A Method for Crowdsourced Hypothesis Generation and Verification for Scientific Research

Toshiki Mera

Graduate School of Library, Information and Media Studies
University of Tsukuba

March 2018

Contents

List of Figures

Chapter1

Introduction

Citizen science is scientific work undertaken by members of the general public, often in collaboration with or under the direction of professional scientists and scientific institutions [1], and it is recognized as a promising approach to advance research. There are different forms of citizen sciences. An extreme example is to work as amateur scientists. Citizens serve as amateur scientists and often make new discoveries [2]. Another extreme example is to participate in a part of the scientific research process, such as data collection and labeling [3][4].

This paper reports our experience of an interesting form of citizen science, which we call the *Crowd Scientist*. In this approach, we distribute many self-contained small tasks to a large number of people to cover *all activities* required in a form of scientific research - making hypotheses and verifying them - to find causes of a specified phenomenon. Although workers performing tasks do not intend to join the scientific research, their small activities are organized *as if the crowd collectively served as a scientist as a whole*.

We designed a Crowd Scientist framework consisting of two phases (Figure 1.1). In the first phase, called the *hypothesis collection and ranking* phase, we ask the crowd workers to provide hypotheses on a specified topic. For example, the task may ask workers to enter what they think are the causes for good sleep. The same task also asks the crowd workers to state their opinion as to whether some of the hypotheses that the other people entered are true. The combination of the two questions effectively ranks the collected hypotheses on the topic. In the second phase, called the *hypothesis verification* phase, we ask the crowd to verify some of the top-ranked hypotheses found in the first phase, by having them take off-line actions. In some cases, the

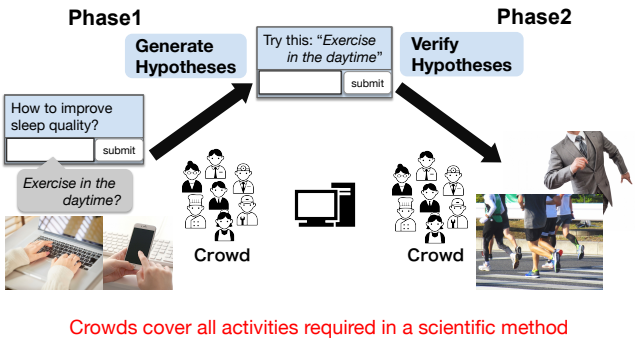

Figure.1.1 A Crowd Scientist framework. In Phase 1, we ask the crowd workers to provide hypotheses on a specified topic. Then in Phase 2, we ask the crowd to verify some of the top-ranked hypotheses found in Phase 1.

crowd themselves serve as test subjects. For example, if the hypothesis is "Having a cup of milk helps a good sleep," the task asks crowd workers to have a cup of milk before going to bed for a week. Later, we ask them to report the results, and we use the reports to decide whether each hypothesis is likely true.

Not only is the concept of the Crowd Scientist interesting, but it has several advantages. First, we can have a tremendous number of people who making hypotheses. One report [5] suggests that hypotheses remain that even experts yet to address. For a certain kind of problems, citizens can make promising hypotheses. For example, assume that we want to know the cause of a certain kinds of allergy that suddenly spread throughout a country. The cause might be newly released soap. Such a hypothesis is expected to be obtained through crowdsourcing. Second, because the approach requires no contribution from experts, we can ask the crowd scientists to answer questions to which we have not yet tried to find answers. Of course, the crowd scientist approach is not a magic wand, but we show that it works well for questions that satisfy certain conditions.

The contributions of the paper are as follows:

Crowd Scientist framework for certain types of questions. This paper provides a Crowd Scientist framework, discusses its limitations, and presents a two-phase workflow that tries to keep the number of tasks relatively small and lets us know when we should stop Phase 1, while the number of hypotheses continues to increase. Although the framework is not suitable for solving all scientific questions, we believe that this is an important first step to develop more powerful frameworks for the crowd scientists.

Experiments with real-world crowd workers. This is the main contribution of the paper. Does the Crowd Scientist framework work well in reality? Do the obtained hypotheses contain interesting ones? Do crowd workers really perform the off-line task for verifying the hypothesis? If so, how many workers? What will be the result? How much did we need for it? We report our experience of applying the Crowd Scientist framework for finding hypotheses about causality for good sleep. Our approach resulted in a lot of hypotheses, some of which were verified, and a few of the verified hypotheses were quite interesting.

The remainder of this paper is organized as follows. Chapter 2 shows related work. Chapter 3 explains a Crowd Scientist framework, the applicability conditions, and its limitation. Chapter 4 presents a workflow design for the framework. Chapter 5 and 6 shows the experimental results. Chapter 7 concludes.

Chapter2

Related Work

There are many attempts of citizen science involving crowdsourcing. Fold It [3] asks crowd workers to play an online puzzle game about protein structures for protein folding. Galaxy Zoo [4] asks crowd workers to classify large numbers of galaxies. The projects exploits *human computation capabilities*, to solve problems for which algorithms are unable to solve. We use the human computation and off-line task capabilities to generate hypotheses and verify them, while controlling the tasks performed by the crowd, which serves as a scientist as a whole.

In the medical research, crowdsourcing has been mainly used to collect data on people. Swan surveyed crowdsourcing approaches to collect data on people's health [5]. Collected data are often used for constructing prediction models [6].

Recently, there have been attempts to obtain data from the crowd to help hypotheses on the health. Bevelander et al. [7] collects information on people's childhood to understand obesity. Aramaki et al. [8] explicitly asked the crowd to provide hypotheses. In those attempts, we need experts to make hypotheses or choose some of them, and to verify the obtained hypotheses. There are online medical diagnosis services where people can ask medical questions to others including ordinary people [9]. In such services, answers are ranked in some way but the rankings are not verified.

Marshall et al. propose to combine machine learning and crowdsourcing to help us perform systematic review of literatures [10] to obtain higher evidence levels. The approach is interesting but it is different from our approach in the problem layer, and can be used to achieve higher evidence levels.

In the first phase of our proposed framework, we need to select a small number of hypotheses that are likely true in order to create questions for microtasks. Unlike top-k selection of objects with the complete data [11], we must select hypotheses under the condition that we do not know their actual probabilities of being true. This is related to the multi-armed Bandit problem [12]. However, our setting is different from a typical multi-armed Bandit, in that we choose more than one arm in each task (as explained in Section 4.1.1, our task in Phase 1 asks workers to choose more than one hypothesis from a list), and that new machines are dynamically added. There are techniques to decide when to stop collecting data [13]. Our problem is unique in that there are a variety of expressions to state the same or similar hypotheses, and thus the number of collected data items increases linearly without sophisticated supports. Although we use the stability of the top-ranked ones at present, applying techniques proposed in other frameworks is an interesting future work.

Chapter3

A Crowd Scientist Framework

This chapter discusses the applicability conditions and the limitation of a Crowd Scientist framework that performs a type of the scientific method; Given a specified phenomenon (e.g., good sleep), the framework tries to identify the cause-effect relationship for the phenomenon. For that purpose, it asks crowd workers to perform two things - generate and verify hypotheses.

We assume that crowd workers are not experts in the filed and do not do two things: first, they do not perform a literature survey and second, they do not perform experiments with specialized equipments. Rather, they answer questions based on their experience and recognition, and they perform what they can do in their daily lives.

3.1 Applicability Conditions

Given the assumptions on crowd workers, the crowd scientist works on the problems that satisfy the following two conditions.

1. Given a specified phenomenon, there must be two types of people in the worker set: people who experience the phenomenon (e.g., good sleepers), and those who do not (e.g., bad sleepers).
2. The hypotheses are related to the things the workers can observe. Examples include the environment around them, their experiences, their actions, and their physical conditions. They cannot deal with, for example, a hypothesis involving genetic disorders because they cannot verify the hypothesis.

3.2 Limitations in the Evidence Level

In order to find the causes of a given phenomenon using a hypothesis, we need to conduct *designed experiments*, in which we have a total control to change variables in the experiments. In contrast, the research method in which we draw inferences from a sample of a population where the independent variable is not under the control of the researcher is called *observational study*.

Unfortunately, we cannot perform designed experiments in a crowdsourcing setting. For example, we cannot perform a random sampling to assign crowd workers to the experimental group (the group of people on which the experimental procedure is performed) and the control group (The group of people on which the experimental procedure is not performed).

However, all observational studies are not the same. In medical research that involves ethical concerns and

Table.3.1 Evidence levels [14].

Level	Description
I	Evidence obtained from at least one properly randomized controlled trial.
II-1	Evidence obtained from well-designed controlled trials without randomization.
II-2	Evidence obtained from well-designed cohort or case-control analytic studies, preferably from more than one center or research group.
II-3	Evidence obtained from multiple time series with or without the intervention. Dramatic results in uncontrolled experiments (such as the results of the introduction of penicillin treatment in the 1940's) could also be regarded as this type of evidence.
III	Opinions of respected authorities, based on clinical experience, descriptive studies, or reports of expert committees.

where it is often difficult to conduct designed experiments, the researchers defined the *evidence levels* of the results of observational studies (Table 3.1), which indicates how strongly the results suggest a cause-effect relationship, according to how each method avoids bias in the results. Better evidence levels are important in making decisions in the medical research area pursue [14].

Studies with the evidence levels at a higher position in Table 3.1 are considered to show stronger evidence for the suggested causal-phenomenon relationship. By introducing tasks for prospective examination, our proposed framework takes the first step in moving the existing crowd-supported data collection framework towards the II-2 evidence level, although we need to carefully design the tasks more carefully in order to actually reach evidence level II-2.

Chapter4

Workflow Design

This chapter explains a workflow design of the crowd scientist framework, which can be used to obtain and verify hypothesis on the phenomena related to people's health condition (called an *outcome* in medical statistics).

Since the budget is not infinite, we need to carefully design the workflow. As shown in the experiment chapters, the number of generated hypotheses linearly increases (Figure 6.6) if we do not show all hypotheses to every crowd worker, and the crowd workers often generate meaningless or useless hypotheses (Figure 5.2, Figure 6.4).

The most costly part of crowd scientist is Phase 2, where we need to ask crowd workers to do off-line tasks (for a long time in some cases) for prospective study. Therefore, the crowd need to choose a small number of hypotheses in Phase 1 that are likely to be true and deserve to be verified in Phase 2.

Effectively choosing good hypotheses with a limited budget is a variation of multi-armed bandit problems: Given a set of hypotheses, we choose ones that are likely to be true; To find them, we ask a certain amount of crowd workers whether each hypothesis is consistent with their past experience. For that purpose, we use odds ratios as explained next.

There are two problems here. First, we cannot do it for all hypotheses because there are too many (and often useless) hypotheses. To reduce the number of hypotheses, which we ask crowd workers whether they think apply to their past experience, we devise a method to limit the number of the questions for choosing likely hypotheses.

Second, we cannot know when to stop collecting hypotheses from crowd workers while the number of hypotheses linearly increases. We solve this problem by implementing into the same task both questions for collecting hypotheses and asking the worker experience on other hypotheses.

4.1 Phase 1: Generating and Ranking Hypotheses

In Phase 1, the crowd generate a ranked list $[h_1, \ldots, h_m]$ of hypotheses on the causes of a specific outcome. The ordering represents how they are likely to be true and deserve to be verified in Phase 2.

Our framework takes as input a sentence representing an outcome A and a set Q_A of questions (in a natural language) for knowing whether each worker experiences the outcome. For example, Q_A could be a direct question, such as "Do you usually sleep good?" but it could be implicit questions to determine whether the worker experiences the outcome such as the Pittsburgh Sleep Quality Index (PSQI).

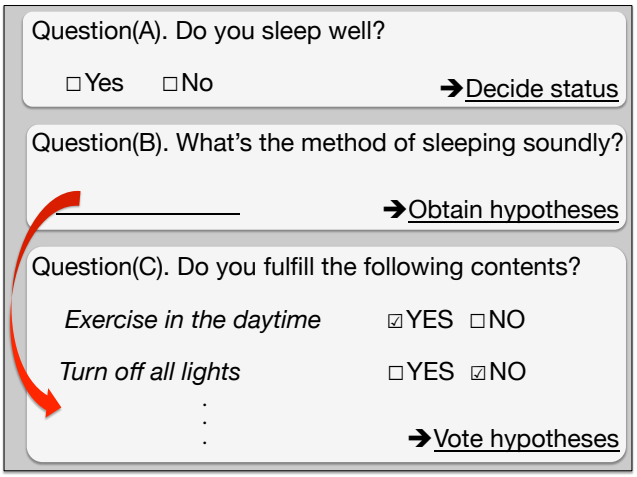

Figure.4.1 An example of questions used in the hypothesis generation tasks.

Table.4.1 A cross-tabulation T used for calculating the hypothesis probability $P(h_i)$.

		Worker's condition	
		A	$\neg A$
Experience is consistent with h_i	Yes	a	b
	No	c	d

4.1.1 Hypothesis Generation Tasks

Given A and Q_A, the framework constructs and submits tasks, called the *Hypothesis Generation Task*, to ask crowd workers to work for the following two things.

- Generate hypotheses on the cause of an outcome A. Question (B) in Figure 4.1 directly asks them to do this.
- Construct a cross-tabulation (Figure 4.1) for each hypothesis. Questions (A) and (C) are implicitly used for that purpose; Question (A) asks workers for answering the question Q_A, which classifies the workers into two groups - those that experiences A and those that do not. Question (C) shows the list of possible causes and asks the worker which causes are consistent with his/her experiences.

This way, the task obtains hypotheses and finds those that are likely to be true from the ones generated by other workers at once.

4.1.2 Exact Ranking of Hypotheses

Assume that we do not have to limit the number of task submissions and workers, and that we have many workers to answer Questions (A) and (C) for every hypothesis obtained through Question (B) enough to populate meaningful numbers to their cross-tabulations. Then, what are the hypotheses that deserve to be

Algorithm 1 Choosing hypotheses for Question (C)

Input: m, H_t

Output: $H_{Q(C)}$

1: $H_{Q(C)} \leftarrow \phi$
2: **while** $(|H_{Q(C)}| < m) \wedge (H_t \neq \phi)$ **do**
3: $h_i \leftarrow WeightedRandomChoice(H_t)$
4: delete h_i from H_t
5: **if** $Correlated(h_i, H_{Q(C)})$ **then**
6: add h_i to $H_{Q(C)}$
7: **end if**
8: **end while**

verified in Phase 2?

A common approach in the medical research area is to use the *odds ratio* [8] [7], which compares the two groups (those who experienced the possible cause and those who do not) in the outcome. Given a cross-tabulation for h_i (Figure 4.1), The odds ratio of a hypothesis h_i is computed by $P(h_i) = \frac{a/b}{c/d}$. The value represents how strongly the presence or absence of the possible cause is associated with the presence or absence of the outcome. The greater value means that the possible cause mentioned in the hypothesis is more likely to be an actual cause. The exact ranking of hypotheses is computed by the values.

4.1.3 Limiting the Number of Hypotheses in Question (C)

While computing the exact ranking requires us to have many hypotheses appear in Question (C), there are two opportunities to reduce the number of them. First, for our purpose, the exact ranking is not important, especially for low-ranked ones. What we want to find is a set of hypotheses that deserve to be verified in Phase 2. Second, since the hypotheses are generated by independent workers, there are often duplicates or equivalent ones even if they are not exact match to each other.

Therefore, we do not need a fair treatment of all hypotheses and can choose fewer number of hypotheses with the following two principles.

Principle 1: Focus on hypotheses that obtained higher odds ratios so far.
Principle 2: Focus on a representative of similar ones.

Algorithm 1 implements the two principles to choose a fixed number of hypotheses shown in Question (C) for a task. The inputs are (1) a set H_t of hypotheses that Phase 1 already obtained at time t, and (2) a natural number $m > 0$. Given them, the algorithm outputs $H_{Q(C)}$ such that $H_{Q(C)} \subset H_t$ and $|H_{Q(C)}| = m$.

The algorithm is simple: While $|H_{Q(C)}| < m$ and $H_t \neq \phi$ (Line 2-7), pick up a hypotheses from H_t with Principle 1 (Line 3-4) and add to $H_{Q(C)}$ with Principle 2 (Line 5-6).

Principle 1 Implementation. Principle 1 is implemented by the $WeightedRandomChoice(H_t)$ function in Line 3, which randamly choose one with the weight $P(h_i)$ for each h_i. This is based on a heuristic that hypotheses that have a higher odds rate at a certain point is also likely to have higer one at the final state.

Principle 2 Implementation. Principle 2 is implemented by the $Corrected(h_i, H_{Q(C)})$ in Line 5. It checks

Table.4.2 Table used for calculating the correlation between h_i and h_j.

Experience is consistent with h_i / h_j		h_i	
		Yes	No
h_j	yes	r	s
	No	t	u

if there is a hypothesis h' in $H_{Q(C)}$ such that h' and h_i are correlated to each other. To compute this, we maintain tables such as Table 4.2. This is based on a heuristic that the answers to questions on similar hypotheses are correlated to each other.

4.1.4 Combining the Three Questions

Theoretically, Question (B), and the combination of Questions (A) (C) can be implemented by distinct sets of tasks; First, we use the task with Question (B) only to generate hypotheses, and use the other tasks to rank the hypotheses. However, this would raise another problem; We cannot know when to stop submitting the first set of tasks. We solve this problem by putting all questions into the same task. This design lets us know when to stop the task; As shown in the experiment chapters (Figure 5.3 and Figure 6.5), newly-added hypotheses do not obtain high odds ratios after a certain number of tasks finished, which means that eventually the top-ranked hypotheses becomes stable. Therefore, we can use its stability as the cue to stop Phase 1.

4.2 Phase 2: Verifying Hypotheses

The output of Phase 1, which is computed based on the past experiences of workers, often shows pseudo-correlation and opposite direction of causality. In Phase 2, the crowd verifies a small number of hypotheses chosen from the result of Phase 1 in the prospective direction.

Given a set $\{h_{i1}, \ldots, h_{ik}\}$ of hypotheses, Phase 2 outputs the test results based on workers' prospective experiences.

Phase 2 employs two types of tasks. The subject-recruitment tasks are used to recruit workers who will perform off-line tasks for verifying each hypothesis. The result-report tasks ask the workers to report their results.

In both tasks, we put the same question as Question (A) in the hypotheses generation tasks, so that we can know that each worker is in the condition described by A or not. Given h_i, let S_{h_i1} be the results of the subject-recruitment tasks that ask workers to follow the instructions for verifying h_i. Let s_{h_i2} be the results of the result-report tasks from those who performed the actions. Our framework checks whether the difference between S_{h_i1} and Sh_i2 is significant.

We pay workers of the subject-recruitment tasks regardless they perform the result-report tasks or not. Therefore, how many workers will report the effects of their off-line tasks is totally up to the workers.

4.2.1 Subject-Recruitment Tasks (First-Step Tasks)

The subject-recruitment tasks ask two things. First, it asks workers to answer Q_A, to know whether s/he is in the condition described as A. Second, it asks workers to follow the instructions of taking off-line actions that can be a cause that leads to A.

The actions are chosen from the causes written in $\{h_{i1}, \ldots, h_{ik}\}$. If the hypothesis is given in some structured form, we can extract the cause and semi-automatically generate the instruction. For each h_{ij}, we submit the task to m workers so that we can obtain a reasonably large number of results of the result-report tasks (The workers for the second task must be a subset of the workers for the first task).

In the task instruction, a prior notice that the next task will be coming later and the payment will be reasonably high is shown.

4.2.2 Result-Report Tasks (Second-Step Tasks)

The result-report tasks ask workers to report the effects of their off-line acts. Instead of asking workers to explicitly report the result, it poses Question (A) again, to know the current condition of workers. The tasks are not open to every workers, because the tasks need to be performed by those who performed the subject-recruitment tasks.

Chapter5

Experiment 1

In order to evaluate the proposed method explained in Chapter 4, experiments for the two phases were conducted.

5.1 Settings

For the experiment, we chose the topic "What makes a good night's sleep", because the topic is familiar to ordinary people. We used the Pittsburgh Sleep Quality Index (PSQI) [15], which is a popular indicator measured with a self-report questionnaire to measure the quality of the workers' sleep in Question (A) and the phase 2 tasks.

The tasks were created on *Crowd4U*[*1], and workers were recruited in *Yahoo! Crowdsourcing* [*2] for the experimental evaluation. *Yahoo! crowdsourcing* is a major crowdsourcing service in Japan, where people of each attribute are indiscriminately included as workers. In this experiment, the tasks were created in Japanese and answers of them also were received in Japanese.

5.2 Parameters

5.2.1 Parameters for Hypothesis Generation Task

Parameters used for implementing the hypothesis generation tasks in Section 4.1.1 are as follows.

Number of hypotheses obtained per a task: Two questions, "Q-B1) How can you get a better night's sleep?" and "Q-B2) What do you think we can do to get a better night's sleep?", were set as Question (B) in the hypothesis generation tasks. The difference between these questions is whether the hypothesis is based on worker's own experience or not. Every worker answered at least one or two hypotheses by these questions.

Test method for calculating correlation: Fisher's exact test was used to judge whether a correlation exists between the hypotheses.

Number of hypotheses used in the hypothesis generation task: For ease of checking the obtained hypotheses, 10 hypotheses are selected and shown in the question (C), that is $m = 10$ in Section 4.1.3.

Value of $P(h_i)$ to be assigned to newly obtained hypotheses: $P(h_i)$ values of newly obtained hy-

[*1] http://crowd4u.org/

[*2] http://crowdsourcing.yahoo.co.jp

potheses become very small using the odds ratio, which is a kind of cold start problem. In order to solve the problem, a fixed value is assigned to $P(h_i)$ until a fixed number of answers are obtained, so that these hypotheses are selected with priority. The maximum value of $P(h_i)$ existing at that time is assigned for hypotheses obtained when the number of their answers is less than 10 in this experiment.

Kind of choice answered by worker in question (C): In the question (C), two choices of 'YES' and 'NO' were set in order to know whether each worker corresponds to the content of hypotheses. In addition, a choice of "the question's sentence is nonsense" was set.

Price Setting and the Period: We paid four JPY (about four cents in USD) to each worker who performed the task. We made the tasks open in the crowdsourcing service for seven days.

5.2.2 Parameters of Hypothesis Verification Task

Parameters set for implementing the hypothesis verification task in the Section 4.2.1 are as follows.

Number of hypotheses to verify in hypothesis verification task: The hypotheses to be verified in Phase 2 are the 10 hypotheses in Table 5.1 among the obtained and screened in Phase 1.

Interval between each step of the hypothesis verification task: In the hypothesis verification task, the interval from the execution of the first step task to the request of the second step task was set to one week.

Price Setting and the period: We paid five JPY (about five cents in USD) to a worker who performed a subject recruitment (First-Step) task, and 100 JPY (about one dollar in USD) for each result-report (Second-Step) Task. We made the first step tasks open in the crowdsourcing service for five days. The second step tasks were put seven days later after we closed the first-step tasks and kept open for five days.

5.3 Results

Number of tasks performed in each phase is shown in Figure 5.1. We paid 90,326 JPY (Tax excluded) in total. We paid 18,600JPY (Tax excluded) for the hypothesis generation tasks in Phase 1, 7,670 JPY and 64,056 JPY for the first and second step tasks in Phase 2, respectively.

Even if we consider the difference in the period of keeping the task open in the crowdsourcing service (seven days for the phase 1 task, five days for each of the phase 2 tasks) the numbers of the tasks performed in Phase 2 was much smaller than that of tasks in Phase 1. In the first-step task in Phase 2, the instruction states that we will submit the result-report (second-step) tasks one week later, and ask you to perform the task. This may be the reason for the small number of the first step tasks. We expected there were few workers who actually perform the second-step tasks. Interestingly, half of them performed the second step tasks, partly because we stated that we would pay 100 JPY for the second-step task in the instruction in the first-step task.

5.3.1 Results of the Hypothesis Generation Tasks

1,546 tasks were performed and 2,619 hypotheses were generated in the hypothesis generation tasks. Of those, 648 hypotheses received more than 10 answers for question (C). After excluding hypotheses for which two or more workers answered "the question's sentence is nonsense," there were 640 hypotheses. The odds ratio values of 297 out of the 640 hypotheses exceeded 1.0 and those of 310 hypotheses was less than 1.0.

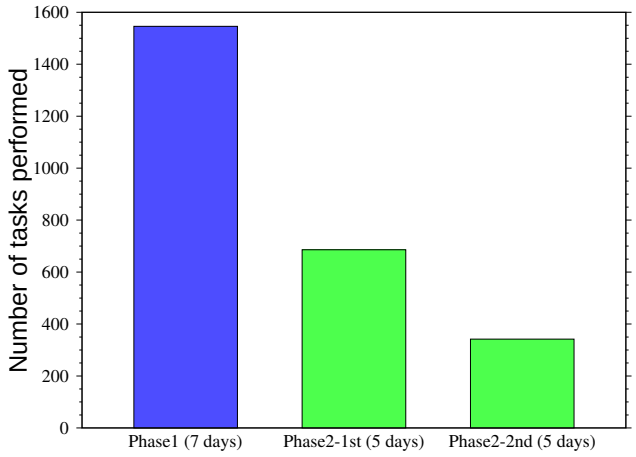

Figure.5.1 Number of tasks performed in each phase.

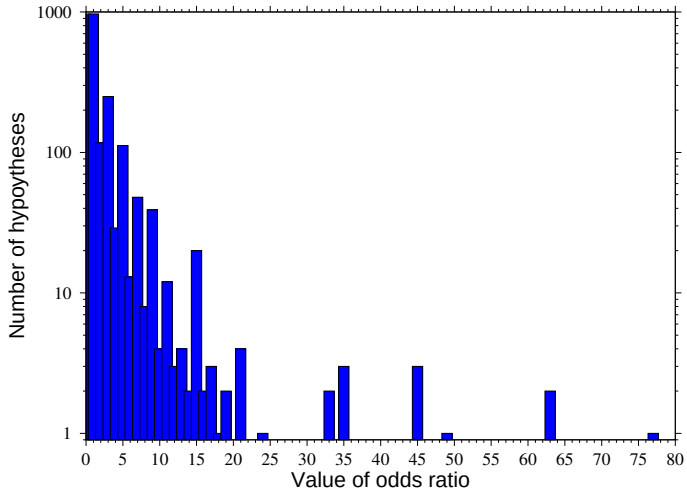

Figure.5.2 Frequency distribution of hypotheses. X-axis represents the odds ratio (per one). There are so many insignificant hypotheses generated by the crowd in Phase 1.

The maximum value of odds ratio was 35.0 and the minimum value was 0.0175.

Figure 5.2 shows the frequency distribution of hypotheses' odds ratios. The result shows that there are so many insignificant hypotheses generated by the crowd in Phase 1. And Figure 5.3 shows scatterplot of odds ratios of all hypotheses. The result shows that hypotheses with extreme values such as some of 1300-1400th were not generated and the ranking of hypotheses becomes stable in the end of Phase 1.

The accumulative number of duplicate hypotheses is shown in Figure 5.4. The number of duplicate hypotheses linearly increases, partly because each task does not show all hypotheses to the workers.

5.3.2 Hypotheses Selection for Hypothesis Verification Task

Figure 5.5 shows the entire workflow for selecting hypotheses in this experiment. (The value of n in the figure is explained in Section 5.3). The upstream part was conducted by crowd workers in Phase 1.

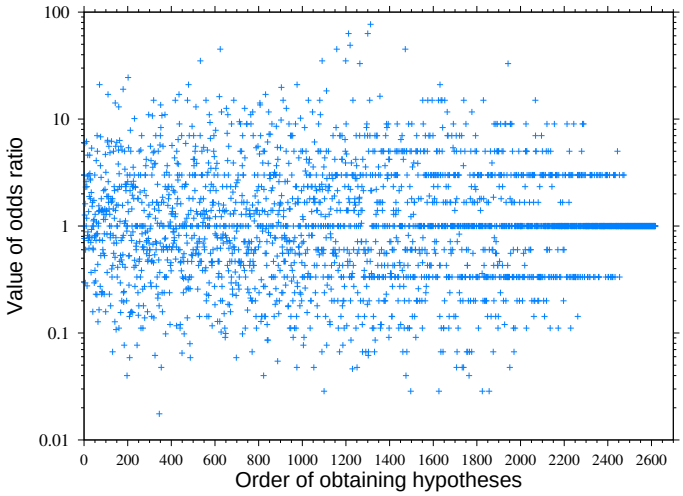

Figure.5.3 Scatterplot of odds ratios of all hypotheses. An point (hypothesis) located at i on the X-axis means that the hypothesis was generated ith in Phase 1. It shows that the ranking of hypotheses becomes stable in the end of Phase 1.

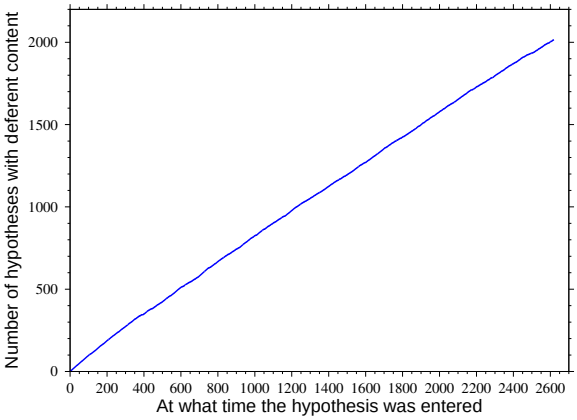

Figure.5.4 The accumulative number of hypotheses (duplicates excluded). X-axis represents when each hypothesis was entered. Although we excluded duplicates, the number of hypotheses linearly increases. This is partly because there is a variation of expressions to state similar hypotheses.

The downstream part was conducted as follows. (1) medical workers first exclude ethically problematic hypotheses out of the top 100; (2) then, experts add one of two labels to each hypothesis: “seems to be effective” or “seems to be ineffective,” and (3) finally, we manually selected ten hypotheses out of the 100 hypotheses. In the selection, we tried to select hypotheses without ambiguous expressions and sometimes added specific numbers such as “(within 5 minutes)” or “(within an hour)” to ambiguous expression like “before sleeping” and “before going to bed”.

We note that the intervention by the people who were not recruited through crowdsourcing is not the essential part of our framework; We needed it to pass the research ethics review and obtain the data for the evaluation. Actually, (1) and (3) could be done by non-expert crowd workers and (2) was not necessary if you do not need to conduct the experiments reported in the paper.

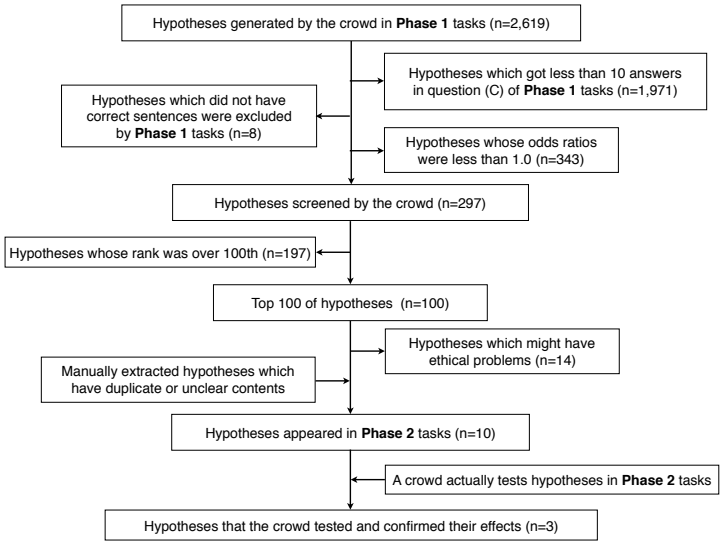

Figure.5.5 A workflow of hypotheses selection in experiment 1.

Table.5.1 Hypotheses shown in the hypothesis verification task.

ID	Contents of hypotheses
1	Take about a 5-minutes bath more than one hour before going to bed
2	Turn off all room lights just before going to bed (within 5 minutes)
3	Walk for more than 15 minutes during the day
4	Do not view bright screens such as a PC or TV before going to bed (within 1 hour)
5	Warm the futon with a futon dryer or a hot water bottle before sleeping (within 1 hour)
6	Use a smartphone before going to bed (within an hour)
7	Do not go to the futon until you want to sleep
8	Before bedtime (within an hour), listen to quiet, relaxing music
9	Take a slow deep breath just before going to bed (within 5 minutes)
10	Make dinner lighter than other meals

5.3.3 Results of the Hypothesis Verification Task

686 workers responded to the first-step tasks and 342 workers responded to the second-step tasks in the hypothesis verification task. Table 5.1 shows the hypotheses selected in the hypothesis verification task and Table 5.2 shows the analyzed answers of the 342 workers who completed the second-step tasks. The meanings of the columns *Improved* and *Cured* are as follows.

Improved: Cross-tabulation answers between the number of workers whether the content of a hypothesis tested for the majority of a week and the number of workers whether the PSQI score decreased. Note that the better the quality of sleep is, the lower the value will be. *Improved* is the p value obtained by Fisher's exact test (two-sided test) using those results.

Table.5.2 Evaluation of task results (Underline: Significance level is less than 5%).

Expert's judgement	ID	workers	*Improved*	*Cured*
Effective	1	21	1.0000	1.0000
	2	55	0.3498	<u>0.0008</u>
	3	28	1.0000	0.1440
	4	40	0.3332	0.6339
	5	26	0.3304	1.0000
Ineffective	6	22	0.2536	0.2281
	7	18	1.0000	0.6447
	8	50	0.3725	<u>0.0201</u>
	9	60	<u>0.0217</u>	0.1534
	10	22	1.0000	0.3512

Cured: Cross-tabulation answers between the number of workers whether the content of a hypothesis tested for the majority of a week and the number of workers whether the PSQI score decreased crosses the threshold (6 points), namely, whether the condition changed from insomnia to sleeping well. *Cured* is the p value obtained by the test using those result.

According to the *improved*, only the hypothesis labeled "Ineffective": "Take a slow deep breath just before going to bed (within 5 minutes)" became significant (with significance level 5%). As for the *cured*, the hypothesis labeled "Effective": "Turn off all room lights just before going to bed (within 5 minutes)", and the hypothesis labeled "Ineffective": "Before bedtime (within an hour), listen to quiet, relaxing music" became significant (with the significance level 5%). Figure 5.6 shows distributions of each worker's PSQI scores for hypotheses whose ID's are one through ten. Plots at $y < x$ indicate that the PSQI score decreased (the quality of sleep improved), and plots at $y > x$ indicate that the score increased (the quality of sleep became worse).

5.4 Discussion

Experimental results demonstrated that this method could collect various hypotheses from over 1,500 workers. One of the advantages of our approach is that we could investigate only the feasible hypotheses. Our approach clearly divided this process into two phases; (1) Phase 1 (generating and ranking hypotheses) and Phase 2 (verifying hypotheses). More than 600 people participated per five days, and about 50% of them performed tasks until the last with a pay of only 105 Japanese yen per a worker.

Although both phases should be conducted by the same workers, workers who performed tasks in Phase 2 were less than Phase 1. This is one of the biggest issues to be solved. Another issue is dropout in Phase 2, which has two tasks (subject recruiting task and result report task). In the experiments, 50% of the workers dropped out before the result report task, causing sampling bias.

The final step (result reporting task) gave us three feasible hypotheses, in which one (hypothesis ID 2) has been already known by medical experts, and the other two (hypothesis IDs 8 and 9) have not been recognized as new hypotheses.

Hypothesis ID 2 ("Turn off all room lights just before going to bed (within 5 minutes)") As mentioned, this hypothesis has been already known by experts. The relation between room light and sleep quality is frequently discussed [16]. Most of these studies are based on the controlled experiments utilizing measuring vital tools, such as a heart rate variability (HRV) and their electroencephalographic (EEG) power spectrum.

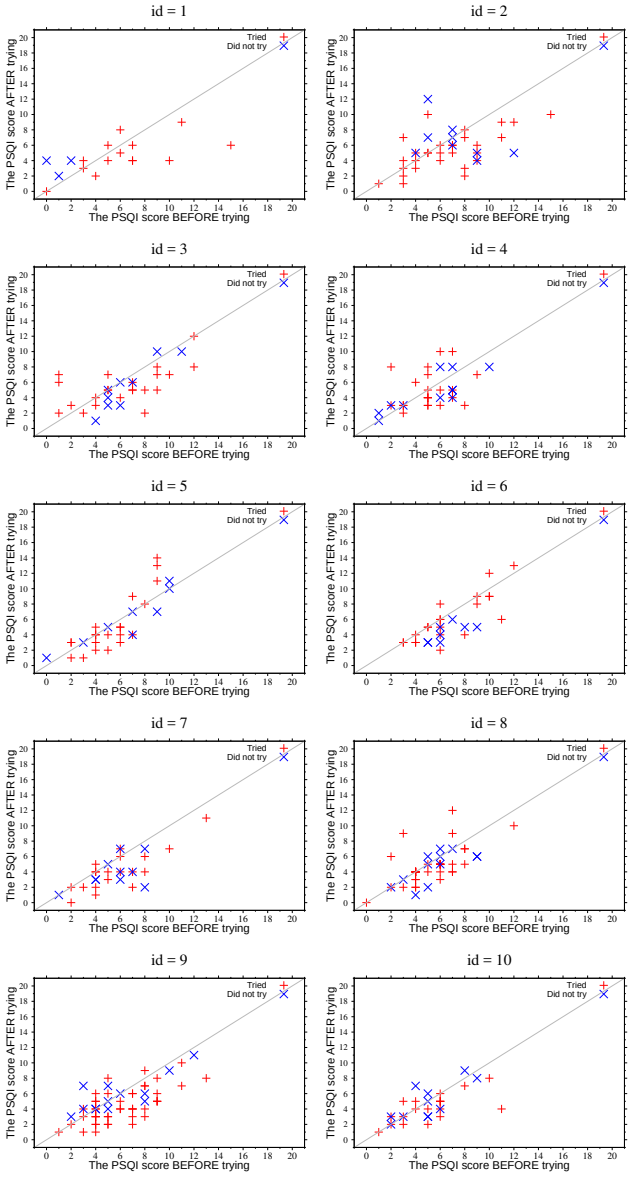

Figure.5.6 Distributions of PSQI scores for each hypothesis. X-axes represent the value of the PSQI score before testing a hypothesis and Y-axes represent it after testing. Plots "Tried" represent workers who tested the hypothesis for the majority of a week, and plots "Did not try" represent workers who did not. Plots at $y < x$ indicate that the PSQI score decreased (quality of sleep improved), and plots at $y > x$ indicate that the score increased (quality of sleep became worse).

The previous studies focus mainly on the mechanism of light or no-light effect. Although our result could not reveal the hidden mechanism of light, but it could directly reach the final result that light improves the sleep quality.

Hypothesis ID 8 ("Before bedtime (within an hour), listen to quiet, relaxing music") is an unknown hypothesis. We, however, could find similar studies on the relation between therapies and the sleep quality [17]. Note that the music listening (music therapy) is one of the targets. Our result encourages the future studies in this field.

Hypothesis ID 9 ("Take a slow deep breath just before going to bed (within 5 minutes)"), which is also a new finding. Like the ID 8, we could find similar studies. One of such studies is mindfulness, which draws much attention recently. The mindfulness of breath exercise has sometimes reported that it improves the sleep quality[18]. The major difference between the mindfulness study and our findings is the simplicity of the hypotheses. While the mindfulness has a formal protocol (breath out timing and breath in timing, the body pose, times of breathing, etc.), our hypothesis simply mentioned on "(to take) a slow deep breath". We believe that one of the advantages of our approach is this simplicity, enabling to collect easy to understand and easy to take off-line actions.

Chapter6

Experiment 2

Experiment 2 was conducted by modifying some parameters from Chapter 5 experiments.

6.1 Settings

Settings of this experiment were the same as in experiment 1. We also chose the question "What makes a good night's sleep?", and used PSQI and recruited workers via *Yahoo! crowdsourcing*.

6.2 Parameters

6.2.1 Parameters for Hypothesis Generation Task

Parameters used for implementing the hypothesis generation tasks in Chapter 4.1.1 were as follows.

Number of hypotheses obtained per a task: One question: "What is the factor that improves sleep quality?" was prepared. In the input form, a tree structure with each hypothesis as a node was shown. (As in Figure 6.1). Workers could input new hypotheses as leaf nodes of the tree structure. Workers did not have to input any hypotheses when they have no idea.

Test method for calculating correlation: Fisher's exact test was also used to judge whether a correlation exists between the hypotheses.

Number of hypotheses used in the hypothesis generation task: As in experiment 1, ten hypotheses were selected and shown in the question (C). (However, this ten hypotheses were selected from leaf nodes of the tree structure.)

Calculation method of $P(h_i)$ values: As in experiment 1, the odds ratio was used for the value of $P(h_i)$. However, values used to calculate the odds ratio were the sum of the values of ancestor and descendant nodes when the hypothesis was represented by the tree structure.

Value of $P(h_i)$ to be assigned to newly obtained hypotheses: The maximum value of $P(h_i)$ was assigned for hypotheses obtained until the number of their answers was less than 10 as in experiment 1.

Kind of choice answered by worker in question (C): In the question (C), three choices of "YES" and "NO" and "The question's sentence is nonsense" were set.

Price Setting and the Period: We paid three JPY (about three cents in USD) to each worker who performed the task. We made the tasks open in the crowdsourcing service for ten days.

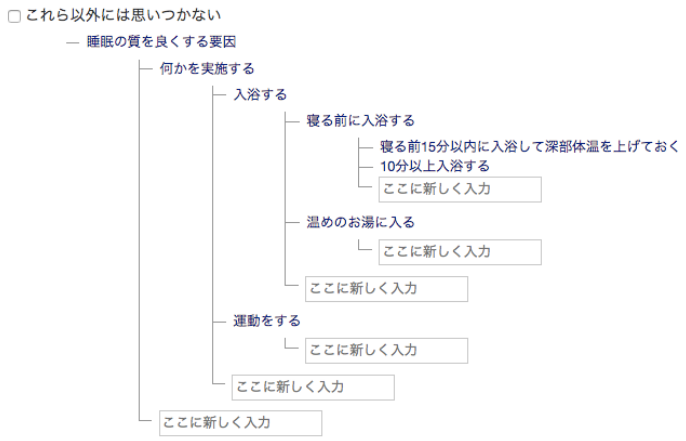

Figure.6.1 An illustration of tree structures in question (C). A tree structure with each hypothesis as a node was shown to workers. Workers do not need to input any number of forms but then they have to check the box at the top.

6.2.2 Parameters of Hypothesis Verification Task

Parameters set for implementing the hypothesis verification task in the Section 4.2.1 were as follows.

Number of hypotheses to verify in hypothesis verification task: The hypotheses to be verified in Phase 2 are the 4 hypotheses in Table 6.1 among the obtained and screened in Phase 1. In order to increase the number of workers assigned to each hypothesis, the number of hypotheses to be verified was reduced more than experiment 1.

Interval between each step of the hypothesis verification task: In the hypothesis verification task, the interval from the execution of the first-step task to the request of the second step task was also set to one week.

Number of times a worker performs Phase 2 tasks: We assigned both a week to test the hypothesis and a week to try to improve the accuracy of the method, and we conducted the task of Phase 2 twice in a row. However, to improve the accuracy of the method, we assigned both weeks to test a hypothesis and not to for each worker, and we made each worker perform the Phase 2's task twice in a row. As in Figure 6.2, we asked three tasks for each worker: (1) Task1 (first Subject-Recruitment Tasks), (2) Task2 (first Result-Report Tasks and second Subject-Recruitment Tasks) and (3) Task3 (second Result-Report Tasks).

In which period workers try hypotheses: We divided workers into two groups: Group A (try a hypothesis between Task1 and Task2) and Group B (try a hypothesis between Task1 and Task2). We tried an analysis considering the effects of placebo by checking the change of state in the same person.

Price Setting and the period: We paid two JPY (about two cents in USD) to a worker who performed Task1, five JPY (about five cents in USD) for each Task2, and 200 JPY (about 2 dollar in USD) for each Task3. We made the Task1 open in the crowdsourcing service for a day. The Task2 was put seven days later and kept open for five days. The Task3 was put seven days later and kept open for five days.

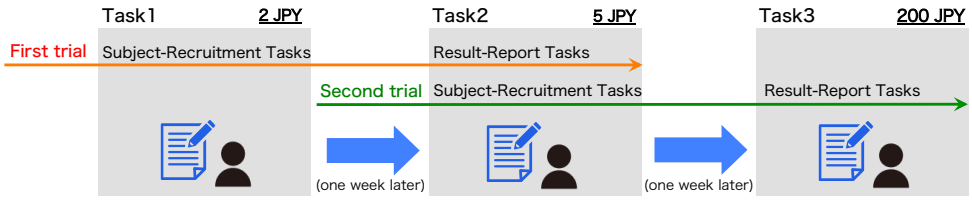

Figure.6.2 The procedure of Phase 2 in this experiment. Using three kinds of tasks, assigned Subject-Recruitment Tasks and Result-Report Tasks in Phase 2 to one worker twice. Workers tried a hypothesis on one of these trials and lived without trying any hypotheses on another trial. Whether each worker tries a hypothesis in which period was randomly determined.

Table.6.1 Candidates of hypotheses shown in the hypothesis verification task.

id	Contents of hypotheses	Majority decisions by experts
1	“Get up early and bask in the sun”	Seems to be effective
2	“Sleep sideways with the heart down”	Neither agree nor disagree
3	“Make own eyes tired by using smartphones”	Seems to be ineffective
4	“Eat three meals a day by chewing well”	Neither agree nor disagree

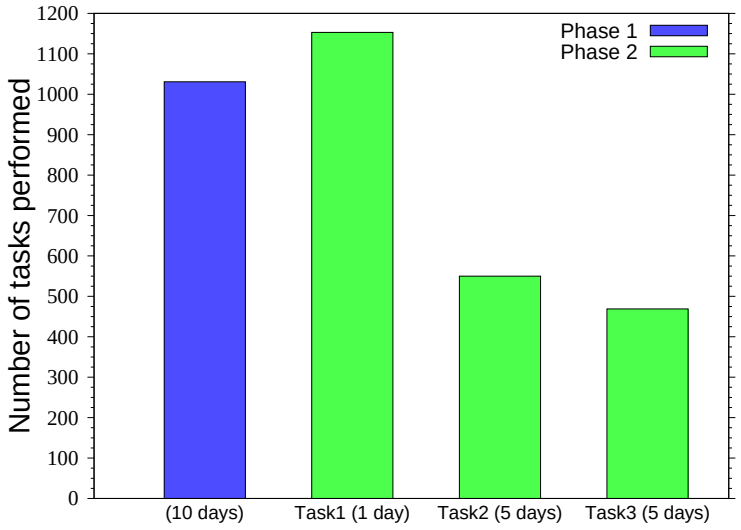

Figure.6.3 Number of tasks performed in each phase.

6.3 Results

Number of tasks performed in each phase is shown in Figure 6.3. We paid 164,863 JPY (Tax excluded) in total. In phase 1, we paid 10,100 JPY (Tax excluded). In Phase 2, We also paid 8,000 JPY (Tax excluded) for Task1, 7,692 JPY and 139,071 JPY for Task2 and Task3, respectively.

The tendency of the number of workers in this experiment was similar to that of Experiment 1. In Phase 2, Approximately half of Task1 participants continued to Task2 and Task3 when paying 200 JPY for reward.

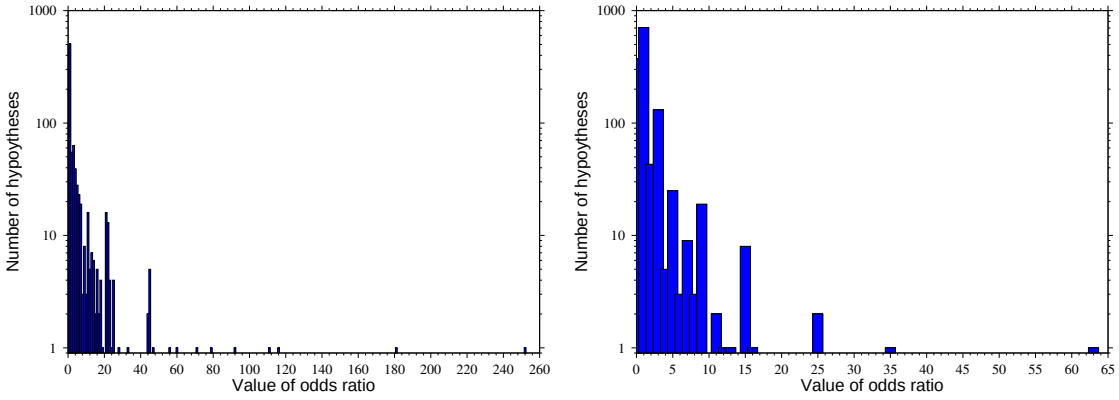

Figure.6.4 Frequency distributions of hypotheses. X-axis represents the odds ratio (per one). There are so many insignificant hypotheses generated by the crowd in Phase 1. The left figure shows odds ratios calculated by using the values of the ancestor and descendant nodes when hypotheses are represented by a tree structure. And right figure shows odds ratios calculated by using the values of a hypothesis itself.

6.3.1 Results of the Hypothesis Generation Tasks

1,031 tasks were performed and 1,337 hypotheses were generated in the hypothesis generation tasks. Of those, 827 hypotheses received one or more answers for question (C). After excluding hypotheses for which two or more workers answered "the question's sentence is nonsense," there were 685 hypotheses. The odds ratio values of 297 out of the 685 hypotheses exceeded 1.0 and those of 312 hypotheses was less than 1.0. The maximum value of odds ratio was 63 and the minimum value was 0.02222.

Figure 6.4 shows the frequency distribution of hypotheses' odds ratios. The left figure used odds ratio values calculated by using ancestor and descendant nodes' values. ($P(h_i)$ value in this experiment.) And the right figure used them calculated by using own node's values. The both of results showed that there were so many insignificant hypotheses generated by the crowd in Phase 1.

And Figure 6.5 shows scatterplot of all hypotheses' odds ratios calculated by using own node's values. The result shows that the ranking of hypotheses becomes stable in the end of Phase 1.

The accumulative number of new hypotheses is shown in Figure 6.6. The number of new hypotheses linearly increases, partly because each task did not show all hypotheses to the workers in the question (C).

6.3.2 Hypotheses Selection for Hypothesis Verification Task

Figure 6.7 shows the entire workflow for selecting hypotheses in this experiment. (The value of n in the figure is explained in Section 6.3). The upstream part was conducted by crowd workers in Phase 1. The downstream part was conducted as follows. (1) We manually selected hypotheses without ambiguous or duplicate expressions out of the top 100. (2) Then, eight sleep researchers excluded ethically problematic hypotheses (3) and they added one of three labels to each hypothesis: "seems to be effective", "seems to be ineffective" or "neither agree nor disagree". (4) Finally, we chose hypotheses in tasks in descending order of the number of votes for each label.

We note that the intervention by the people who were not recruited through crowdsourcing is not the

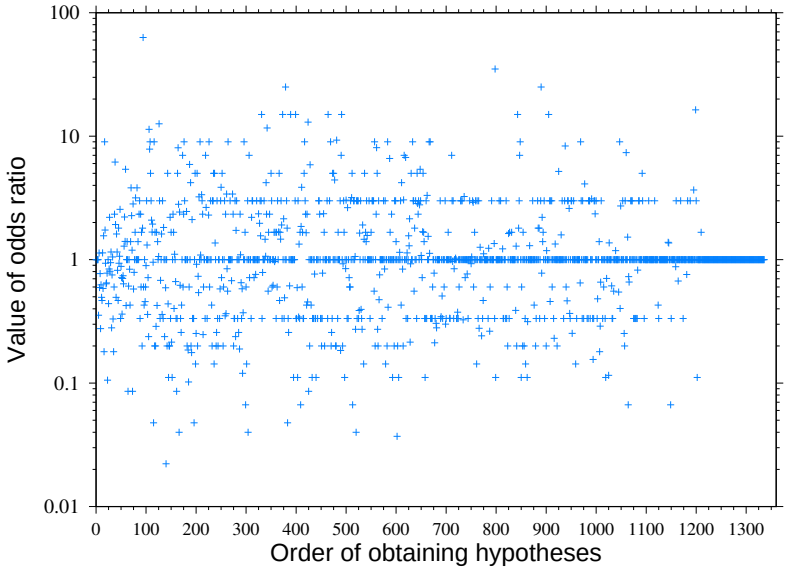

Figure.6.5 Scatterplot of odds ratios of all hypotheses. An point (hypothesis) located at i on the X-axis means that the hypothesis was generated ith in Phase 1. It shows that the ranking of hypotheses becomes stable in the end of Phase 1.

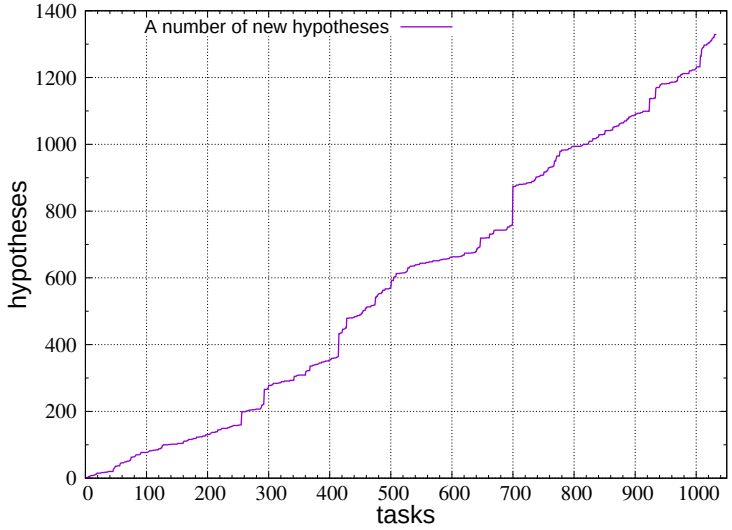

Figure.6.6 The accumulative number of hypotheses (duplicates excluded). X-axis represents the number of tasks performed in Phase 1. Although we excluded duplicates, the number of hypotheses linearly increases. This is partly because there is a variation of expressions to state similar hypotheses.

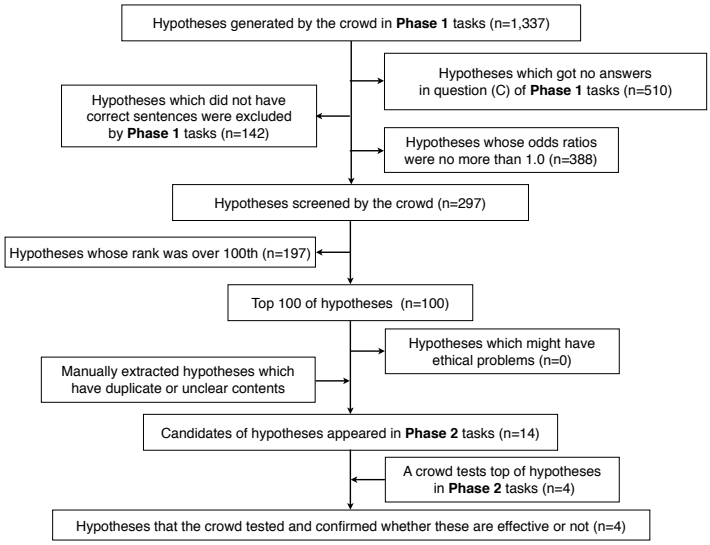

Figure.6.7 A workflow of hypotheses selection in experiment 2. The upstream part was conducted by crowd workers in Phase 1. The downstream part was conducted by the author and sleep researchers for the evaluation.

essential part of our framework; We needed it to pass the research ethics review and obtain the data for the evaluation. Actually, (1) and (2) could be done by non-expert crowd workers and (3) was not necessary if you do not need to conduct the experiments reported in the paper. (4) was conducted for this experiment to limit the number of hypotheses in Phase 2 tasks according to time and cost limitations.

6.3.3 Results of the Hypothesis Verification Task

1153 workers responded to the Task1, 550 workers responded to the Task2 and 469 workers responded to the Task3 in the hypothesis verification task.

Table 6.1 shows the hypotheses used in the hypothesis verification task.

Figure 6.8 shows the transition of the PSQI scores for each group of workers for each hypothesis. We conducted the paired-t test with difference of PSQI scores at any two points of the three kinds of tasks. Overall, the differences between the PSQI scores at the point of Task1 and Task3 in all four hypotheses were significant. According to the comparison between Task1 and Task2, or Task2 and Task3, the hypothesis of id1 has significant differences only in the week when the hypothesis was tested. In contrast, the hypothesis of id3 has significant differences only in the week when the hypothesis was not tested.

Figure 6.9 shows the relationship between the number of days the workers actually tried and the difference of PSQI scores. As the number of days tested increased, the PSQI scores were improved in the hypothesis of id1. In contrast in the hypothesis of id3, the PSQI scores were worsened as the number of days tested increased. In the hypothesis of id2, the PSQI scores were improved and worsened significantly, and there was no especial changes in the hypothesis of id4.

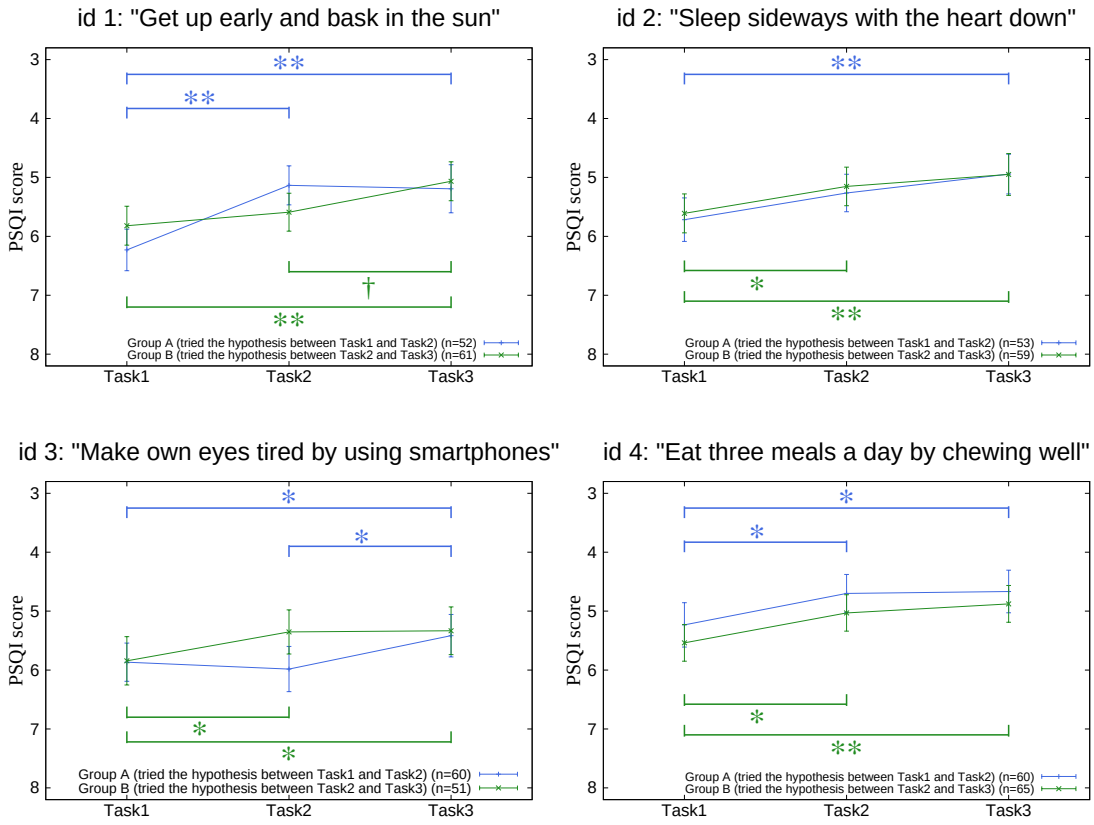

Figure.6.8 The transitions of PSQI scores' mean in Phase 2 task (with standard error bars). The two lines represent Group A (tried the hypothesis in the first week) and Group B (tried in the next week), respectively. The lower the PSQI score, the better the quality of sleep. In the figure, n represents the number of workers who performed all three tasks. And symbols represent the following: † P<0.1, $*$ P<0.05, $**$ P<0.01 .

6.4 Discussion

Experimental results demonstrated that this method could collect various hypotheses from the crowd as in experiment 1. One of the advantages of our approach is that we could investigate only the feasible hypotheses. Our approach clearly divided this process into two phases; (1) Phase 1 (generating and ranking hypotheses) and Phase 2 (verifying hypotheses). More than 1100 people participated per a day, and about 40% of them performed tasks until the last with a pay of only 207 JPY per a worker.

In contrast to Experiment 1, workers who performed tasks in Phase 2 were more than Phase 1. From this result, it can be considered that the number of participants for each phase varies widely depending on the seasons or rewards.

We used three tasks in Phase 2 in this experiment. About 60% of the workers dropped out before the Task3, also causing sampling bias as in experience 1.

We obtained more important results than experiment 1 from Phase 2 of this experiment. For all four hypotheses, the mean of the PSQI scores improved significantly from Task 1 to Task 3 in both groups, so it

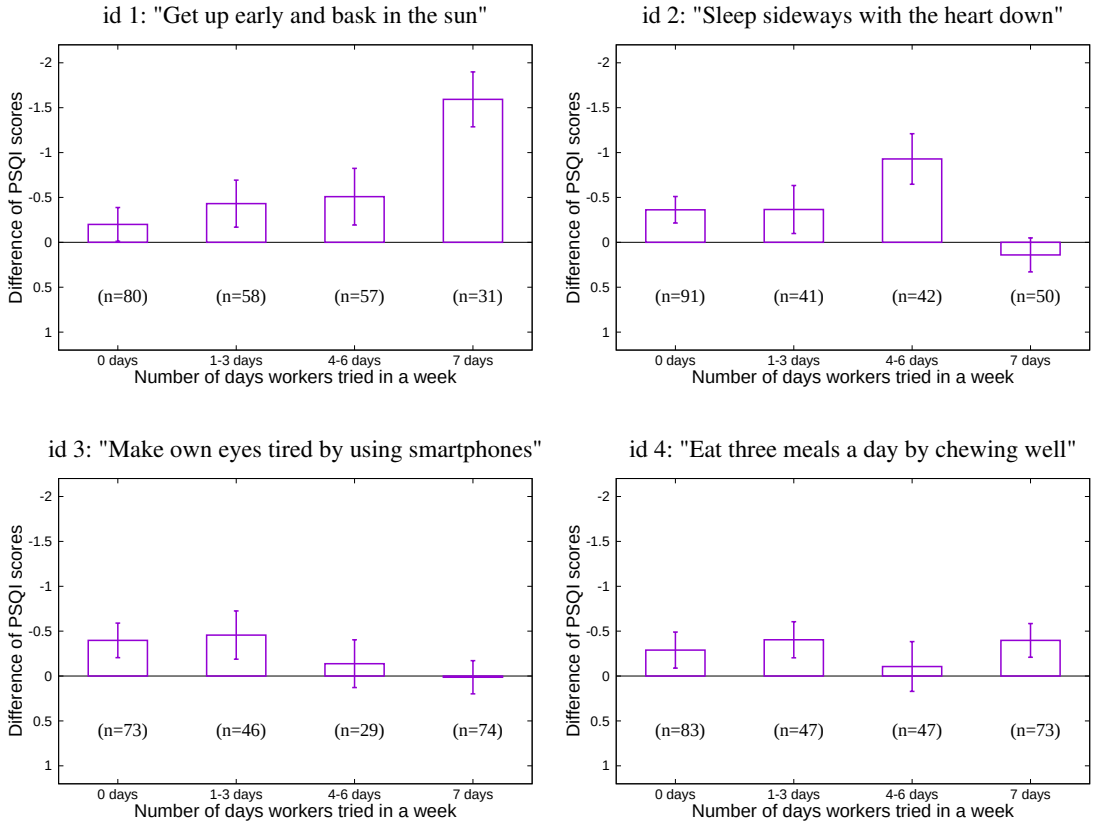

Figure.6.9 Means of differences of PSQI score for each number of days workers tried the hypothesis within a week (with standard error bars). The difference: (PSQI score before the trial) − (PSQI score after the trial)

is considered that there was the effect by placebo and self-selection bias. For the hypothesis of id1, the mean of the PSQI scores improved significantly only in the week when the hypothesis was tested. From this, id1 is considered effective for improving sleep quality. This conclusion is consistent with the majority vote of experts in Table 6.1. In contrast for the hypothesis of id3, the mean of the PSQI scores improved significantly only in the week when the hypothesis was not tested. If there were workers who already corresponds to the contents of a hypothesis, we asked them not to do it in the week and do it as usual in another week. Workers who were assigned the id3 had more workers (42.3% of all) already corresponding to the content than others. Since the sleep quality was improved only when the workers usually doing the content of id 3 stopped it, id2 is considered counterproductive for improving sleep quality. This conclusion is not contradictory to the majority vote of experts in Table 6.1. For the id2 and id4, there was no significant difference between the period of testing the hypothesis and the period of not tried. These hypotheses were also judged ”Neither agree nor disagree” by majority vote of experts in Table 6.1. These results suggested that hypothesis verification by crowd workers in our method may lead to the same results as that by experts.

Still our method has much room to improve. One of them is to decide the heuristic parameters used in the crowdsoucing. When workers vote to others' ideas in crowdsourcing, there is also a study to use weighting process according as their voting history [19]. Those knowledge might be used to select the contents of question (C) in Phase 1 of our study.

Chapter7

Conclusion

This study proposed a method consisting of two phases and evaluated the method in an experiment that actually requests tasks to the crowd in order to discover hypotheses about causality by crowdsourcing toward a higher evidence level than previous works.

The results of Experiment 1 suggested that the crowd could discover hypotheses about causality that were not found by experts, but there is still room for further verification and evaluation on this. One of them is what to compare with experiments by the crowd. This experiment was conducted as a first step to realize *observational study* toward the future under the control with crowd. Individual opinions of the expert ware regarded as comparison objects, but comparing them with results of more reliable results of experiments (such as II-2 in Table 3.1) should be conducted in the future.

The results of Experiment 2 suggested that hypothesis verification by crowd workers in our method may lead to the same results as that by experts. Therefore, our approach might be used to greatly accelerate the existing expert's research or lower its cost, not replacing it. For that reason, further analysis on the accuracy of results and investigation and comparison on costs of various methods will be necessary in the future.

Acknowledgment

I would like to express my gratitude to many people around me.

Prof. Morishima and Prof. Matsubara made a lot of time to discuss the research with me and gave me various useful suggestions despite being very busy. In regular meetings, Prof. Aramaki and Ms. Wakamiya (NAIST) gave me constructive comments and advice from viewpoints of different research fields. I got various opinions from IIIS researchers as experts in sleep research about experiment settings and evaluations. FusionCOMP lab's members always supported me not only in research activities but also in my everyday life. I also appreciate the workers of *Yahoo! crowdsourcing* who participated in my experiments.

I am deeply grateful to all people who supported me. Thank you.

Reference

[1] Eric Hand. Citizen science: People power. *Nature*, Vol. 466, No. 7307, pp. 685–687, 2010.

[2] Caren Cooper. *Citizen Science: How Ordinary People are Changing the Face of Discovery*. The Overlook Press, 2015.

[3] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, et al. Predicting protein structures with a multiplayer online game. *Nature*, Vol. 466, No. 7307, pp. 756–760, 2010.

[4] Chris J Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M Jordan Raddick, Robert C Nichol, Alex Szalay, Dan Andreescu, et al. Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, Vol. 389, No. 3, pp. 1179–1189, 2008.

[5] Melanie Swan. Crowdsourced health research studies: an important emerging complement to clinical trials in the public health research ecosystem. *Journal of medical Internet research*, Vol. 14, No. 2, 2012.

[6] Josh C Bongard, Paul D H Hines, Dylan Conger, Peter Hurd, and Zhenyu Lu. Crowdsourcing predictors of behavioral outcomes. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, Vol. 43, No. 1, pp. 176–185, 2013.

[7] Kirsten E Bevelander, Kirsikka Kaipainen, Robert Swain, Simone Dohle, Josh C Bongard, Paul D H Hines, and Brian Wansink. Crowdsourcing novel childhood predictors of adult obesity. *PloS one*, Vol. 9, No. 2, p. e87756, 2014.

[8] Eiji Aramaki, Shuko Shikata, Satsuki Ayaya, and Shin-Ichiro Kumagaya. Crowdsourced Identification of Possible Allergy-Associated Factors: Automated Hypothesis Generation and Validation Using Crowdsourcing Services. *JMIR Research Protocols*, Vol. 6, No. 5, 2017.

[9] CrowdMed. https://www.crowdmed.com, 2017.

[10] Iain J Marshall Aaron M Cohen Neil R Smalheiser Byron C Wallace, Anna Noel-Storr and James Thomas. Identifying reports of randomized controlled trials (rcts) via a hybrid machine learning and crowdsourcing approach. In *Journal of the American Medical Informatics Association*, 2017.

[11] Luca de Alfaro, Vassilis Polychronopoulos, and Neoklis Polyzotis. Efficient techniques for crowdsourced top-k lists. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*, 2016.

[12] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, Vol. 47, No. 2-3, pp. 235–256, 2002.

[13] Beth Trushkowsky, Tim Kraska, Michael J Franklin, and Purnamrita Sarkar. Crowdsourced enumeration queries. In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, pp. 673–684. IEEE, 2013.

[14] U S Preventive Services Task Force. *Guide to clinical preventive services: Report of the US Preventive Services Task Force*. DIANE Publishing, 1989.

[15] Daniel J Buysse, Charles F Reynolds, Timothy H Monk, Susan R Berman, and David J Kupfer. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. *Psychiatry research*, Vol. 28, No. 2, pp. 193–213, 1989.

[16] Tae-Joon Kim, Byeong Uk Lee, Jun-Sang Sunwoo, Jung-Ick Byun, Jangsup Moon, Soon-Tae Lee, Keun-Hwa Jung, Kon Chu, Manho Kim, Jong-Min Lim, Eunil Lee, Sang Kun Lee, and Ki-Young Jung. The effect of dim light at night on cerebral hemodynamic oscillations during sleep: A near-infrared spectroscopy study. *Chronobiology International*, pp. 1–14, 2017.

[17] Iris Rawtaer, Rathi Mahendran, Hui Yu Chan, Feng Lei, and Ee Heok Kua. A nonpharmacological approach to improve sleep quality in older adults. *Asia-Pacific Psychiatry*, 2017.

[18] Rinske A. Gotink, Paula Chu, Jan J. V. Busschbach, Herbert Benson, Gregory L. Fricchione, and M. G. Myriam Hunink. Standardised mindfulness-based interventions in healthcare: An overview of systematic reviews and meta-analyses of rcts. *PLOS ONE*, Vol. 10, No. 4, pp. 1–17, 2015.

[19] Manas S. Hardas and Lisa Purvis. *Bayesian Vote Weighting in Crowdsourcing Systems*, pp. 194–203. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.