within the minimum time specified for the task. The minimum time was defined as 15 seconds for both headline creation and validation. The minimum time defined in the validation task is sufficient to exclude those workers who quickly click on a random option and move on to the next one for easy money. Otherwise, the minimum time should be a little longer, because especially on headline creation tasks, writing four headlines may take longer than 15 seconds, no matter what the content is. Too-fast responses were punished with a 15-day ban.

Poorly motivated workers were also limited by preventing them from skipping tasks. Skipping two tasks resulted in a 15-day ban. In the headline creation task, all four input fields were required to complete the task. The task could not be submitted without at least one four-letter word in all input fields.

A captcha was added to ensure that a human completes the tasks and not, for example, robots. The captcha had to be completed with 65 percent accuracy or the punishment was a 15-day ban.

Regarding the quality of the answers, one important aspect was the language used in the tasks. The workers must have passed an English language test to be able to perform tasks. The task assignment was in English, which is why it was important for the workers to understand and be able to use it. This could at least minimize misunderstandings or low-quality responses due to linguistic reasons.

Toloka has a speed/quality balance feature that can be used to select the most suitable workers. According to Toloka's documentation, the selection is based on a large amount of data about user behavior in the system, how other users completed your tasks, and the task itself. An assessment is made in real-time of how well each worker would perform the given task. The workers who best respond to this assessment are selected among all workers. Normally all workers have access to a task, but by setting the speed/quality balance filter, access can be given to workers with a certain reputation. The filter defines a certain percentage of workers who are given access to the task. Presumably, the lower the percentage, the higher the quality of the responses, but it takes longer to get them. In the tasks of this thesis, the speed/quality balance was initially 10 percent. This did not generate enough responses, after which it was raised to 20 percent.

5 Data Gathering

Data gathering was done using Toloka's platform. The project, pool, and tasks were entered through the tool to Toloka after which waited for the answers from the workers. A total of three projects were created during the research: one for Phase 1 and two for Phase 2. Phase 1 was held on the 18th of April 2022 and Phase 2 on the 28th of June 2022.

Each of the Phase 1 tasks produced four different headlines. The first headline had to be correct, the second misleading, and the other two incorrect. There was a total of four tasks and 30 different workers for each task. Generally, every response was accepted, except for 34 headlines that had to be rejected because they did not correspond to the assignment in any way. In rejected responses, random letters were used, or words were copied and pasted from the assignment. Accepted responses were paid the price specified in the assignment. The same tasks were performed in two different pools, one with 25 cents as the reward and the other with one dollar. In the end, there were 42 approved 25-cent assignments and 42 one-dollar assignments, so there was a total of 84 approved assignments in the first phase. With the tool, it was possible to download a CSV file of the responses and utilize it in the next phase. Below is the data-gathering workflow in phase 1 (Figure 3.).

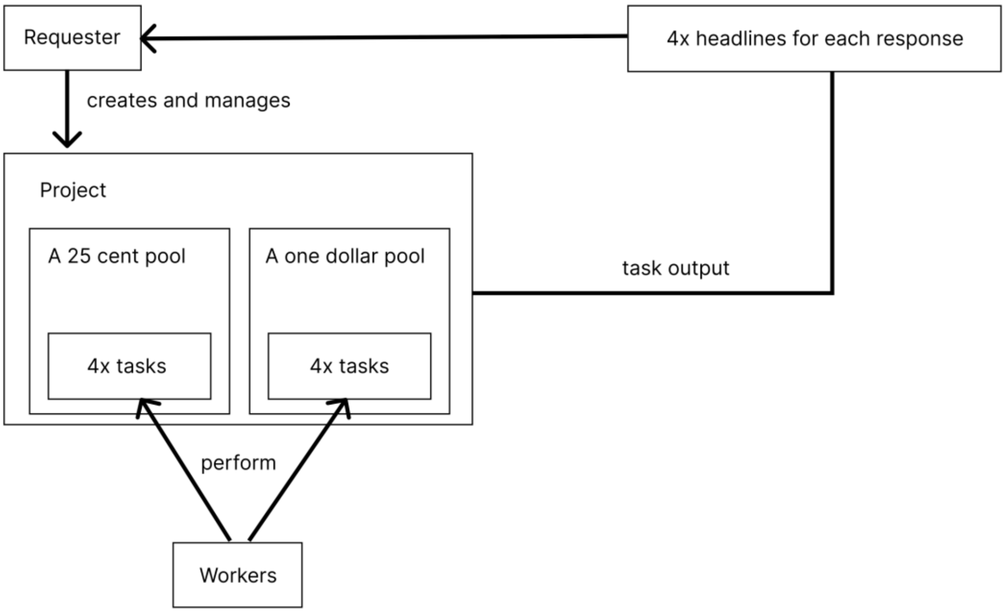


Figure 3. Phase 1 data gathering.

In Phase 2, the task was to validate the responses of the first phase. Two projects were formed for this phase. The task of the first project was to determine whether the headline is correct or incorrect and the second project was whether the headline is misleading or not. The CSV file of the responses obtained from the previous phase could be entered into the tool, which automatically generated the validation tasks. The first project used three of the four headlines of the task: correct and two incorrect. There were a total of 2550 responses to these tasks. The second project used only a misleading headline and resulted in 850 responses. All validation tasks were worth two cents.

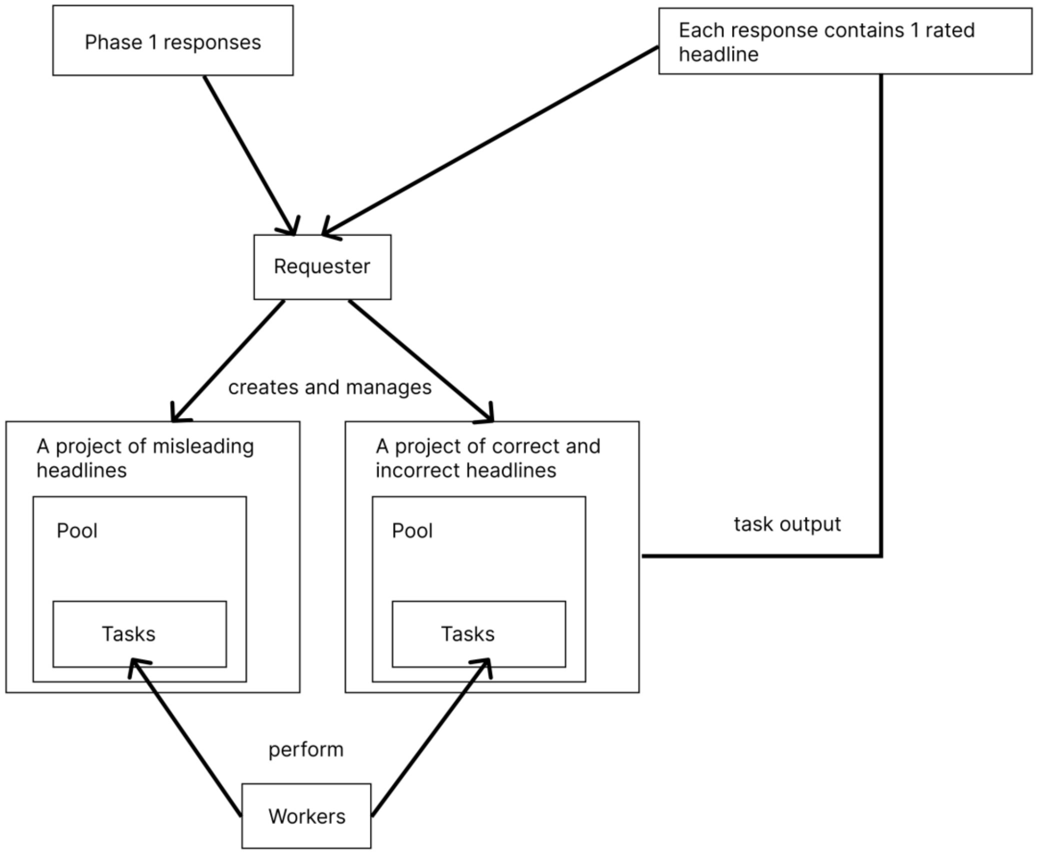

Phase 1 responses
Each response contains 1 rated headline
Requester
creates and manages
A project of misleading headlines
Pool
Tasks
A project of correct and incorrect headlines
Pool
Tasks
task output
perform
Workers

Figure 4. Phase 2 data gathering.

6 Data Analysis

This thesis used crowdsourcing for headline creation (phase 1) and validation (phase 2). In addition, there were two experts who participated in the thesis project: a Doctor of Science, age 33, male, and a Bachelor of Science, age 30, male. Between the two phases, a brief review of each of the responses of Phase 1 was performed by a requester, who was one of the two experts. In addition to managing crowdsourcing projects, the experts were responsible for evaluating the workers' responses, which is why they are called expert evaluators. At this point, the response was accepted if it was somehow reasonable text, which means, for example, that it was not just copy-pasted from the assignment or random text which means nothing. Validation tasks for phase 2 were formed from the accepted responses. One headline creation task produced four different headlines, and for each headline, a separate validation task was formed. In the validation task, the worker evaluates whether the headline corresponds to the assignment. After receiving enough responses from workers, the part of crowdsourcing ended here. In the validation phase, ten workers responded to the same task, which is why they were aggregated by majority voting. Each task thereafter had only one response based on the majority opinion. These responses were further reviewed separately by two expert evaluators. Figure 5 describes the main points of the data analysis workflow.

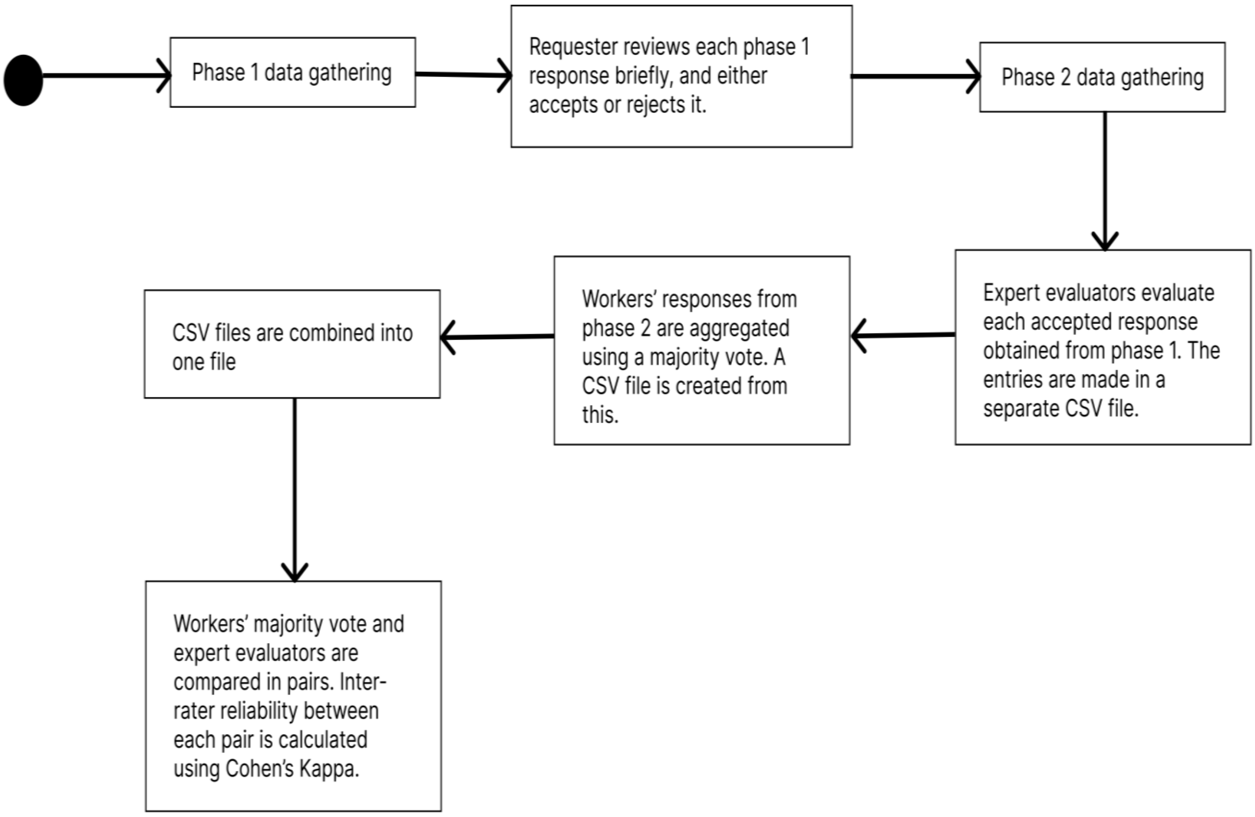

Phase 1 data gathering
Requester reviews each phase 1 response briefly, and either accepts or rejects it.
Phase 2 data gathering
Expert evaluators evaluate each accepted response obtained from phase 1. The entries are made in a separate CSV file.
Workers' responses from phase 2 are aggregated using a majority vote. A CSV file is created from this.
CSV files are combined into one file
Workers' majority vote and expert evaluators are compared in pairs. Inter-rater reliability between each pair is calculated using Cohen's Kappa.

Figure 5. Data analysis workflow.

Expert evaluators had three different options for validating the responses: yes, no, and unclear. The options were practically the same as those used by crowdsourcing workers but without the suffix. The expert evaluators went through two CSV files, and each row contained information about the assignment and the headline created by a worker. The first CSV file contained responses for the misleading theme and the second contained responses for the headline correctness theme. The expert evaluators compared the assignment and related graph to the headline and selected the most suitable option. The evaluator's own responses were listed in a CSV file next to the headline, after which all the responses were exported to one CSV file. This made it possible to compare the responses with each other.

According to Sun (2011), "inter-rater reliability refers to the consistency of ratings given by different raters to the same subject". Cohen's Kappa is well-suited for measuring inter-rater reliability in this kind of research. To quote Sun (2011) even more "Cohen's Kappa determines whether the degree of agreement between two raters is higher than would be expected by chance". The measurement of Cohen's Kappa is done between two raters. However, in this research, there were three of them: two experts and the majority. In this case, Cohen's Kappa was used to measure the degree of agreement between each possible pair. The validation task had three answer options, but Cohen's Kappa only has yes and no. For this reason, answers per headline were removed from the comparison, if even one rater had used an unclear option.

Each pair of raters was reviewed separately. Each rating was assigned a score of 1 for agreement and 0 for disagreement. In Cohen's Kappa Hypothetical probability of chance agreement (Pe) is calculated using four different values, which can be, for example, A, B, C, and D.

The value **A** represents the number of instances where both raters gave a yes answer, which also means that the raters agree.

The value **B** represents the number of instances where rater 2 gave a no answer when rater 1 gave a yes answer, in which case they disagree.

The value **C** represents the number of instances where rater 1 gave a no answer when rater 2 gave a yes answer, in which case they disagree.

The value **D** represents the number of instances where both raters gave a no answer, which means that the raters agree.

P(yes) = (A + B / A + B + C + D) * (A + C / A + B + C + D)

P(no) = (C + D / A + B + C + D) * (B + D / A + B + C + D)

Pe = P(yes) + P(no)

Next Relative observed agreement among raters (Po) is calculated as follows:

Po = number in agreement/total number of answers

According to Sun (2011), the formulas can finally be used in Cohen's Kappa formula presented in Eq. 1.:

$$kappa(\mathcal{K}) = \frac{Po-Pe}{1-Pe} \quad (1)$$

The results of Cohen's Kappa can then be evaluated using Table 1 below. The table shows the level of agreement between the raters.

Table 1. Ho et al., 2019. Interpretation of Cohen's Kappa value.

Cohen's Kappa value	Interpretation of Cohen's Kappa value
≤ 0	No agreement
0.1-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-0.99	Near-perfect agreement
1	Perfect agreement

7 Results

This section goes through the results obtained during the research. After the first phase, 84 headline creation assignments were accepted and 34 were rejected. Validation tasks were formed from the output of the accepted tasks that is, a validation task was created for each headline, in which case four validation tasks were formed for one headline creation assignment. Workers produced a total of 3400 responses in the headline validation phase, of which 850 were about misleading and 2550 were about correctness. The validation responses were aggregated using a majority vote, so for the same validation task, only the majority response remained as the correct answer. Eventually, there were a total of 84 validation responses on the misleading headline topic and 248 validation responses on the headline correctness topic.

Table 2. The distribution of the responses of expert evaluators and majority vote, where the topic was the correctness of the headlines.

Correctness of the headlines – the distribution of the responses

	"YES – correct"	"NO – incorrect"	"UNCLEAR"
Majority vote	217	31	0
Expert evaluator 1	104	131	13
Expert evaluator 2	58	167	23
Total	379	329	36

Table 3. The distribution of the responses of expert evaluators and majority vote, where the topic was misleading headlines.

Misleading headlines – the distribution of the responses

	”YES – misleading”	”NO – not misleading”	”UNCLEAR”
Majority vote	29	52	3
Expert evaluator 1	33	44	7
Expert evaluator 2	31	40	13
Total	93	136	23

The expert evaluators went through the headlines and gave their responses to each headline. Cohen's Kappa accepts only two response options, so headlines that received at least one "unclear" response were removed from the comparison. Because of this, 13 responses were removed from the misleading topic and 26 from the correctness topic. So, in the end, a total of 71 misleading topic responses and 222 correctness topic responses were taken into comparison. Tables 2. and 3. show the distribution of the responses of each rater.

The following values are obtained between the raters by entering the collected numbers and calculating according to the formulas introduced in the data analysis chapter. Collected numbers refer to numbers 1 "agreement" and 0 "disagreement" between the raters.

Table 4. Cohen's Kappa values between raters and the majority vote, where the topic was the correctness of the headlines.

Cohen's Kappa value - Correctness

VS.	Expert evaluator 1	Expert evaluator 2	Majority vote
Expert evaluator 1		0.59	0.15
Expert evaluator 2	0.59		0.07
Majority vote	0.15	0.07	

The values between the raters can be interpreted with the help of Table 1 which indicates the level of agreement. Table 4 shows a moderate agreement between the expert evaluators and substantial agreement is not too far off either. The result is not exactly bad, but the evaluators still have some differing opinions. When moving on to compare the expert evaluators with the majority vote, the differences increased considerably. At its lowest point, the level of agreement was only 0,07 and at its highest, only 0.15. This means that even at best there was only a slight agreement between the expert evaluator and the majority vote. That is, the workers' responses differed from the experts' responses almost completely.

Table 5. Cohen's Kappa values between raters and the majority vote, whether the graph was misleading.

Cohen's Kappa value - Misleading

VS.	Expert evaluator 1	Expert evaluator 2	Majority vote
Expert evaluator 1		0.405	0.056
Expert evaluator 2	0.405		0.089
Majority vote	0.056	0.089	

If there was a lot of disagreement on the first topic, there was even more on the validation of the misleading topic. The disagreement between the expert evaluators and the majority did not increase significantly. In fact, for expert evaluator 2, the agreement even increased slightly. However, the biggest difference occurred between the expert evaluators. They agree fairly well on the topic of correctness, but on the misleading topic, agreement dropped from moderate agreement to fair agreement.

Table 6. shows how the reward has affected the agreement. If the headline is rated as "UNCLEAR", it has only been considered in the last column. An agreement can be either a "NO" or a "YES" answer. All the table headlines should have been rated as "YES" if they were formed correctly, that is, misleading. Therefore, the table also shows the proportion of “YES” ratings out of the total number where the evaluators agree.

Table 6. The effect of reward.

Reward per headline creation task – Misleading headline

	“YES” / expert evaluators 1 & 2 agree	“YES” / all agree	Expert evaluators 1 & 2 disagree	At least one was rated as unclear
25 cents	9 / 26	4 / 16	8	10
Dollar	11 / 22	4 / 13	15	3
Total	20 / 48	8 / 29	23	13

8 Discussion

This chapter discusses the research and its findings. The research questions will be reviewed, and at the end, there are limitations and future work.

8.1 Summary

During this thesis, two interfaces were designed for two different user groups. The prototype of the tool was intended for the requesters and the UIs of the Toloka tasks for the workers. The prototype was designed in such a way that it will be used by people involved in the development of the critical reading comprehension game. Their background is already partly known, and it is possible to get more information if necessary, and most importantly they are motivated towards using the prototype. In terms of designing the prototype, it was enough that the UI was easy to use.

The user interface design for Toloka's task was completely different because tasks will be done by people who are unknown. The basic idea behind the design of the UI was that it would encourage workers to perform tasks as well as possible. The idea was based on previous studies that have shown that the right kind of task design and incentives can improve workers' motivations, and on the other hand, the wrong kind can weaken motivation. (Feng, 2018; Deng, 2016; Daniel, 2018). According to Deng (2016), the most negative emotions were evoked by complex tasks and unfair compensation. It is important for the worker to be able to choose the tasks that suit him or her, in which case the instructions and timelines must be clear and well thought out.

The Toloka task's UI went through different iterations, after which it made its final form as clear and descriptive as possible because according to Daniel (2018), the task description and user interface have a direct impact on workers' performance. Among the features of the UI, especially user-friendliness and understandability affect the attitude of workers. In the instructions for the task, the purpose was described and a concrete example of the type of responses was to be obtained. The choices related to the UI are based on the fact that they would reduce the ambiguity related to the task and the misleading topic. As Daniel (2018) states, the worker needs to understand the assignment correctly to get the right kind of data.

The data collection of both phases took one day, so in total it took only two days to collect the data. There were only four headline creation tasks, but there were many contributors to one task. Even after removing completely unusable headlines, there were 336 headlines left. After this, the validation tasks were performed, which produced a total of 3400 responses. This all happened in two days without recruiting anybody. Initializing tasks, creating projects and quality control certainly took their own time, but the time spent on them decreases over time when a sufficiently good level in the project and task definitions and quality control has been reached.

This sample is not yet enough for the content of a critical reading comprehension game, but headlines on, for example, 100 graphs would be at least a good start. If these tasks are responded to in the same way as these four tasks, 2094 headlines are obtained after filtering the unusable responses. In crowdsourcing, a large number of people around the world perform these tasks at the same time for a small reward.

8.2 R1: Can crowd workers create a variety of news headlines based on misleading graphs?

According to Garcia-Molina et al. (2016), humans are better suited to image or natural language interpretation tasks than a computer. Both are included in the tasks of this thesis, but in addition to these, there is a subject, misleading graphs, that is as presented by Ramly (2021) also difficult for humans to interpret.

During the phase 1 crowdsourcing project, where the task was to create headlines, a total of 118 responses were received from workers. The crowdsourcing project requester reviewed each response quickly. This was done so that the completely invalid responses could be removed before the validation phase. A total of 34 responses were rejected, and this was because the responses were unclear text, or the responses were copied and pasted from the task description. At this point, 28,8 percent of the responses had been rejected.

The responses to the headline creation tasks contained four headlines, resulting in a total of 336 validation tasks. After two expert evaluators and a majority vote of workers, 39 headlines were marked, by at least one rater, as unclear. Unclear means a low-quality response that does not correspond to the assignment. 11,6 percent of the responses were of poor quality by this measure. The remaining responses can be assumed to have somewhat followed the assignment and provided meaningful content. The quality of the responses is discussed more in the following research question.

8.3 R2: How to determine that the material produced is usable and correct for the critical reading comprehension game?

According to Vuurens (2011) evaluating data by crowdsourcing is not that simple due to dishonest workers. There may be spammers among the workers who are not even trying to respond correctly. This, of course, has a direct effect on data quality. The fact that makes it even more challenging is that both types of tasks, headline creation, and validation, are done by crowdsourcing. Thus, both tasks must be evaluated with the idea that they could have been done by a dishonest worker.

The responses to the headline creation tasks were mostly accepted. If there were immediately noticeable poor-quality responses, they were rejected, but no further assessment of their quality was made at this stage. After this, each approved headline was formed into its own validation task, where the workers could give their assessment of whether the headline corresponds to the assignment. The responses were then aggregated using a majority vote. The purpose of the majority vote was to get the general opinion of the workers about the quality of the response to the previous task. According to Yung (2014), voting does not in itself improve the quality of the answers, but it can distinguish between good and bad answers. Voting was used for exactly this purpose: to get an answer as to whether the response was good or bad.

The majority vote does not guarantee that the ratings are good and carefully considered. In the validation tasks, there were three different answer options to choose from, so, if wanted to, the tasks could be done very quickly by clicking on a random option. Therefore, two expert evaluators gave their own ratings, which were compared separately to each other, and the workers' ratings were based on the majority vote. The ratings of each pair were compared using Cohen's Kappa. The result was that the ratings of both expert evaluators hardly agreed with the majority vote. The same result was repeated in both topics, whether the headline was correct or incorrect, or whether the headline was misleading or not. When the expert evaluators were compared with each other, they agree fairly on the headline correctness part, but agreement decreased slightly on the misleading part.

A noteworthy point is that the workers and expert evaluators answered the tasks differently. If the worker did the task as planned, he or she went through the instructions, the graph description, the graph itself, and the headline based on it. All this could be seen in the same view. The expert evaluators did not make ratings through the Toloka user interface but went through a CSV file from top to bottom, going through each headline created. Own entries were made to the file. As a result of going through such a large list, it is possible that the concentration has been lost at some point. In addition, it is possible that the topics have been mixed up, in which case the ratings have been made based on the instructions of another task.

When looking at the workers' responses, there were 248 responses after the majority vote about the correctness of the headlines. There were 217 "yes - correct", 31 "no -incorrect", and 0 "unclear" responses. The number of "yes" responses was high considering that two-thirds of the headlines should have been incorrect according to the instructions for headline creation tasks. The ratings of expert evaluator 1 were 104 "yes – correct", 131 "no – incorrect", and 13 "unclear". In this case, the number of "yes" responses was in better proportion. The expert evaluator 2 was even more critical with 58 "yes – correct", 167 "no – incorrect", and 23 "unclear" responses. Here, the proportion of "no" responses was correct, but the proportion of "yes" responses was slightly lower than expected, which is due to the high number of "unclear" responses.

When looking at whether the headlines were misleading or not, each headline was intended to be misleading in the context of the graph and its description. Misleading means were introduced to the workers in creating and validating headlines. By default, if the headlines meet the requirements, all responses in this type of task should be" yes – misleading". There was a total of 84 responses for each rater. The majority vote had 29 "yes – misleading", 52 "no – not misleading", and 3 "unclear" responses. Evaluator 1 had 33 "yes", 44 "no", and 7 "unclear" responses. Evaluator 2 had 31 "yes", 40 "no", and 13 "unclear" responses. All raters had almost the same number of "yes" responses, but this cannot be considered a good result, because their percentage was low. At best "yes" responses were 39 percent of all responses from a single rater, which is very low when 100 percent is expected. In the responses to this topic, there were relatively more "unclear" responses than to the headline correctness topic. 15 percent at most is quite a lot compared to another topic's 9 percent.

Based on the statistics, there seemed to be many low-quality responses in both task types. Poor-quality responses clearly related to crowdsourcing can be detected, especially in the task of creating headlines, where the headlines do not in any way correspond to the assignment. There were also such low-quality responses in the headline creation tasks, for which it is impossible to know the exact reason why they were formed that way. Such were, for example, the responses that followed the context of the graph were misleading but not the way presented in the assignment. Low-quality responses are not necessarily associated only with dishonest workers, but according to Vuurens (2011), even an ethical worker can produce a low-quality response. Even though an ethical worker tries to follow the instructions and perform as well as he or she can, it is possible that the worker misinterprets the task or is not capable enough for the task. (Vuurens, 2011.)

The task of creating headlines was rewarded with either 25 cents or a dollar. After validation, the reward had no effect on how well the raters agreed or whether the response was in accordance with the assignment. The only clear difference between the rewards was that in the lower reward task, the response was more likely to be considered unclear. Of all responses that were found to be unclear, 73 percent about the correctness of headlines and 77 percent about misleading headlines, came from 25 cent tasks. In addition, there were few completely identical responses or nearly identical responses, but they were more likely to occur in tasks with a lower reward.

Based on the expert evaluators' validation, the workers had the most difficulty in forming a misleading headline of the four headline options. There was also less agreement between expert evaluators' ratings, suggesting a different interpretation. According to the majority vote of the workers, only a good third of the headlines were properly misleading. This is in line with previous research as Ramly (2021) stated that people have difficulty interpreting misleading graphs even if they are aware that the graph is misleading. Although all four headlines were formed from the same graph, the correct and incorrect options were easier for workers to understand than the misleading ones. This conclusion is supported by the fact that the expert evaluators' ratio of "yes" and "no" ratings was nearly correct, and their agreement was at a moderately good level when the correct and incorrect headlines were validated.

8.4 R3: How to integrate crowdsourcing project management into a part of a critical reading comprehension game's content creation by utilizing a crowdsourcing platform's API.

In this thesis, the content created with the prototype does not cover the entire content production of the critical reading comprehension game and is limited to headline creation only. The headlines were supposed to be formed by crowdsourcing, for which the Toloka platform was chosen. Crowdsourcing projects should be mainly managed from the prototype's own graphical user interface by utilizing Toloka's API and not so much its web interface. Crowdsourcing is included as a part of the content creation workflow and should no longer be a separate step of its own.

Simple graphical user interfaces can be made relatively quickly with Tkinter. Considering that the prototype made in this thesis did not have any design-related requirements, it seemed a suitable choice. This was largely a good choice until there were more requirements for the tool and more features wanted that were more complex than the original. Some other solutions than Tkinter would have been suitable for implementing those complex features, so it was decided to abandon these features for this prototype. The next version could use, for example, a web-based implementation.

However, creating and managing crowdsourcing projects is possible with this prototype, which was one of the requirements. A faster and more user-friendly way was made to add and handle tasks and images. The images no longer had to be stored in a different place, but it was possible to define the entire assignment from start to finish in one place.

From the efficacy perspective, the goal was reached in the implementation of the GUI. The prototype can be used to reduce the intermediate steps of project definition and the number of clicks when creating tasks. Review and approval of responses were omitted from this implementation, and these steps had to be done in Toloka's user interface. However, with further development of the current features and the implementation of the missing features, it would be possible to reach an even more efficient result.

8.5 Limitations

Although previous research recommended training workers before actual tasks it was still not used in this research (Le, 2010). The lack of training meant that the workers relied only on instructions, which advised what type of answers are wanted by a concrete example of how to perform a similar task. Unlike the training tasks, this was easy to skip, and no data could be collected about the workers before the tasks.

Training is one way by which feedback can be given to the workers even before the actual tasks are performed. There is no single correct answer to the tasks created in this research, which is why the training tasks could not be implemented in the same way as the actual tasks and this was also one of the reasons why the training was left unimplemented. Another reason was the desire to get versatile responses because it is possible to create the right kind of headlines in several ways. Teaching workers to respond in a certain way could have led to a narrowing of diversity. If it had been decided to implement the training tasks, and as they require a correct answer, they could have been, for example, choosing from the answer options. This could have influenced the quality of the responses because it is also one way to detect spammers. (Vuurens, 2011). Workers are also not paid for training tasks, which can reduce less motivated workers. In addition, it is possible to allow workers to do actual tasks only after they have completed the training task set successfully enough. Mandatory training tasks would have been useful, especially for the validation tasks in mind, as these tasks contained ready-made options that were easy to click to earn money without much effort. With training tasks, it would have been possible to at least reduce spammers and poorly performing workers.

Toloka collects information on workers' performance, including the time spent on completing the task. The quality of the responses could also have been examined using this metric, as the data exists and was available. The interpretation of the graph and its misleading methods requires some time, in which case, in the validation tasks, more weight could have been given to those responses that took at least a certain amount of time to complete. In addition, time spent is yet another way to detect spammers. (Vuurens, 2011).

The completion of the task could have been made more meaningful for the workers to increase the workers' motivation. Previous research shows that workers are particularly motivated when they can make contributions to scientific research and societal initiatives. (Deng, 2016) A better expression of what the responses will be used for later could have been added to the task descriptions.

8.6 Future Work

Crowdsourcing and matters related to its management such as quality control are broad concepts. In this thesis, some compromises and limitations had to make so that the project would not become too large. In the future, crowdsourcing could be studied in such a way that it produces data from both well-formed graphs and misleading graphs. In this way, the effect of a misleading graph can be better separated from the factors of crowdsourcing. In addition, quality control needs more research. Many existing methods of quality control were excluded from this research. What kind of responses would have been obtained, for example, after the training tasks?

9 Conclusion

In this thesis, a prototype of a tool was made to manage crowdsourcing projects by utilizing Toloka and its API. Crowdsourcing was used to create and validate headlines that are related to misleading graphs. The validation responses were aggregated by majority vote. In addition, two expert evaluators gave their ratings to the headlines, and finally, the rater's ratings were compared with each other using Cohen's kappa.

Analyzing the data was complex because there were two diverse aspects, crowdsourcing, and misleading graphs, that can affect the quality of data, to be considered. Workers' response quality can be considered poor if the responses are compared with expert evaluators' responses and if it is assumed that the headlines were in accordance with the assignment. In evaluating the correctness of headlines, most responses were "yes" although two-thirds should have been "no" responses. Expert evaluators' ratings were closer to the expected outcome, but they did not completely agree with each other. Neither expert evaluator agrees with the majority vote. In evaluating misleading headlines, most responses of the majority vote were "no" even though there should have been only "yes" responses in these tasks if the formed headlines had followed the assignment. However, most of the expert evaluators' responses were also "no". Still, the expert evaluators did not agree with the majority vote, suggesting that the workers also performed poorly on this as well. Also, expert evaluators agree less, indicating the differences and difficulty in interpreting the misleading graphs and headlines related to it.

The results of this thesis show that it is possible to efficiently obtain data by crowdsourcing. This is in line with previous research. The results also show that ensuring data quality is a major challenge. Quality control is a significant factor when trying to remove poorly performing workers. Selection tasks are particularly difficult, as it can be hard to identify poorly performing workers because they do not produce anything unique and may occasionally respond correctly. In the future, tasks like this will require better ways to select workers and detect spammers.

In the task of creating headlines, the reward seemed to have only a small effect on quality as it mainly reduced the number of responses that did not correspond in any way to the assignment. Otherwise, no difference was noticed in the quality of the responses. Most of the created headlines were reasonably good as they mainly correspond to the assignment. Graphs and related topics appeared in the responses, however, interpretations formed based on them distinguished responses from correctly and incorrectly formed headlines. This thesis shows that it is possible to improve the existing crowdsourcing processes' efficiency by integrating them into part of other implementations, in this case in the graphical user interface, which utilizes existing crowdsourcing platform functions along with their own processes. When a crowdsourcing project is well-defined and tested, it can produce meaningful data.

References

Baba, Y., & Kashima, H. (2013, August). Statistical quality estimation for general crowdsourcing tasks. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 554-562).

Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., & Allahbakhsh, M. (2018). Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, *51*(1), 1-40.

Davis, A. J. (1999). Bad graphs, good lessons. *ACM SIGGRAPH Computer Graphics*, *33*(3), 35-38.

Deng, X., Joshi, K. D., & Galliers, R. D. (2016). The Duality of Empowerment and Marginalization in Microtask Crowdsourcing. *Mis Quarterly*, *40*(2), 279-302.

Feng, Y., Ye, H. J., Yu, Y., Yang, C., & Cui, T. (2018). Gamification artifacts and crowdsourcing participation: Examining the mediating role of intrinsic motivations. *Computers in Human Behavior*, *81*, 124-136.

Frees, E. W., & Miller, R. B. (1998). Designing effective graphs. *North American Actuarial Journal*, *2*(2), 53-70.

Garcia-Molina, H., Joglekar, M., Marcus, A., Parameswaran, A., & Verroios, V. (2016). Challenges in data crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering*, *28*(4), 901-911.

Garzón-Guerrero, J. A., Valenzuela, S., & Batanero, C. (2020). STATISTICAL SENSE AND GRAPHS IN THE COVID ERA. *Financial Times*.

Geiger, D., & Schader, M. (2014). Personalized task recommendation in crowdsourcing information systems—Current state of the art. *Decision Support Systems*, *65*, 3-16.

Hettiachchi, D., Berkel, N. V., Hosio, S., Kostakos, V., & Goncalves, J. (2019, September). Effect of cognitive abilities on crowdsourcing task performance. In *IFIP Conference on Human-Computer Interaction* (pp. 442-464). Springer, Cham.

Hettiachchi, D., Van Berkel, N., Kostakos, V., & Goncalves, J. (2020). CrowdCog: A Cognitive skill based system for heterogeneous task assignment and recommendation in crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1-22.

Hettiachchi Mudiyanselage, D. E. (2021). Task assignment using worker cognitive ability and context to improve data quality in crowdsourcing (Doctoral dissertation).

Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, *14*(6), 1-4.

Jin, Y., Carman, M., Zhu, Y., & Xiang, Y. (2020). A technical survey on statistical modelling and design methods for crowdsourcing quality control. *Artificial Intelligence*, *287*, 103351.

Kiili, K., Lindstedt, A., Ninaus, M., & Nylén, T. (2021, December). Using a multi-step research approach to inform the development of a graph literacy game. In *International Conference on Games and Learning Alliance* (pp. 78-88). Springer, Cham.

Kittur, A., Chi, E. H., & Suh, B. (2008, April). Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 453-456).

Lauer, C., & O'Brien, S. (2020, October). The deceptive potential of common design tactics used in data visualizations. In *Proceedings of the 38th ACM International Conference on Design of Communication* (pp. 1-9).

Le, J., Edmonds, A., Hester, V., & Biewald, L. (2010, July). Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation* (Vol. 2126, pp. 22-32).

Lease, M. (2011, August). On quality control and machine learning in crowdsourcing. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the national academy of sciences*, *108*(22), 9020-9025.

Nguyen, V. T., Jung, K., & Gupta, V. (2021). Examining data visualization pitfalls in scientific publications. *Visual Computing for Industry, Biomedicine, and Art*, *4*(1), 1-15.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, *24*(3), 45-77.

Offermann, P., Levina, O., Schönherr, M., & Bub, U. (2009, May). Outline of a design science research process. In *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology*(pp. 1-11).

Qiu, C., Squicciarini, A. C., Carminati, B., Caverlee, J., & Khare, D. R. (2016, October). CrowdSelect: increasing accuracy of crowdsourcing tasks through behavior prediction and user selection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (pp. 539-548).

Ramírez, J., Baez, M., Casati, F., Cernuzzi, L., & Benatallah, B. (2020, October). Challenges and strategies for running controlled crowdsourcing experiments. In *2020 XLVI Latin American Computing Conference (CLEI)* (pp. 252-261). IEEE.

Ramly, C., Sen, A., Kale, V., Rau, M. A., & Zhu, J. (2021). Digitally Training Graph Viewers against Misleading Bar Charts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Sun, S. (2011). Meta-analysis of Cohen’s kappa. *Health Services and Outcomes Research Methodology*, *11*(3), 145-163.

Toloka API. How does Toloka work? Retrieved May 26, 2022, from toloka.ai, https://toloka.ai/docs/guide/concepts/overview.html

Tran-Thanh, L., Huynh, T. D., Rosenfeld, A., Ramchurn, S. D., & Jennings, N. R. (2014, May). Budgetfix: budget limited crowdsourcing for interdependent task allocation with quality guarantees. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*(pp. 477-484).

Venetis, P., & Garcia-Molina, H. (2012, August). Quality control for comparison microtasks. In *Proceedings of the first international workshop on crowdsourcing and data mining* (pp. 15-21).

Vuurens, J., de Vries, A. P., & Eickhoff, C. (2011, July). How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11)* (pp. 21-26).

Yang, B. W., Restrepo, C. V., Stanley, M. L., & Marsh, E. J. (2021). Truncating bar graphs persistently misleads viewers. *Journal of Applied Research in Memory and Cognition*, *10*(2), 298-311.

Yung, D., Li, M. L., & Chang, S. (2014). Evolutionary approach for crowdsourcing quality control. *Journal of Visual Languages & Computing*, *25*(6), 879-890.

Zheng, Y., Li, G., Li, Y., Shan, C., & Cheng, R. (2017). Truth inference in crowdsourcing: Is the problem solved?. *Proceedings of the VLDB Endowment*, *10*(5), 541-552.

Zhu, S., Kane, S., Feng, J., & Sears, A. (2012). A crowdsourcing quality control model for tasks distributed in parallel. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems* (pp. 2501-2506).