Field Experiment on Incentive of Compilation and Behavioral Evaluation in Peer Review System

Takuya Nakaizumi, Mitsuharu Watanabe *

December 31, 2009

Abstract

We set up a Web Bulletin Board System [1] (BBS) in the U.S. in 2008 and in Japan in 2007 and experimented whether an incentive

*Takuya Nakaizumi, Associate Professor, Kanto Gakuin University, and Visiting Scholar, UC San Diego International Relations and Pacific Studies, tnakaizumi@ucsd.edu, Mitsuharu Watanabe, Professor, Kanto Gakuin University, light@kanto-gakuin.ac.jp : 1-50-1 Mutsurahigashi, Kanazawa-ku, Yokohama-city, 236-8501, Japan, Preliminary version, We are grateful to Masahiro-Okuno Fujiwara, Hideshi Itoh, Tatsuo Tanaka, Dan Sasaki, Masuyuki Nishijima for helpful comments and Lubor Jelinek for English proof reading. All mistakes are belonging to Authors. The support of the Grant-in-Aid for Scientific Research (B) from the Ministry of Education, Culture, Sports, Science and Technology in Japan and The Japan Securities Scholarship Foundation are also gratefully acknowledged.

[1] A Bulletin Board System, or BBS, is a computer system running software that allows users to connect and log in to the system using a terminal program. Once logged in, a user can perform functions such as uploading and downloading software and data, reading

affected the editing of the conversation in the BBS, when the editing was evaluated by all the other participants and when the rewards of the editor depended on the evaluation of the participants. We provided two kinds of rewards, one of which was lower, the other was higher and also provided two kinds of topic one was difficult to edit and the other easy. The difficulty was based on the variance of the opinions among the participants. The results of the experiments were counter-intuitive and thought provoking, that is, from orthodox incentive theory, an easy topic with higher rewards should lead to the highest score and a difficult topic with lower rewards should lead to the lowest and the evaluation score of an easy topic with lower rewards and a difficult topic with a higher rewards should be between the first two. However the result of our experiment was that a difficult topic with lower rewards led to highest score and a difficult topic with higher rewards led to the lowest score. And the score of an easy topic with lower rewards and an easy topic with higher rewards are between the first two. By using total number of letters of editing as a proxy variable of editor's effort, effort is shown to be higher when rewards are higher. These counter-intuitive results are shown to be mainly based on the evaluation behavior by other participants.

news and bulletins, and exchanging messages with other users, either through electronic mail or in public message boards. Many BBSes also offer on-line games, in which users can compete with each other, and BBSes with multiple phone lines often provide chat rooms, allowing users to interact with each other.

1 introduction

Everyone has a strong desire to communicate and express his or her own opinion. And editors or writers have their own opinion. They seldom adopt the opinions of others into their writing or compilation when they do not agree with them or do not like it. For example, it is very hard for the editor to adopt and reflect the opinions that they do not want to when they edit the entire conversation in a Web Bulletin Board System (BBS). We may be glad that our opinions are reflected in some composition or writings. It is natural for the participants to want their opinion to be reflected in the editing. Presumably they may appreciate the editing more, if it reflects their opinion.

However, if the editing is conducted with the intent to flatter the participants, it is more difficult and requires much more effort, especially when there are conflicts of opinion between the editor and some participants. Even in such circumstances, incentives may make him or her edit in a more flattering way to the other participants against the editor's original intent.

The famous web information system, Wikipdia, is a completion of some information. However without rewards, it cannot always prepare adequate information. Thus we must set an adequate incentive system for writing. Our primary concern is how to prepare incentives to write or edit.

If the editing is conducted with the intent to flatter the participants, however, it is quite difficult to give an incentive of adequate evaluation. Without incentive, in some cases evaluator might not make a proper evaluation and could not give a proper incentive of editing, while in other cases, evaluators may not need to make an specific effort to evaluate because all the evaluators

should do is only the truthful revelation of their preference. In many specific models, these preference revelation is assumed to be done honestly if it is costless and without breaking IC. condition. In natural setting, there are also many cases in which there is no incentive of evaluation. For example, juror would not be blamed even if they made a mistake to judge.

We experiment how the evaluator make an evaluation without incentive, and whether such an evaluation gives an proper incentive to editing by a field experiment on the BBS. We set up a Web Bulletin Board System (BBS) in the U.S. in 2008 and in Japan in 2007 and experimented whether an incentive affected the editing of the conversation in the BBS when the editing was evaluated by all the participants and the rewards of the editor depended on the evaluation of the participants.

We can verify this by a field experiment on the BBS. We set up a Web Bulletin Board System (BBS) in the U.S. in 2008 and in Japan in 2007 and experimented whether an incentive affected the editing of the conversation in the BBS when the editing was evaluated by all the participants and the rewards of the editor depended on the evaluation of the participants, because the editing should reflect the conversation that participants made in each thread in BBS.

We will explain the experiment briefly below. We conducted an online experiment using a representative sample of the nations' adults recruited by a partner research firm. All the experiment participants were gathered randomly. After completing a pre-test survey, participants were directed to a Web message board devoted to discussing just one topic. The topic belonged to one of the following domains: investment, marketing, religion and politics,

(investment and marketing were in Japan only) . The participants were registered to a domain according to their preference.

In addition we choose two topics of each domain, one of which was the most controversial and the other of which was the least controversial among the several topics of each domain. The controversial level was based on the variance of participants' opinions that were answered in pre-survey. The most controversial topic was considered to be the most difficult to sum up while least controversial was considered the easiest to sum up.

While they were discussing a specific topic, SIS "social influence scores" were calculated for each participant which measured community acceptance and acclaim. After the conversation, the participant with the top-ranking SIS was automatically granted editorial privileges which allowed them to summarize the discussion.

After the conversation was edited by the editor, who was the participant with the highest SIS, they were rewarded according to the evaluation of the editing by all the other participants, that is, if the editor was rated less than 'not satisfactory' (A -2 out of a scale of 7(-3 to 3)) on average, then the bonus for editing would be reduced by half. The 7-point scale ranged from ' Strongly Disagree' to 'Strongly Agree'. The procedures of the experiment were fully disclosed to all the participants.

All the participants have a conversation trying to impress others so that they can be the editor and get more rewards. If they become an editor, it is natural that they try to edit the conversation to flatter others in order to get more rewards even if they do not like to edit in such a way.

Our initial hypothesis was that if the rewards were higher, they would

try to flatter other participants more, even if they were reluctant. And these editor's effort make the evaluator to give higher evaluation scores honestly even if they have no incentives. This would result in a higher actual evaluation if the rewards were higher. We also considered the difficulty to edit, that is, if the editing required more effort, this would result in a lower actual score. To check this, we had two types of BBSs, one in which rewards were lower for the editor and the other in which the rewards were higher. And we also provided two kinds of topic: one was difficult to edit and the other was easy. We tested whether the evaluation of the editing tended to be higher when the rewards were higher with the easier topic.

From orthodox incentive theory, an easy topic with higher rewards should lead to more effort and it is easy for the editor to achieve the highest score, while a difficult topic with lower rewards lead to the lowest score. And the evaluation score of an easy topic with lower rewards and a difficult topic with higher rewards should be between the first two. However the result of our experiment was far from that and thought provoking. The difficult topic with lower rewards led to the highest score and the difficult topic with higher reward led to the lowest score. And the score of the easy topic with lower rewards and the easy topic with higher rewards were between the first two.

To our best knowledge, this is the first field experiment on editing incentive on editing using web thread. Although there are some papers about the experiments of incentive in Lab such as Gneezy and Rustichini[2000a,b] and Gneezy [2003], there are two main additional contributions of our paper. One is that as a field experiment, we check the incentive of editing in a more realistic situation. As pointed out in the excellent survey by Harrison and

List [2004], "Our primary point is that dissecting the characteristics of field experiments helps define what might be better called an ideal experiment, in the sense that one is able to observe a subject in a controlled setting but where the subject does not perceive any of the controls as being unnatural and there is no deception being practiced".

As a result of multi-disciplinary collaboration between an economist, a computer scientist mathematical sociologist, and with the assistance of market research companies who were responsible for paying the rewards to our experiment's participants, we succeeded in creating an environment that closely models the real world. In our experiment the goal of all the participants was not to participate in an experiment of incentive for getting credits but to sum up their opinions as the results of the communication in the BBS and also to be paid rewards (hopefully with a bonus) from the research companies. This means they are motivated by both personal interest and a monetary incentive. It seems obvious that this situation approximates reality more than a classroom where professors control their students with extra credits.

Our second contribution is that we show that the participant exhibit a behavioral nature as evaluators of the editing. In terms of the incentive for evaluator, we did not pose incentive because it is just subjective preference revelation without any need to manipulate and there are many cases in which no incentive is posed to evaluation. And we test some kinds of reciprocity, experimenting how the participants evaluate and the evaluation given any incentive for editors.

By using total number of letters of editing as a proxy variable of editor's effort, effort is shown to be higher when rewards are higher. And no editor's

bonus was reduced. This shows that editor thought that even those evaluations without incentive should prepare some adequate performance measure and give them incentives.

And the main result of the paper is that average evaluation score shows an counter-intuitive and thought provoking results. From orthodox incentive theory, an easy topic with higher rewards should lead to the highest score and a difficult topic with lower rewards should lead to the lowest and the evaluation score of an easy topic with lower rewards and a difficult topic with a higher rewards should be between the first two. However the result of our experiment was that a difficult topic with lower rewards led to highest score and a difficult topic with higher rewards led to the lowest score. And the score of an easy topic with lower rewards and an easy topic with higher rewards are between the first two. Although even if we try to apply some behavioral economics theory, it is not easy to explain. For example, if the fairness behavior of evaluators was considered, the highest evaluation score might be a difficult topic with lower rewards while easy topic with higher rewards lead to the lowest evaluation score. The result of the experiment was consistent in that a difficult topic with lower rewards led to the highest score but inconsistent in that a difficult topic with higher rewards led to the lowest score. The other example is spiteful behavior whereby participants evaluate more severely when the editor got higher rewards. Even if the spiteful behavior of evaluator is considered, the highest score might be an easy topic with lower rewards while a difficult topic with higher rewards lead to the lowest. The result of the experiment was consistent in that a difficult topic with higher rewards led to the lowest score but inconsistent

in that a difficult topic with lower reward led to the highest score. Such behavioral evaluation by the participants should be considered in a real world organization. We can apply such results to organizational design where there are many occasions similar with the editor's situation in the experiment.

Finally, we set 72 threads and chose 72 editors. But only 44 editors edited actually and other 28 editors did not edit and not get the editing rewards automatically. But we still ask the participants to evaluate the 28 editors who do not editing although it does not affect the rewards and participants do know that.

The average evaluation score of the editors without editing differs in Japan and US. Namely, while the average evaluation score without editing in the US is almost the same as with editing in the US, the average evaluation score without editing in Japan is apparently lower than with editing in Japan. The results are also thought-provoking because it anticipate that national identity make the difference in such a off-path situation.

Next section we explain the experiment in detail. In section 3, we will introduce the benchmark model. In section 4, we will describe the main results of the experiment and We refer the average evaluation score of the editors without editing in Section 5 and Section 6 is concluding remarks in which we will also sum up the implications.

2 Procedure of the Experiment

In this section we will describe the procedure of the experiment in detail. We set up a Web Bulletin Board System (BBS) in the U.S. in 2008 and

in Japan in 2007. We can divide the experiment into 4 stages. At first the participants were gathered and a pre-survey was conducted. In the next stage, participants were registered to a BBS and assigned a s particular forum and to discuss a specific topic in the BBS. After the conversation was over, an editor was chosen among the participants and they edited the conversation in the BBS in the third stage. In the final stage, when the editing was completed, the editor's rewards were paid according to the evaluation by the other participants.

2.1 0 stage: pre-survey

Before starting the BBS, we made a pre-survey of the attitude of all the participants as to which topic they would prefer to discuss.

We hired a market research company to manage the BBS, recruitment, selection of the participants, compensation and the reward process. As part of their recruiting process, participants, who had to be more than 18 years old, were first asked their interest in four domains in Japan, religion, politics, marketing, and finance and two domains in the U.S., religion and politics.

They were assigned to a specific domain out of the four domains in Japan and two in the U.S. Although we limited the number of participants in each domain, we attempted to assign them to the domain they were most interested in as much as possible.

All the participants were asked to fill out a pre-survey about their basic information and their opinions on the specific domain to which they were assigned. An example of a question on the basic information section would be: "in the past 12 months, have you used the Web to look for information

about a product that you might want to buy? If so, how often?" They did not need to complete the survey in order to be paid a reward (which will be described later). They might choose "Not prefer to answer" to any of the questions or items in the survey. The survey took approximately 15-20 minutes to complete.

In the pre-survey, all the participants were also asked their opinions on the specific domain to which they were assigned. They were asked their opinion regarding some topics of the domain.

We could ascertain each participant's viewpoints on each topic and then assess whether they were likely to agree and disagree with other participants.

From some questions in the pre-survey, we could pick the most controversial topic, in which participants' opinions of the topic were the most likely to be diverse and the least controversial topic, in which the participants' opinions were least likely to be diverse. We call the most controversial topic the most difficult and the least the easiest. The participants were assigned either the most difficult or easiest topic. For example, in the religion domain in the U.S. the most variant topic was "The Bible is the word of God and is to be taken literally, word for word". The variance of the topic was 2.33 (average is 0.02 on a scale of 7(-3 to 3)). It had the highest variance among all the topics.

2.2 1st stage: Discussion on BBS

All the participants were requested to carry out a debate on one thread in the bulletin board, writing at least one post per week during a period of four weeks (with a total of at least four posts to their assigned forum domain by

the end of the four week period). They were rewarded $10.00 in the U.S. and 1,000 yen in Japan, if they contributed more than four times.

Each post had to contain over 100 words and was expected to take about 10 minutes to complete. They needed to write a total of at least four posts in the forum domain assigned to them in order to be paid a reward (more on rewards below). They could have logged on and off the forum to participate as many times as they liked during the four-week period. At the beginning of the discussion each group contained 10 members in Japan and 20 members in the U.S. But all the participants did not continue by the end of the experiment, because they did not post more than four times.

In the BBS, the observer calculated the "Social Influence Score"(SIS) of each participant based on the total combined evaluation of each particular post. In addition, "Social influence score" was automatically calculated by a software algorithm according to the level of social interaction within a debate. Each participant's social influence score was based upon evaluations from other participants. For example, if one of their posts received a reply, their social influence score increased. If someone rated one of their posts or their profile, their score changed. Thus, they should expect their social influence score to fluctuate throughout the experiment as they continued to contribute to the forum. We use SIS for mainly two reasons. One is that when comparing with a simple voting rule, SIS could reduce the chance of collusion among specific groups in the forum. The other is that we would like to conduct the experiment in more realistic way such as the Google search system. Like in Google, the SIS increases more when evaluations are made by participants with a higher SIS, as any other social resource, more of it is

distributed by those who have more social influence. Then it shows long tail distribution.

2.3 2nd stage: editing by the editor Chosen According to the SIS

After completion of the BBS, the participant with the highest SIS score was chosen as an editor and to summarize the main points of the domain's debate using the editing software function of the site. The editing was a report detailing the discussion of a particular forum topic of a particular domain.

The editor received a bonus reward of $20.00 or $80.00 in the U.S. and of 2,000 yen or 8,000 yen in Japan for her or his effort. The exact amount of the bonus was decided based on the group s/he belonged to as an experimental condition. S/he was assigned to a group at the beginning of the experiment and they had a 50-50 chance of belonging to either the $20 (2,000 yen) reward group or the $80(8,000 yen) reward group. All members of a group were notified of the amount when they started the discussion. If the others' assessment of the editing was low, then the reward was reduced in both cases as will be explained below [2].

[2]Although only the individual selected as the editor received the bonus reward, all the group members other than the selected editor had the option of summarizing discussions in another setting of the same software module. However there was no use of summarizing other than rewarded editor

2.4 3rd stage: Evaluation of the editing and rewarded

When all the participants completed participating in the forum, they were required to fill out a final survey (post-survey) about their experiences with the experiment and evaluate the editor. An example of a question from this survey was: "Are you satisfied with the information provided by other users? " They had the option to choose "Not prefer to answer" to any of the questions or items in the survey.

The editor's rewards were paid according to the evaluation. Namely, if the average of score of all the other participants' evaluation of the editing was less than "not satisfactory" (the second worst on a scale of 7), then the bonus for the editing was reduced by half. We should note that there was no monetary incentive to evaluate the editing.

A summary of the experiment conditions:

- **Domain:** Economics: Marketing, Investment (in Japan only)

  Non-Economics: Politics, Religions

- **Difficulty level of topic:** based on variance of opinion
  - **High:** The topic with the most variant opinion
  - **Low:** The topic with the least variant opinion
- **Amount of reward to the editor:**
  - **High:** $80 (8,000 yen in Japan)
  - **Low:** $20 (2,000 yen in Japan)

- **Number of participants in each conversation:** 10 at the beginning in Japan and 20 in the U.S.

- **Total number of conversation groups:**

  48 domains in Japan and 24 in the U.S.

  In Japan, 4 domains, 2 topics, 2 types of rewards, which means 16 forums each forums contained 3 separate groups, thus 48 groups. In the U.S., 2 domains, 2 topics, 2 types of rewards, which means 8 forums each forums contained 3 separate groups, thus 24 groups, and a total of 72 groups and chose 72 editors although only 44 editors edited actually and other 28 editors did not edit and not get the editing rewards automatically. But we still asked the participants to evaluate those editors who do not edit in spite that participants do know that it does not affect the rewards. We got the 21 editor's evaluation scores.

- **Total number of participants:** 480 at the beginning in Japan and in the U.S., thus a total of 960 participants.

Table 1: **BBS**

Country	**Rewards**	**Domain**	**Diffiulty**	**Total BBS**
Japan	2	4	2	48
The U.S.	2	2	2	24
Total	2	2 or 4	2	72

3 Basic Model of the Experiment

First set up the model to clarify the hypothesis of the experiment. The experiment can be divided into 4 stages as we explained in the previous section. In the first stage, after the 0-stage pre-survey, participants of the BBS discuss of the specific topic in which the "social influence scores" (SIS) of each participants is calculated. After the conversation is over, an editor who has the highest SIS is chosen among the participants.

In the second stage, the editor edits the conversation in the BBS. In the third stage, after the completion of the editing, all the other participants evaluate the editing. And the evaluation affects the rewards of the editor.

For simplicity we assume that the quality of the editing by editor 0 can be expressed by the integer, X_0.

There are $N+1$ participants where $i = 1, ..., N$. The editor is notated as 0. All the other participants would evaluate the editing based on their own preferences.

We give no incentive for the other participants when evaluating the editing. From standard economic theory, they should reflect only their actual preference for the editing on their evaluation. Our initial assumption is that their preferences would depend on only the quality of the editing.

Assumption 1 ***Evaluation of Editing***

The preference of other ith participants for the editing X_0 would be expressed as an utility function that depends only on the quality of editing, $u_i(X_0)$ And the evaluation score of editor 0 by the other participants i de-

pends on their preference,

$$s_0^i = s(u_i(X_0))$$

which is non decreasing function of X_0

And s_0, the evaluation score of editor 0 by all the other participants i is average of their scores in the final stage. Thus in the second stage, when editor edit the conversation, the editor does not know the evaluation of the other participants. Thus he would make an effort to infer the evaluation score. We consider the uncertainty as the inference error of the editor and simplify the uncertainty by adding aggregated error term ϵ to s_0. And we assume that cummulative distribution function of ϵ is $F()$ and distribution function of ϵ is $F'() = f()$ and $E[\epsilon] = 0$. Thus s_0, the evaluation score of editor 0 by all the other participants i is

$$s_0 = \frac{1}{n}\sum_{i=1}^{n} s_0^i = \frac{1}{n}\sum_{i=1}^{n} s(u_i(X_0)) + \epsilon = r(X_0) + \epsilon$$

The editor's utility is $u_0(X)$ and the editor's ideal editing can be expressed as $\hat{X} = \arg\max u_0(X)$. We also define the well behaved metric $\rho(\hat{X}, X_0)$ between X_0 that is a quality of actual editing and $\hat{X}$ that is a quality of ideal editing for him. And we define $\hat{X} = 0$, namely, there exists the well behaved real number x whereby $x = \rho(0, X_0)$. Thus $s_0 = 1/n \sum_{i=1}^{n} s(u_i(x)) + \epsilon = r(x) + \epsilon$. [3]

The cost of editing is assumed to be $\alpha c(x)$ ($\alpha > 0$), because the ideal editing for him should cost zero , $c(0) = 0$. And $c'(x) > 0$, $c''(x) > 0$.

[3]We omit small 0 because of simplifying.

In the third stage, other participants evaluate the editing and the evaluations determine the reward W of the editor. The rewards function is as follows:

$$W(s_0) = \begin{cases} w & \text{if } (s_0 \geq \hat{s}) \\ w/2 & \text{if } (s_0 < \hat{s}) \end{cases} \tag{1}$$

And as for the domain of ϵ, we introduce following assumption.

Assumption 2 ***domain of the distribution function of*** ϵ

for any $\epsilon \in [\underline{s}, \overline{s}]$, $f() > 0$. $\underline{s}$ *is the least score while* $\overline{s}$ *is the best score in the experiment.*

That means the uncertainty is large enough for the editor to make an effort to edit for any cutoff score $\hat{s}$ in order to maximize her or his net payoff while facing uncertainty as to the evaluation by other participants.

With assumption 1and 2, her or his net expected payoff with x is $E[B(x)]$ that is as follows,

$$E[B(x)] = E(W(s_0)) - \alpha c(x) = E(W(r(x) + \epsilon)) - \alpha c(x)$$

Then we derived the following Proposition 1.

Proposition 1 *With Assumption 1 , a higher* w *should lead to a higher* s*. And a higher* α *should lead to a lower* s

Proof*: with assumption 1, the editor maximizes her expected utility by choosing* $\hat{x} = \arg\max E[B(x)]$*. From (1),*

$$\begin{aligned} E[B(x)] &= wP(s_0 \geq \hat{s}) + w/2P(s_0 < \hat{s}) - \alpha c(x) \\ &= wP(\epsilon \geq \hat{s} - r(x)) + w/2P(\epsilon < \hat{s} - r(x)) - \alpha c(x) \\ &= w - w/2F(\hat{s} - r(x)) - \alpha c(x) \end{aligned}$$

Thus

$$\frac{dE[B(x)]}{dx} = w/2f(\hat{s} - r(x))r'(x) - \alpha c'(x) = 0$$

From Assumption 2, $f(\hat{s} - r(x)) > 0$ *and* $r'(x) > 0$. *There is interior solution of* x. *Thus* $\hat{x}$ *of the solution is an non-decreasing function of* w *and non-increasing function of* α. *Thus* $s_0 = r(x) + \epsilon = r(x(w, \alpha)) + \epsilon = s(w, \alpha)$ *is also an non-decreasing function of* w *and non-increasing function of* α. *(Q.E.D.)*

This proposition shows that a higher incentive tends to bring a higher actual evaluating score. The next section we will describe the results of experiment and test the hypotheses.

4 Results of the experiment

In this section we will show the results of the experiment and test the hypothesis derived from Proposition 1.

The most important data the experiments provided is based on data from 322 participants in Japan and 347 participants in the U.S. all of whom completed their mission in the experiment. Each participant belonged to one of 48 domains in Japan and 24 domains in the U.S. But only 44 editors edited actually and other 28 editors did not edit and not get the editing rewards automatically. But we still ask the participants to evaluate the 28 editors who do not editing although it does not affect the rewards and participants do know that. And we got the 21 editor's evaluation scores.

Table 2: **Number of Participants**

Country (Total BBS)	**each BBS**	**Total without editing**	**Total with editing**
Japan (48)	10	97(5)	225(31 editors)
The U.S. (24)	20	172(16)	175(13 editors)
Total (48 or 24)	10 or 20	269(21)	400(44 editors)

Proposition 1 predicts that higher rewards should provide a higher evaluation score while a higher for the editor should lead to a lower evaluation score. Thus we derive following hypothesis.

Hypothesis 1 *An easy topic with higher rewards should lead to the highest score and a difficult topic with lower rewards should lead to the lowest and the evaluation score of an easy topic with lower rewards and a difficult topic with higher rewards should be between the first two.*

To clarify the experiment's results, we will show the average score of editing evaluation of each domain in table 3-7. And we can apply a Generalized Wilcoxon Test (GWT) in a nonparametric test to all the evaluation scores of editing of each participant. A Generalized Wilcoxon Test is applied in the case of a missing sample among the score data, which occurred because not all of the participants continued until the end of the experiment. It is applied to all the evaluation score data of each participant including those with missing data.

Table 3 shows the average score of editing evaluation based on total samples who continued to the end. If the difference was significant at the 5%

level in the test, we noted it with a asterisk. And we show the P-value of (GWT) in brackets. The results of the experiments were far from what we had expected and thought provoking. An incentive had a counter-related effect on the editing and higher rewards led to less positive evaluations than lower one.

Table 3 Average Evaluation Score

	Reward for editing		Difference
	20	80	
average	1.129	0.673	-0.456^{*}
standard error	1.254	1.586	
sample	205	151	

Table 4 shows the average score of editing evaluation based on both rewards and difficulty. The result shows that the difficult topic with lower rewards led to highest score (1.23) and the difficult topic with higher rewards led to the lowest score (0.564). And the score of the easy topic with lower rewards (0.972) and the easy topic with higher rewards (0.88) are between the first two. The difference between the score of the easy topic with lower rewards (0.972) and the difficult topic with higher rewards (0.564) shows significance at the 5% level in the GWT test. P-value of (GWT) is 0.007.

Table 4 Average Evaluation Score for Both Easy and Difficult and high and low reward

Difficulty of topic	Reward for editing 20	80	Difference
difficult(average)	1.23	0.564	-0.67^{*}
(standard error)	1.157	1.576	(0.007)
(sample)	121	101	
easy (average)	0.972	0.88	-0.092^{*}
(standard error)	1.39	1.59	(0.02)
(sample)	84	50	
Difference	-0.157	0.316^{*}	
	(0.16)		

Thus we get the following result 1

Result 1 *From the results of the experiment, we can reject the initial Hypothesis.*

Then we applied the behavioral theory to the evaluation activity. First we consider the spiteful behavior on part of the other participants. By spiteful behavior we define that the evaluation of the editor with higher rewards became much harder than that with lower rewards. Then we should check following hypothesis

Hypothesis 2 *Consider the spiteful behavior on part of the other participants, the easy topic with lower rewards should lead to the highest score and the difficult topic with lower rewards lead to second highest and the easy topic with a higher rewards should be the third and the difficult topic with higher rewards should lead to the lowest.*

However from table 4, the difficult topic with lower rewards led to highest. Thus we get the following result 2

Result 2 *From the results of experiment, the difficult topic with lower rewards led to highest, which contradict the Hypothesis 2 that predict that the easy topic with lower rewards should lead to the highest score. Thus we can reject the second Hypothesis.*

Finally, we consider fairness behavior on part of the evaluators. By fairness evaluation, we define that the evaluation of the editor with difficult topic with lower rewards should be much more appreciated than the easy topic with higher rewards. Then we should check following hypothesis.

Hypothesis 3 *Consider fairness behavior of the other participants, the difficult topic with lower rewards should lead to the highest score and the easy topic with higher rewards should lead to the lowest.*

However from table 4, the difficult topic with higher rewards led to the lowest. Thus we get the following result 3

Result 3 *From the results of experiment, the difficult topic with higher rewards led to the lowest, which contradict the Hypothesis 3 that predict that easy topic with higher rewards should lead to the lowest score. Thus we can reject the third Hypothesis.*

We consider the effort of the editor. By using total number of letters of editing as a proxy variable of editor's effort, effort is shown to be higher when rewards are higher. And no editor's bonus was reduced. We show that total

number of letters of editing is increasing function of rewards. And all editors who write an editing got a full rewards.

Table 5 Average total number of letters of editing

Difficulty of topic	Reward for editing 20	80	Difference
difficult(average)	880.3	1307.5	427.2
(standard error)	711.7	1463.0	
(sample)	13	12	
easy (average)	1053	1791.9	738.9
(standard error)	859	1345	
(sample)	9	9	
Difference	172.7	484.3	

Although those difference does not show significant difference with GWT test because of the small sample. Difference between 20 with difficult and 80 with easy is 911.5 and show the significant difference with $p-$value is 0.025.

We conclude that these counter-intuitive results might be attributed to be mainly based on the evaluation behavior by other participants, although we still try to develop better proxy variable of a effort.

5 Difference of evaluation between Japan and the US participants

We set 72 threads and chose 72 editors. But only 44 editors edited actually and other 28 editors did not edit and not get the editing rewards automat-

ically. But we still ask the participants to evaluate the 28 editors who do not editing although it does not affect the rewards and participants do know that.

The average evaluation score of the editors without editing differs in Japan and US. Namely, while the average evaluation score without editing in the US is almost the same as with editing in the US, the average evaluation score without editing in Japan is apparently lower than with editing in Japan. We think that the results are also interesting because it anticipate that national identity makes the difference in such off-path situation.

Table 6 Evaluation Score with and without editing

	With editing	Without editing	Difference
Japan (average)	0.75	0.28	0.471
(standard error)	1.41	1.69	
(sample)	194	92	
the U.S. (average)	1.18	1.15	−0.02
(standard error)	1.41	1.48	
(sample)	162	156	
Total	0.93	0.77	

6 Concluding Remark

Communication is one of the basic activities of human beings and it is essential for organizational activities. The experiment to check whether incentive works where the communication is essential. And it reflects an real situation of organizational activity especially considering peer reviewed evaluations.

We provided two kinds of rewards, one of which is lower, the other is higher. The results of the experiments were far from what we expect. And It was extremely interesting. When opinions for the topic were easy to sum up, rewards did not affect the evaluation. when opinions for the topic is difficult to sum up, on the other hand, rewards affected the evaluation significantly. But the results were counter-intuitive, that is, higher rewards brought fewer positive evaluations.

There sometimes happen that member of the organization must obey an decision of the organization that s/he does not want to. In such a situation, more rewards may bring an more conflict among the members in the organization. The experiments show that we should be more careful of organization design, especially for peer reviewer.

References

[1] Burger, Nicholas, Gary Charness and John Lynham, [2008] , "Three Field Experiments on Procrastination and Willpower" UCLA discussion paper

[2] Bandiera,Oriana, Iwan Barankay and Imran Rasul, [2005], "Managerial Incentives in Hierarchies: Evidence From a Field Experiment " (mimeo)

[3] Dewatripont, M.,and J. Tirole, [2006] ,"Modes of Communication", Journal of Political Economy

[4] Fehr, Ernst and Simon Gachter, [2002]"Altruistic punishment in humans," Nature, 415 , 137-140.

[5] Fehr, Ernst and Simon Gachter, [2000]"Cooperation and punishment in public goods experiments," American Economic Review, 90 , 980-994.

[6] Fehr, Ernst and John A. List, [2004] "The hidden costs of incentives - trust and trustworthiness among CEOs," Journal of the European Economic Association, 2 , 743-727.

[7] Fehr, Ernst and Klaus M. Schmidt, [1999] "A Theory of Fairness, Competition, and Cooperation," Quarterly Journal of Economics, 114 3 , 817-868.

[8] Gachter, Simon and Benedikt Herrmann, [2006] "The Limits of Self-Governance in the Presence of Spite: Experimental Evidence from Urban and Rural Russia," Institute for the Study of Labor (IZA), Research Paper Series, No. 2236.

[9] Gneezy, Uri and Aldo Rustichini, [2000a],"A fine is a price" ,Journal of Legal Studies,Vol.29,No.1,pp.1-17

[10] Gneezy, Uri and Aldo Rustichini [2000 b], "Pay enough or Don't pay at all" ,The Quarterly Journal of Economics,Vol.115,pp.791-810.

[11] Gneezy, Uri [2003] , "The W effect of incentives" ,University of Chicago working paper

[12] Harrison, G. W. and List, J. A. [2004], "Field experiments", Journal of Economic Literature 42, 1009 55.

[13] Landier, Sraer and Thesmar, [2008], "Optimal Dissent in Organizations, " (forthcoming) Review of Economic Studies

[14] Levine, David K., [1998] "Modeling Altruism and Spitefulness in Experiment," Review of Economic Dynamics, vol. 1(3), 593-622, July.

[15] Plott, Charles R. and Vernon L. Smith (edit) [2008], "Handbook of Experimental Economics Results" , Volume 1, North Holland

[16] Reiley, David H. and John List, [2008],"Field Experiments in Economics", 'The New Palgrave Dictionary of Economics' 2nd edition, edited by S.N. Durlauf and L.E. Blume. Palgrave Macmillan

[17] Saijo, Tatsuyoshi and Hideki Nakamura [1995], "The "Spite" Dilemma in Voluntary Contribution Mechanism Experiments" Journal of Conflict Resolution, Vol. 39, No. 3, 535-560

[18] Takuya Nakaizumi, Watanabe Mitsuharu,[2008] "Field Experiment on Incentive of Communication and Compilation and Evaluation", 5th Japanese-German Frontiers of Science Symposium 2008, 30 October - 2 November 2008 in Mainz, Germany by The Alexander von Humboldt-Foundation and the Japan Society for the Promotion of Science

[19] Takuya Nakaizumi, Watanabe Mitsuharu,[2009] "Field Experiment on Incentive of Communication and Compilation and Evaluation", 2009 Pacific Rim Conference of Western Economic Association International , March 24-27, 2009 Ryukoku University, Kyoto, Japan.

[20] Watanabe Mitsuharu, Takuya Nakaizumi, Noboru Sonehara,[2008] "Distribution of social resources in a community of dialogue and editing.-An integrated system for values and knowledge on the internet -" Internet

Research 9.0: Rethinking Community, Rethinking Place, IT University of Copenhagen, October 15, to18