Punishments for the Common Good: Increasing Motivation for Content Evaluation using Reputation Systems with Penalties

Marta Juźwin, Paulina Adamska, Adam Wierzbicki, Grzegorz Kowalik

Polish-Japanese Academy of Information Technology, ul. Koszykowa 86, 02-008 Warszawa, Poland

marta.juzwin@pjwstk.edu.pl, paulina.adamska@pjwstk.edu.pl, adam.wierzbicki@pjwstk.edu.pl, grzegorz.kowalik@pjwstk.edu.pl

**Abstract**

Using penalties in reputation systems associated with online platforms is an uncommon way to increase users’ effort. Here we show, based on experiments and a social simulation model, that penalties can enhance both users’ effort and accuracy of results. First, we asked online platform users to evaluate the credibility of web pages, previously modified and evaluated by an expert, concerning the Java programming language. Users received reputation points for each rating: the closer their evaluation was to the expert’s evaluation, the greater their reward. In the control group, users could not lose reputation points; in the treatment group, users lost points when their ratings were far from the expert’s. The comparison of results in the treatment and control groups confirmed the hypothesized positive effect of penalties on users’ effort and accuracy. Next, we created a social simulation model to verify that penalties have a positive impact on users’ effort in scenarios with different payoff schemes, as well as in a scenario when only 25% of Webpages had expert ratings. Both sets of results prove penalties have a positive effect on users’ effort and accuracy. From the model, which analyzed a situation whereby a user can choose not to participate in evaluation, we learned that penalties can have an adverse effect on participation.

Keywords: crowdsourcing work evaluation, quality control, reputation

1. **Introduction**

Reputation systems are widely used mechanisms for evaluating a user’s services or work and for motivating users to provide high-quality input. Examples of such usage include reputation systems for web content evaluation, such as in popular Q&A systems (Stack Overflow, Yahoo Answers, etc.). Reputation is also used in crowdsourcing to evaluate workers [1]. Despite recent theoretical results that indicate the usefulness of such an approach [2], penalties are rarely used in reputation systems. While popular sites such as Stack Overflow allow the use of downvotes, in order to use one, a user has to sacrifice some reputation points, and thereby downvotes are rarely used by community members. In this paper, we consider the designs of reputation systems that use penalties. Basing on theoretical results, we hypothesized that penalties would be an effective motivator for users to increase their effort, leading to work of improved quality. In contrast to Witkowski’s theoretical work [2], we validated our hypothesis empirically and by using a simulation model designed based on empirical data. We chose to evaluate a simple reputation system rather than adopting more complex approaches inspired by peer prediction [2, 3, 4] and studied by means of theoretical research. The choice of a simple reputation system was motivated by the desire to study a system that could be used in real-content evaluation systems.

We studied the impact of negative payoffs on users’ effort in the evaluation of the credibility of web pages, in a situation where ground truth is known. Our study consisted of two parts. The first part used an experiment in which college programming students evaluated web pages about the Java programming language in terms of credibility, using a 5-point Likert scale. Pages had been previously prepared (and their content modified) and evaluated by an expert; pages could contain varying amounts of errors, leading to a reduced credibility evaluation. Participants in our experiment received reputation scores depending on how well their evaluations agreed with the expert’s judgments. In the experiment, two settings, with penalties and without penalties, have been considered. Empirical results confirm the effectiveness of penalties and are used to design a simulation model that allows us to examine further scenarios.

The second part of our study was based on an analysis of a simulation model that was strongly grounded in our experimental results. Data from the experiment was used for design and validation of the multi-agent simulation model. The goal of this analysis was to broaden the scope of investigated payoff schemes utilizing penalties, as well as to study the effect of penalties on participation and effort. A scenario when only a part of the Webpages had expert evaluations was also studied using simulation.

* 1. **Research Contributions**

The contribution made by our work is the validation of the hypothesis that penalties in reputation systems lead to increased effort and enhanced quality of work. This validation is based on an empirical study and on simulation, thereby complementing the theoretical results regarding effectiveness of penalty mechanisms. Our study also considers simple, intuitive structures of reputation payoffs, which would be easier to use in a practical reputation system. Finally, our study contributes to the understanding of the risks and limitations of using penalties. We show how penalties can lead to reduced participation by users, thus creating a tradeoff between quality (through the increased effort of participating users) and quantity of results (through reduced participation by users who had previously contributed to an information system and who might choose not to participate). We also confirm a second hypothesis proposed by Witkowski [2]: that penalties would lead to the selection of workers with superior skills, since those with inferior skills would drop out of the system. This hypothesis is confirmed by the results of our simulation, which used a more general and realistic model than that employed in theoretical research.

The article is structured as follows. In the next section, we describe related work. Section 3 describes our empirical experiments, the reputation payoffs in these experiments, and the experiment results showing the effectiveness of using penalties. Section 4 describes our simulation model, which was closely based on experiment results, but enabled the study of additional variants of reputation payoffs; the results of the simulation experiments are also discussed in this section. Section 5 concludes the paper.

1. **Related work**

The problem of using reputation scores to reward members of an online community for content provided is very complex, as it involves the evaluation of the quality of this content. By content provided by users, we mean questions or answers posted on a question-and-answer site (for example, Stack Exchange); user’s opinions about purchased products published on an e-commerce portal (for example, Amazon); or completing a task on a crowdsourcing portal (for example, Mechanical Turk or Upwork). There are many studies focusing on performance of algorithms designed to motivate members of online communities, especially for members of question answering websites. Question and Answer portals are treated as knowledge sharing platforms, where the most important is a spread of expertise. The aim of studies on experts’ identification on such websites is an attempt to rank users of platform regarding to their expertise. Jun Zhang and others presented several algorithms useful in finding members with expertise in online communities [5]. Wei-Chen Kao and others [6] designed the hybrid algorithm considering user subject relevance, user reputation and authority of a category in finding experts. Their experiment results have shown that their model outperform other conventional methods. Cigdem Aslay [7] proposed a community expertise network structure based on the principle of competition among the answerers of a question. They also found that the ability to identify experts depends on the type of community. Other common approach is to use link analysis in order to find experts in online network [8, 9, 10]. This technique takes into account users’ degree of authority. Bouguessa et al. [11] proposed a model of authority scores using estimations based on Bayesian Information Criterion (BIC) in the combination with the Expectation-Maximization (EM) algorithm.

The main problem is to find an appropriate and credible point of reference for evaluating users’ activity on such portals. Thus, to evaluate this content, existing online platforms use opinions about content generated by members of the community instead of automatic algorithms. Commonly used algorithms automatically enhance or damage the reputation of the author of content after receiving a single opinion about this content (positive or negative) from members of the relevant online community. In order to prevent manifestations of mindless and unjustified criticism, which can discourage authors from uploading new content, portals use a variety of techniques, such as:

1. Meta-moderation, i.e., evaluations of evaluators (e.g., as introduced by the portal Slashdot);

2. A form of “fee” or “payment,” paid by a user from his or her reputation points, for expressing a negative opinion;

3. A daily limit on the number of negative opinions expressed by an individual user (this solution, like solution (2), is used on Q&A portals, for example, on sites from the family of Stack Exchange portals).

Potentially, meta-moderation is the most effective of the techniques used to control negative phenomena associated with the direct use of community members’ opinions to calculate reputations. Unfortunately, this technique can be also the most expensive from the point of view of the owner of the platform (assuming that the platform employs second-order evaluators). Additionally, it is also susceptible to manipulation; for example, if second-order evaluators are elected from ordinary members of the community, they can include a dishonest person (for instance, an evaluator who promotes the content of certain users and criticizes content published by others). The phenomenon of publishing biased opinions is known even on portals which enable users to publish reviews of products. The idea of peer prediction [3, 12] is a response to these situations. This concept assumes that by using a specific payoff structure (and if certain assumptions are met), the reputation algorithm is capable of motivating users to post honest opinions, even if the platform does not have a reliable point of reference which, for each users’ opinion, would provide a basis for the validation process. However, the effectiveness of alternative models proposed to date, based on the peer-prediction model, has been confirmed mainly by theoretical analysis.

Of the models proposed so far, only one, as far as we know, investigates the effect of negative payments. Witkowski et al. [2] created a model, based on game theory and using the peer-prediction approach, demonstrating that the introduction of penalties can discourage users with less knowledge from participating in the game, while those with more extensive knowledge are motivated to stay in the game and to invest effort into completing tasks. However, this model has not been verified under real conditions.

Our approach can be interpreted as the generalization of downvoting. In our model, users’ ratings of user-generated content are compared to an expert’s rating. Users are rewarded for accurate ratings. In our model, we also use users’ effort, similarly to Witkowski et al. [2] and Dasgupta and Ghosh [13]. However, our model is clearly not an effort-incentivizing mechanism; moreover, effort is a continuous, rather than a binary, variable. We assumed that agents are lazy and want to maximize reputation points while minimizing their investment of effort. Reputation is a function of invested effort, but the agents’ aim is to find an equilibrium between invested effort and accumulated reputation points. It has been found that a high reputation may actually lead to decreased activity of coders in crowdsourcing contests for software development [14], which suggests that approaches that decrease reputation (such as downvoting or our approach) may result in increased motivation. In contrast to Witkowski [2] and Dasgupta and Ghosh [13], we considered the situation in light of known ground truth. To better illustrate our downvote approach, we present it in Figs. 1 and 2. The diagram of downvotes shows the standpoint known from popular question-and-answer sites such as StackOverflow.com.

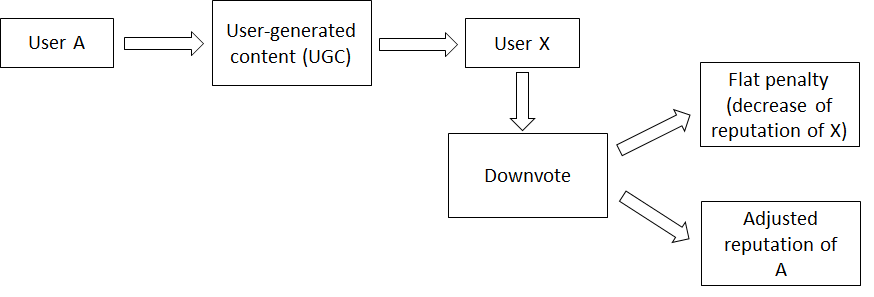


Fig. 1. Downvote approach

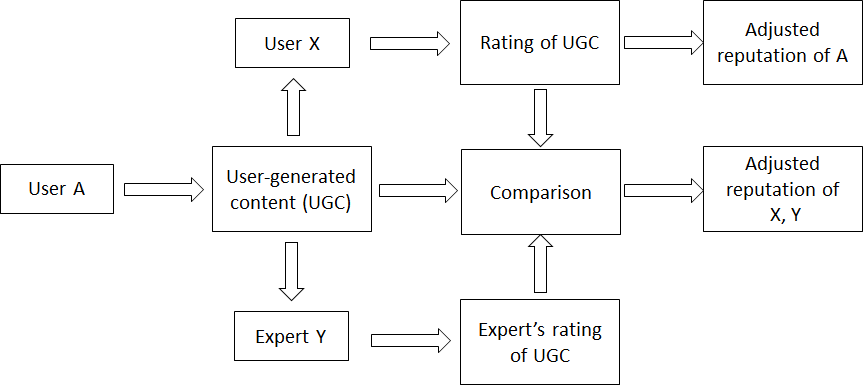


Fig. 2. Our approach

Recent research on crowdsourcing [15,16] uses worker models capable of applying the results of our research (especially the model proposed in section 4.1). Worker models are used by algorithms for quality estimation and prediction. Algorithms proposed by Baba [15] take into account the fact that workers may have various skills or biases in quality evaluation. However, workers also may use various degrees of effort in order to create results. The notion of worker effort is derived from models of human cognition created by psychologists [17]. A human being can carry out any mental task using one of two systems: the automatic, heuristic System 1 and the intentional, deliberative System 2. Using System 2 requires mental effort, but can produce more correct results. Using System 1 requires no mental effort and can produce correct results in many cases as well, but can also lead to systematic biases and reasoning fallacies.

Peer-prediction algorithms and similar approaches are relevant for crowdsourcing systems that use pairwise comparisons for quality evaluation [18, 19, 20]. Our research concerns the design of payoffs in such algorithms for increased worker motivation. Moreover, research on crowdsourcing has investigated the issue of motivating workers [21, 22, 23] to exert greater effort, which was our goal as well. Peer-prediction approaches can also be thought of as methods that stimulate the interaction of peers who may be knowledge workers. Recent research has found [24] that interaction among high-, and low-performing knowledge workers is beneficial for the lower performing knowledge workers and does not adversely impact the high-performing ones. Mechanisms similar to peer-prediction are peer comparison or peer evaluation, which have been found to operate in crowdsourced innovation platforms [25]. Contributors of low-potential ideas eventually become inactive due to the peer evaluation mechanism, while high-potential idea contributors remain active, which is similar to the self-selection mechanism postulated by Witkowski.

1. **Experimental evaluation of penalties**

In this section, we present the design and results of our experiment. We also present the most important results for creating the agent model and the agent strategy used in our simulation model that is described in Section 4.

The experiment has been conducted at the Faculty of Computer Science of Polish-Japanese Academy of Information Technology. We divided the experiment into two parts. Two different and independent groups of students (students taking the same class in different semesters) have been invited to participate in each part of the experiment. The first part was carried out between December 18, 2013 and January 15, 2014 (with the first group of students – a control group), the second between November 27, 2014 and January 12, 2015 (with the second group of students – a treatment group). Students participating in the first part of our experiment could not participate in the second part. In the first part, users could not receive penalties (negative reputation points) for completing tasks; in the second part, penalties were used. The accuracy of student answers has been evaluated by a comparison to the results provided by the expert. We observed that users’ results in the second part of the experiment were more accurate than in the first. We also observed that the average time a user spent on the evaluation increased impressively, from 20 seconds per web page in the first part of the experiment to one minute in the second part. We concluded that the positive changes in these indicators were the result of introducing penalties in the second part.

* 1. **Experiment design**

In the experiment, participants became users of the online platform Reconcile http://www.reconcile.pl/ and were asked to evaluate web pages about the Java programming language in terms of credibility. They were informed that they would evaluate various web pages, some of which would be credible, others not. They also knew that all of the web pages had been pre-evaluated by an expert. The choice of the Reconcile platform (a prototype designed for supporting evaluation of web content credibility) was motivated by the desire to test the impact of a modified reputation system on a realistic web content evaluation task. Within the framework of the Reconcile platform, we could change the reputation algorithm as desired; on the other hand, the platform resembles many real-content evaluation systems.

Our experiment was divided into two parts using two different payoff schemes, as described below. In each part, 300 pages were available for evaluation by every participant. Users evaluated web pages using a 5-point ordinal (Likert) scale, which is a typical choice for Web content credibility evaluation. Participants received reputation points that reflected the difference between their rating and the expert’s evaluation.

* + 1. **Webpage Credibility Preparation**

Web pages related to the Java programming language were specially chosen and prepared. First, their design was standardized: all web pages had the same font and only necessary graphics (without advertisements, etc.) in order to avoid the non-measurable impact of other factors on evaluation of the credibility of web pages in the system. Secondly, the webpages presented to the users were modified to provide a certain amount of false information. The expert’s credibility evaluations were proportional to the number of these modifications, so that each false statement reduced the website’s credibility by one point. Following our modification, a credible page was one with no errors (the expert’s rating equaled 5); a page which had one mistake received an expert’s rating equal to 4; a page with 2 errors received an expert’s rating equal to 3; and so on. We excluded the influence of other factors, such as the layout or design of the web page. All pages looked the same: text only, on a white background. Users evaluated Web content in terms of credibility. A 5-point ordinal scale was used, where the question was:

*Rate the credibility of this webpage/website/text. By credibility we mean information which is, in your opinion, true information that you can trust.*

*1. completely not credible;*

*2. mostly not credible;*

*3. somewhat credible, although with serious doubts;*

*4. credible, with some doubt;*

*5. completely credible.*

* + - 1. **Factors Affecting Credibility Evaluation**

According to the Prominence-Interpretation theory of Web content credibility evaluation [26], users evaluate Web page credibility in two stages. In the first stage, users investigate the Web content and notice some credibility cues. On the other hand, other (possibly important) cues may go unnoticed at this stage. Prominence is the likelihood that a Webpage element will be noticed or perceived. Five factors influence prominence:

* the motivation (“effort”) of the user,
* the ability (“experience”) of the user,
* the task of the user,
* a user's individual characteristics,
* topic of the Web page.

Since the task and topic were the same for all pages and for both the treatment and control group, these two factors are controlled. This means that only the user’s ability, motivation or individual characteristics could have influenced prominence. In the design of the treatment and control groups described below, we have evaluated that the ability and individual characteristics that have the largest impact on credibility evaluation have been similar. This leaves motivation as the only factor that could have influenced prominence of cues used in Webpage credibility evaluation.

During the second stage, Interpretation, users make judgments about the cues they have noticed in the first stage. Since users were evaluating Webpages about programming, and the pages contained erroneous statements, the only possible interpretation of a discovered error was to reduce the credibility evaluation of the Webpage. This means that in our experiment design, only participants’ motivation could impact the correctness of the credibility evaluation. A greater amount of incorrect information increased the chance that the users would notice errors, and therefore a higher difference from the user’s to the expert’s rating was indicative of a lower user motivation that adversely impacted rating accuracy, because a user did not read the Webpage with sufficient attention. For example, if the Webpage was modified to contain 2 errors, it received an expert credibility rating of 3 on the 1-5 Likert scale. If the user chose a rating of 5, this was evaluated as a more significant mistake than if the user chose a rating of 4, since a rating of 5 meant that the user did not notice any one of the two errors.

* + 1. **Participants and Control Group**

Participants of our experiments were assigned into two groups: the control and treatment group. Both groups consisted of students of the same course, but taken by students in different years. The control group consisted of students who took a course of the Java programming language in the academic year 2013/2014. The treatment group consisted of students who attended the same course in the academic year 2014/2015. A pre-test that measured the level of knowledge of Java, as well as the final grades of the Java programming course, have not shown significant differences between these two groups of students. The average number of points earned by each class was 49/100 for the class of students invited to the experiment as the control group and 46 points for the class of students invited to the study as the treatment group. The ability of student participants to evaluate credibility of Webpages concerning Java programming was therefore similar in the two groups.

Rafalak et al. [27] has shown that trust and risk-taking are psychological traits allowing to assign crowdsourcing platform users into one of three groups: group 1 – users with a tendency to underestimate websites’ credibility, group 2 - users with a tendency to overestimate websites’ credibility, group 3- users with a tendency to give adequate credibility judgments. We have measured individual characteristic of students using a psychological questionnaire. All students invited to our experiment took psychological test before running our experiment. Intensity of psychological characteristics among the students were assessed using International Personality Item Pool (http://ipip.ori.org). Risk-taking was measured by JPI: Rkt (Cronbach’s alpha coefficient = 0.78) and trust was measured using NEO: A1 (Cronbach’s alpha = 0.82). Results have shown that there were no significant differences in individual traits between students who decided to take part in our experiment (control vs treatment group). Thus, the only factor that may affect the differences in the results achieved in our experiment by the two groups of students is motivation (and the resulting cognitive effort).

* + 1. **Payoff schemes**

The aim of our experiment was to learn how different payoffs affect the users’ effort, measured as the time spent on the evaluation of the credibility of websites in a real-content evaluation system. Our experiment was divided into two parts. In the first, participants received only positive reputation points for completing tasks (the control group); in the second, they also received negative payoffs (in terms of reputation points) in the case of incorrect evaluations (the treatment group). The table below presents the payoff schemes used in our experiment.

Table 1: The difference in reputation payoffs between the first and the second part of the experiment

|  |  |  |
| --- | --- | --- |
|  | Payoff schemes | |
| Difference between user’s and expert’s rating | Reputation payoff  Control group | Reputation payoff  Treatment group |
| 0 | 8 | 8 |
| 1 | 4 | 0 |
| 2 | 0 | ‒8 |
| 3 | 0 | ‒8 |
| 4 | 0 | ‒8 |

* + 1. **Motivation for Participation and Effect of Penalties on Participation**

Participants in the experiment were motivated using grading points for the Java programming course. The same motivation system has been used in both parts of the experiment (for the control and treatment group). The rankings for both groups were independent.

Participants received additional grade points in a course in programming. The number of grade points was decided by the lecturers. The best 20% of students received 8 grade points, the second 25% of students 6, the next 30% 4, and the last 25% 2. Users could access a ranking (shown in Fig. 3 below) on their reputation points on the Reconcile platform. The ranking was divided into four parts; each part corresponded to the quartile used to determine students’ rewards. Participants also had access to their total reputation and activity scores, but these scores were not reflected in the rankings of all users, since the most important thing for students was their position in the four-part ranking. The reputation score consisted of the sum of user’s points received for completing the tasks, and depended on whether the user’s evaluations were consistently close enough to the expert’s evaluations. Activity scores were awarded for every task (activity in the system) independently of the quality of evaluations or answers. For completing each task, a user received 1 point on their activity score. Taking part in our experiment was not obligatory and all additional work done by students was rewarded.

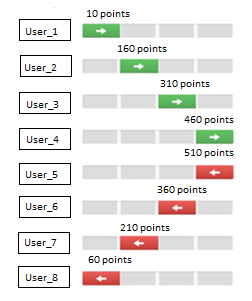


Fig. 3: Users’ ranking of reputation scores

The users’ ranking was divided into four parts. The part showing the 20% of the best users is on the far right. The arrows show whether a user’s position has improved or worsened after completing a task and receiving reputation points. Following completion of each task, a user’s position changes.

The table below presents the number of users invited to take part in the experiment along with the number of participants in each part of the experiment. Participation in our experiment was voluntary, although students were motivated by points that influenced their final grades. We observed significantly less participation in the second part, in which penalties were introduced. Therefore, reduced participation in the second part of experiment may be an effect of penalties. However, in both parts of experiment, participants were rewarded based on their position in the final ranking and not directly on their reputation scores (this is an approach chosen in realistic reputation systems, were the desire to increase a user’s position in the reputation ranking serves as an external motivation).

Table 2: Number of invited students and participants in the experiment

|  |  |  |
| --- | --- | --- |
|  | Number of students | |
|  | Control group | Treatment group |
| Invited students | 140 | 124 |
| Experiment participants | 65 | 26 |
| Response rate | 46% | 21% |

* + 1. **Experimental Task: Web Page Evaluation**

When evaluating web pages, users had access to the distribution of previous ratings given by others. Web pages were evaluated in terms of the credibility of their content, their appearance, the expertise and intentions of their authors, and the completeness of the information they contained. Users received reputation points only for evaluating the credibility of a given web page.

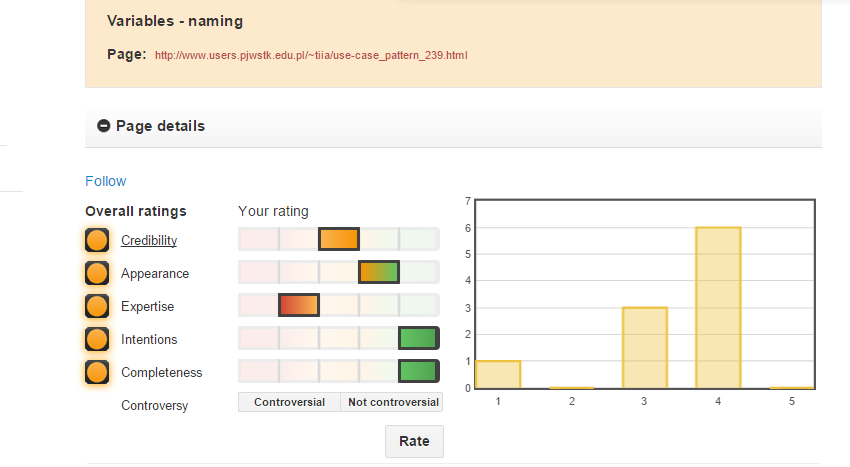


Fig. 4: Evaluation of a web page in the X system. The chart on the right presents the distribution of ratings of credibility of a web page on the 5-point scale mentioned previously.

* 1. **Results**

In this section, we present the results of our experiments. The next subsection contains the basic data regarding users’ ratings, distributions of ratings, their accuracy, and analyses of users’ reputation scores.

* + 1. **Number of ratings**

The following table presents the number of ratings provided by users. In the first part of the experiment, in which we did not use penalties, 65 active users submitted 7,036 ratings (the control group). In the second part of the experiment, in which the influence of penalties was tested, 26 active users submitted 1,684 ratings (the treatment group).

Table 3: Number of ratings

|  |  |  |
| --- | --- | --- |
|  | Control group | Treatment group |
| Number of ratings | 7,036 | 1,684 |
| Average ratings per user | 108 | 65 |

We can see that users from treatment group were less active than from control group (this can be interpreted as a participation effect). This effect could be due to penalties, since the motivation offered to both student groups was the same. Moreover, students from control and treatment groups were attending the same course but at different times. However, we cannot conclude with certainty that the observed effect is due to the change in payoffs. For this reason, we validated this hypothesis using a simulation.

* + 1. **Distribution of ratings**

The following charts present the difference of aggregated distributions between users’ ratings and expert’s ratings. Fig. 5 shows the rating distributions from the two parts of the experiment.

Fig. 5: Rating distribution for a 5-point scale

As we can observe from the chart above, users’ ratings are more positive than the expert’s ratings. However, the distribution of ratings of treatment group is closer to the uniform distribution of the expert’s ratings than the distribution of ratings of control group. The reason the evaluations of users were so different from the expert’s evaluation may be a bias observed for many types of credibility ratings [28]. The observed bias of users’ ratings was included in our simulation model.

* + 1. **Comparison of the accuracy of the two parts of the experiment**

The charts below present the accuracy of the submitted ratings. Numbers on the horizontal axis correspond to the error between the users’ and the expert’s ratings.

Fig. 6: Accuracy of the ratings in the two parts of the experiment

Ratings of the treatment group were more accurate than those of the control group. The improvement was achieved mostly for small errors, since the frequency of error equal to 1 decreased by 5%, while the frequency of error equal to 0 increased by 9%. This led us to the conclusion that the enhanced effort of users was largely effective in correcting small mistakes. We also tested hypothesis 0, which stated that the two distributions obtained from the two parts of the experiment were identical. The hypothesis was rejected by a Mann-Whitney U test, with a significance level equal to 0.05.

* + 1. **Users’ effort (measured by time)**

In this subsection we present results regarding the time spent on evaluation in the two groups. We consider time spent as users’ effort. We show that the introduction of penalties had the effect of prolonging the time spent on evaluation. In the control group, the average time spent on the rating process was 20 seconds; in the treatment group, it was almost 1 minute (we did not include results reflecting over 10 minutes spent per evaluation). Users spent, on average, a total of 00:34:29 for evaluation of web pages in the system in the control group and 1:11:48 in the treatment group.

Fig. 7: Distribution of total time spent by users in the two parts of the experiment

The chart above shows the percentage of users who spent a given amount of total time in the system on the evaluation of standard web pages. Fig. 8 presents the percentage of users who submitted a certain number of standard ratings. We can see that in the treatment group users evaluated fewer pages, but spent more time on evaluation.

Fig. 8: Distribution of the number of pages evaluated by users included in the analysis of time spent

The increased amount of time for evaluations in the treatment group also shows which of the two possible factors: user’s motivation, or user’s individual characteristics, was the cause for the improved accuracy. The increase of user’s motivation is clearly observed in our experiment.

* + 1. **Participation and activity**

As we mentioned before, participants in the second part of our experiment (treatment group) were less active. Moreover, the response rate to the invitation to participate in the experiment was lower in the second part than in the first (in both invitations, the existence of penalties was explained to invited students). In the first part of the experiment, the response rate was 56% (the control group), in the second 27% (the treatment group). Fig. 9 presents the distributions of numbers of ratings in the two parts of our experiment, and also reflects the distribution of activity scores (users received 1 activity point per rating).

Fig. 9. Distribution of the number of ratings in the two parts of the experiment

* + 1. **Users’ reputation**

This section is devoted to differences in reputation scores earned by users in the two parts of the experiment. We show here how the introduction of penalties changed users’ scores.

In our experiment, users were awarded two kinds of points, namely reputation and activity scores. A user was awarded or penalized reputation points for each evaluation. Reputation scores depended on the accuracy of his or her ratings. Activity scores measured users’ activity in the system. Every evaluation, irrespective of its accuracy, was counted as 1 point of activity.

Table 4: Descriptive statistics of reputation scores

|  |  |  |
| --- | --- | --- |
| Reputation scores | Control group | Treatment group |
| Average | 450 | 84 |
| Min. | 0 | ‒208 |
| Max. | 1,772 | 592 |

Table 5: Descriptive statistics of activity scores

|  |  |  |
| --- | --- | --- |
| Activity scores | Control group | Treatment group |
| Average | 108 | 65 |
| Min. | 1 | 1 |
| Max. | 300 | 258 |

In the control group, the average reputation score was higher than in the treatment group. Users from treatment group were also less active. Moreover, there were users who did not improve the quality of their evaluations, receiving negative scores for almost every evaluation.

* 1. **Conclusions drawn from the experiment**

Based on the experimental results, we can conclude that the introduction of penalties can have a positive effect on the accuracy of evaluations and users’ effort measured in terms of time.

In the next section we present the results obtained using our simulation model. Data obtained in our experiment was used to design and validate the proposed model. We decided to design a simulation model in order to test scenarios regarding different payoffs, and to conduct a sensitivity analysis the impact of penalties on user’s effort. The two effects observed in the experiment, namely, the influence of penalties on the accuracy of users’ ratings and on their effort, were used to validate our simulation model.

1. **Description of the simulation model**

In this section, we present the design of our simulation model and the results of tests of various scenarios. Our goal in the simulation study was to verify more scenarios, using different reputation payoffs (with varying levels of penalties), than the payoffs studied in our experiment. Using simulation, it was also possible to verify the effect of reducing the number of expert ratings available to the reputation algorithm. In the experiment, all Webpages had expert ratings; on the other hand, in a realistic scenario of a Web-based platform for Web content credibility evaluation, only a minority of Webpages would be likely to be evaluated by an expert.

We based the agent model and strategies in the simulation on experimental data. In this section, we describe our simulator design and agent strategies, namely, effort and participation.

In our model, agents submitted ratings of an object generated by a platform using a 5-point scale () with a level of probability representing a combination of the agent’s effort and bias distributions. We assumed that an agent’s decision about a rating reflects that agent’s effort and a certain tendency in the assignment of ratings. This tendency is called bias. The platform generates objects with a given ground truth rating with uniform probability for all possible values from the scale . The higher the level of the agent’s effort, the greater the probability of accurate assessment. Effort is an agent’s strategy; it is continuous and takes values from the set <0, 1>. Moreover, our model assumes the existence of bias, which is a global variable equal for all agents. In the literature, as well as in our studies, there are many examples of existing bias in people’s judgments [29]. Our experimental results have shown that respondents evaluating objects are more positive than experts. Overly positive assessments are also observable in e-commerce [30]. Some investigations have proved that this skewness of distribution of human judgments is an effect of using the Likert scale and is an intrinsic feature of this scale [31]. The bias in our simulation model reflects the probability of submitting ratings based on the distribution of ratings of all users from our experiment. As in our actual experiment, agents receive reputation scores for each rating depending on how much their ratings diverge from those generated by the platform. We also considered a situation in which agents resign from evaluation of web pages. This agent strategy is called a participation strategy, and is defined as the probability that the agent will evaluate a given web page. The lower the probability, the greater the possibility that an agent will resign from evaluation.

Agents evolve using a utility function that increases with reputation, but decreases with higher effort. When an agent resigns from evaluation, the utility function is unchanged. We studied stable strategies (values for effort and participation) for various designs of reputation payoffs (various levels of penalties for errors).

* 1. **Simulator design**

The design of our simulator closely resembles content evaluation systems such as the Reconcile platform, which was used in our experiment. The model assumes that users receive a signal and then provide information, for instance, by evaluating a web page. The agent sees a web page and decides whether or not he or she wants to evaluate it. Then, the agent’s evaluation is compared to an expert’s rating and the agent receives a payoff (reputation points) depending on the difference between the agent’s evaluation and the expert’s.

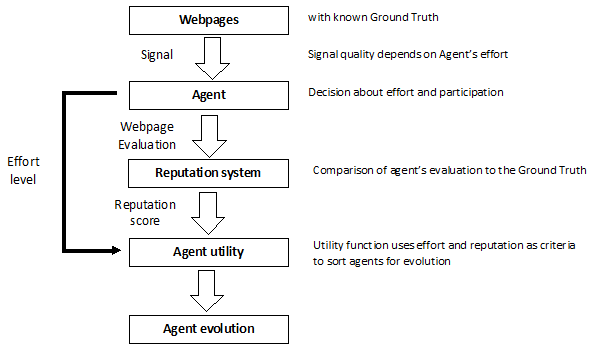


Fig. 10: Schematic of the simulation model

* 1. **Agents’ strategies**

In the proposed model, agents model users of an online platform, for instance a crowdsourcing platform. Based on the results of our experiment and in-depth interviews with experiment participants, we formulated an assumption regarding the preferences and utilities of simulated agents. We assumed that the aim of such agents is to simultaneously maximize their reputation scores and minimize their effort. Users maximize gains while minimizing costs. In our model, reputation score represents gain; effort represents cost. Where users have the same level of expertise and create a uniform group, effort can be measured by time spent on evaluation of a web page. In general, the effort parameter captures all factors that may contribute to the improvement of a user’s work (content evaluation), including the skill or experience of the user. In our agent model, the parameter called “skill” affects the standard deviation of the agent’s effort distribution. A high value for the skill parameter causes a decrease in the standard deviation value of the agent’s effort distribution, which means that an agent with a higher level of skill does not need to invest much effort to obtain good results. Conversely, an agent with a lower skill level needs to invest more effort to obtain good results. However, “skill” is also a separate parameter, one that we used for sensitivity analysis of the “effort” variable. A change in “skill” level modifies the influence of “effort.”

Another strategy available to agents is participation in the evaluation of a given site. In our model, the participation strategy is expressed by the probability that an agent will evaluate a page. The greater the probability of participation, the greater the possibility that the agent will evaluate web pages.

* + 1. **Effort**

In the proposed model, effort is an agent’s strategy. Our model assumes that an increase in effort increases the probability of providing the correct evaluation. Therefore, effort affects reputation, and reputation is a function of effort; hence, the user’s fitness criteria are not independent. The effort value is in the range <0, 1>. We assumed that every user sets his or her own effort value within that range.

The final distribution of an agent’s evaluations is a combination of two distributions. The first is an effort distribution, with a mean equal to the correct rating of an evaluated website and a standard deviation which is a global variable for the population (a parameter). The users’ standard deviation of effort’s distribution depends on the level of another global variable, the parameter called “skill.” This parameter is in the range <0, 1> and affects the standard deviation of effort distribution. The higher the value of skill, the lower the observed standard deviation (more accurate distribution). The second distribution is the users’ bias distribution, which is equal for all agents. This is a skewed distribution with a predefined mean and standard deviation, equal for all agents, as parameters. The user’s bias indicates the option most preferred by users. The bias distribution reflects the distribution of ratings by the participants in the experiment. In our previous studies, such as in the described experiment, we observed that users’ ratings were skewed towards positive values. This means that people tend to evaluate objects, especially web pages, as better than they really are. In terms of the distribution of users’ evaluations, we observed that they were negatively skewed. This is why we use bias distribution in the reported model.

The final distribution, which determines the agent’s behavior, is calculated using these two distributions: effort and bias. The more effort the agent uses, the stronger the impact of the effort distribution on the final results (with 100% effort, only the effort distribution will be used).

where:

– the correct answer (the mean of effort distribution)

– the standard deviation of effort distribution (a parameter)

– a biased answer (the mean of bias distribution, a parameter). The rating most preferred by users, constant for all users.

– the standard deviation of bias distribution (a parameter)

The shape of bias distribution was taken from experimental data.

– skill

– (parameter) effort scaled to the range <‒6, 6>, equal to

Examples of the agent model:

Example 1

Table 6. Settings for Example 1

|  |  |
| --- | --- |
| Correct answer/rating | 2 |
| Effort ‒ standard deviation | 1 |
| Effort value | 0.6 |
| Skill | 1 |
| Bias ‒ standard deviation | 1 |
| Biased answer | 4 |

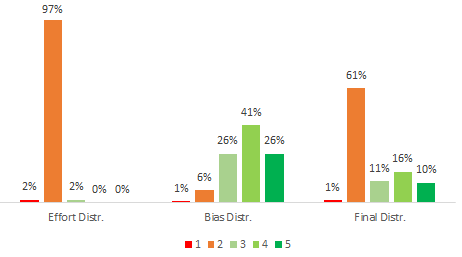


Fig. 11: Final distribution of user’s behavior, Example 1

Example 2

Table 7. Settings for Example 2

|  |  |
| --- | --- |
| Correct answer/rating | 2 |
| Effort ‒ standard deviation | 1 |
| Effort value | 0.6 |
| Skill | 0 |
| Bias ‒ standard deviation | 1 |
| Biased answer | 4 |

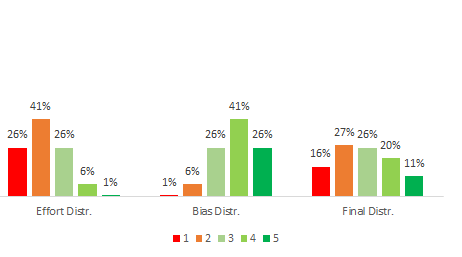


Fig. 12. Final distribution of user’s behavior, Example 2

* + 1. **Participation in evaluation**

Another user strategy is participation in website evaluation. This is a continuous variable from the range <0, 1> expressing the probability that an agent will submit a rating. When this probability is a low number, the agent frequently resigns from evaluating sites; when high, the agent will frequently submit a rating.

* 1. **Agents’ model: the utility (fitness) function**

The function based on the reference point method was used as an objective function [32, 33]. The reference point method of preference modeling in multi-criteria decision problems has been found to accurately express the preferences of decision-makers faced with multiple criteria, such as users in our experiment who simultaneously try to maximize their reputations and minimize their effort. This function integrates several criteria (in our case, two criteria that are not independent: reputation score is a random function of effort). This method enables the selection of Pareto-optimal solutions with respect to all criteria. For each criterion, a reservation and an aspiration level are specified expressing the preferences of the decision-maker (agent) with regard to each criterion. We assumed that the criteria are maximized (with respect to effort, we transformed the criterion to a maximized one as will be described below). The aspiration level belongs to a set of values defined for a criterion considered acceptably high (sufficient) by an agent. In the reported model, we set the aspiration level to the maximum of a criterion (in effect, this means that this parameter is not used). The reservation level is a value considered unacceptably low by an agent. The reservation level is treated in our model as a parameter.

The utility (fitness) function of an agent, is:

{;

where:

*rep*: reputation scores – criterion 1

*1 – eff*: 1 minus effort – criterion 2

: parameter, in our settings equal to 0.05

: a piecewise linear scaling function for each criterion including the reservation level of the criterion and the aspiration level of the criterion under consideration.

We assumed value 1 as the aspiration level for the variable (1 – effort), which means that users strive to exert value 0 of effort. As an aspiration level for reputation score, we assumed the maximum value which could be calculated for reputation scores for one generation. In the reported model, we used 100 iterations for one generation, which corresponded to a release of 100 pages for evaluation by one agent. For the payoff used in our experiment, the maximum value for evaluation of one web page is 8. In the case of 100 iterations, we obtained 800 as the maximum value of reputation for one generation.

Reservation levels of the criterion (1 ‒ effort) and reputation scores are set to half of the aspiration values. This means that agents will be equally satisfied by decreasing their effort to 0.5 and by achieving a reputation of 400. This setting of the reservation value gives the two criteria an equal impact on the agents’ utility function. This means that the agents can choose the strategy of decreasing effort, if they are able to do so without a decrease in reputation (this is possible if, for example, agents can avoid participating in content evaluation).

The function for effort is:

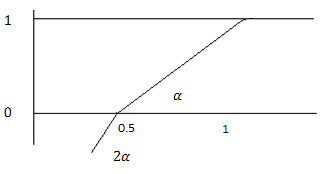


Fig. 13: Form of scaling function for effort used in the agents’ utility function

The function for reputation is:

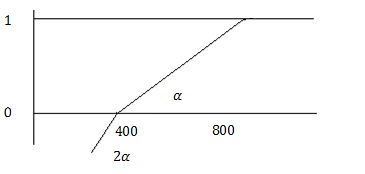


Fig. 14: Form of scaling function for reputation used in the agents’ utility function

The aim of agents with this utility function is to maximize their reputation scores while simultaneously minimizing their effort. However, agents may value effort and reputation differently, as expressed by different reservation levels for these criteria (aspiration levels are both at the maximum). Also, the utility function imposes a more severe penalty for letting a criterion drop below the reservation level than it does for letting it drop below the aspiration level. This shape of the scaling function reflects the behavior of decision-makers.

* 1. **Evolution of agents’ strategies**

Evolution of agents’ strategy involves their effort and participation. We applied the stochastic universal sampling technique to select strategies potentially useful for the next generation, using the objective function of the reference point method as a fitness metric. The stochastic universal sampling technique enables us to avoid genetic drift [34]. The evaluation function orders users by taking into account reputation scores and users’ effort, minimalizing their effort while maximizing their reputations based on reservation and aspiration levels.

* 1. **Parameter settings for the reported simulation**

The table below presents the parameter settings used in our simulation.

**Table 7. Parameter settings**

|  |  |
| --- | --- |
| Parameter | Value |
| Number of runs | 10 |
| Number of generations | 500 |
| Number of iterations | 100 |
| Number of agents | 1000 |
| Reservation level (effort) | 0.5 |
| Aspiration level (effort) | 0 |
| Reservation level (reputation) | 400 |
| Aspiration level (reputation) | 800 |
| Skill | <0, 0.3, 0.5, 1> |
| Epsilon | 0.05 |
| Number of ratings (for one page) | 5 |
| Standard deviation of bias distribution | 1 |
| Standard deviation of effort distribution | 1 |
| Bias ‒ on which rating | 4 |
| s-curve parameters | 12\*effort ‒ 6 |

* 1. **Results**
     1. **Impact of payoffs on agents’ effort without including the participation strategy**

In this section, we present the effects of different payoff schemes on agents’ effort. We consider scenarios involving various sets of parameters. Parameters such as standard deviation of effort and bias are taken from the experimental data, more specifically from general distributions of users’ ratings. We do not include our model participation strategy in this section. Parameters are fixed for various scenarios beside skill and payoffs. We considered four levels of skill parameters and three different payoff schemes.

Table 8. Considered payoffs

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Payoffs | | | | | | |
| Error | Probability of error | Proportional Heavy (PH) | Only (O) | Proportional (P) | Heavy (H) | Medium (M) | None (N) |  |
| 0 | 20% | 10 | ‒8 | 10 | 8 | 8 | 8 |  |
| 1 | 32% | ‒15 | ‒8 | ‒5 | ‒8 | 0 | 4 |  |
| 2 | 24% | ‒20 | ‒8 | ‒10 | ‒8 | ‒8 | 0 |  |
| 3 | 16% | ‒25 | ‒8 | ‒15 | ‒8 | ‒8 | 0 |  |
| 4 | 8% | ‒30 | ‒8 | ‒20 | ‒8 | ‒8 | 0 |  |
| Expected value of game | | ‒14 | ‒8 | ‒6 | ‒4.8 | ‒2.24 | 2.88 |  |

We tested more payoff schemes; the greater the penalty, the more effort is put into evaluations. Here we show the results for the two payoff schemes which were used in our experiment (M and N) and four other scenarios (PH, O, P, and H); one of them is similar to the scenario from the second part of experiment, but payoffs are more severe (scenario H). The table below shows the average effort for the last generation for 24 scenarios. The scenarios do not include participation strategies. Effort stabilizes quickly, after about 30 generations.

Table 9. Average effort for the last generation for 24 scenarios without the participation strategy

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Skill | Payoff | | | | | |  |
|  | **PH** | **O** | **P** | **H** | **M** | **N** |  |
| 0 | 0.99 | 0.003 | 0.89 | 0.98 | 0.75 | 0.45 |  |
| 0.3 | 0.99 | 0.003 | 0.78 | 0.89 | 0.69 | 0.45 |  |
| 0.5 | 0.90 | 0.003 | 0.72 | 0.79 | 0.65 | 0.44 |  |
| 1 | 0.71 | 0.003 | 0.61 | 0.64 | 0.59 | 0.43 |  |

From the table above we can see that the average effort is highest when the most severe penalties are used (scheme PH and H). Moreover, the higher the skill level, the lower the average effort. Agents do not need to use more effort if their skill level is high, as a higher skill level makes the effort distribution more effective. For example, the probability that an agent with skill = 1 and effort = 0.4 will choose the correct answer is almost the same as probability that an agent with skill = 0.3 and effort = 1 will do so. Contrastingly, the lowest level of effort is observed when the game makes no sense, as in the case of a payoff scheme including equal payoffs for every decision.

Table 10. Average reputation scores for each scenario

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Skill | Payoff | | | | | | |
|  | **PH** | **O** | **P** | **H** | **M** | **N** |  |
| 0 | ‒406 | ‒800 | 73 | ‒68 | 176 | 415 |  |
| 0.3 | ‒50 | ‒800 | 181 | 75 | 246 | 426 |  |
| 0.5 | 152 | ‒800 | 259 | 180 | 288 | 434 |  |
| 1 | 332 | ‒800 | 390 | 330 | 369 | 449 |  |

The table above includes the average reputation scores for each scenario. We can see that the values are lower for scenarios incorporating penalties. Higher skill levels reduce effort, but at the same time result in higher average reputation scores. This is because skill affects agents’ effort distribution and makes their evaluations more accurate.

* + 1. **Impact of payoffs on agents’ effort, including the participation strategy**

In this section we present results that take into account another strategy used by agents: the participation strategy. We considered the same payoff schemes. In this case as well, effort and participation strategy stabilize quickly. Participation strategy is the probability that agent will take part in the game and will evaluate a page. If the agent decides not to participate, he or she receives 0 points (0 is better than a penalty). The fitness function in this case also optimizes for effort and reputation (and does not include the participation strategy). Effort is a constant value for one generation. For this designed model, we obtained only two values of participation strategy for the last generation for every scenario, converging to 0 or 1. The convergence depends on the utility function of agents. Below a certain level of an agent’s utility function, his or her effort and participation strategy converge to 0.

**Table 11. Average effort for the last generation for 24 scenarios including participation strategy**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Skill | Payoff | | | | | |  |
|  | **PH** | **O** | **P** | **H** | **M** | **N** |  |
| 0 | 0.03 | 0.02 | 0.89 | 0.02 | 0.75 | 0.46 |  |
| 0.3 | 0.03 | 0.02 | 0.79 | 0.88 | 0.69 | 0.45 |  |
| 0.5 | 0.90 | 0.02 | 0.72 | 0.79 | 0.66 | 0.45 |  |
| 1 | 0.71 | 0.02 | 0.61 | 0.64 | 0.59 | 0.44 |  |

The table above shows the average effort for the last generation for the scenarios including participation strategy. Effort for which the participation strategy is close to 0 is indicated in gray. The addition of participation strategy affects agents’ effort only when it is low (converges to 0). In other cases, effort is the same as it is without the inclusion of participation strategy, because the value of participation strategy is always close to 1, meaning that agents play and receive reputation scores. Agents with low levels of skill are willing to resign from a game. Penalties cause agents with low levels of skill to give up the game.

Table 12. Average reputation scores for each scenario including participation strategy

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Skill | Payoff | | | | | |  |
|  | **PH** | **O** | **P** | **H** | **M** | **N** |  |
| 0 | ‒12 | ‒7.24 | 67 | ‒4.5 | 172 | 410 |  |
| 0.3 | ‒11 | ‒7.22 | 175 | 71 | 241 | 421 |  |
| 0.5 | 139 | ‒7.28 | 252 | 174 | 282 | 428 |  |
| 1 | 329 | ‒7.21 | 378 | 322 | 362 | 444 |  |

The table above presents average reputation scores for scenarios including participation strategy. We can see that, in scenarios in which average reputation scores are lower than the expected value from the game (see Table 9), agents decide not to play.

* + 1. **Scenarios including lack of expert’s ratings for chosen percentage of pages**

In this section we present result for scenarios including incomplete assessment of sites by experts. We considered two options: 25% of pages in each generation have experts’ ratings, and 0% of pages in generation have experts’ rating. Agents did not receive reputation points for the evaluation of a website without known expert’s rating. As in sections above, we show results including and not including participation strategy.

* + - 1. **Impact of payoffs on agents’ effort, without including the participation strategy**

The most important observation is that reducing the number of expert ratings does not affect the main observed effect of penalties that increase agent’s effort. However, some differences are visible. Results presented below have shown that skill does not affect effort when only 25% percent of webpages have expert’s evaluation. Effort does not react to an increase in skill with a small amount of expert evaluations, while effort decreases with increasing skill if 100% of pages have the experts' evaluations. This can be explained as follows: effort "reacts" to increased skill when agents "notice" that they can reduce effort without losing reputation. However, in this case reputation must be sensitive and responsive to changes of effort. Meanwhile, when there are fewer ratings, reputation is less variable in general and therefore does not react to changes. The agent can itself reduce effort, but his reputation will not change, because there are no expert’s evaluations. Thus, the agents effort will remain high.

The second observation is that when there are no webpages with expert’s rating in the system, agents’ effort is close to 0. It could mean that even small number of expert’s ratings and a low risk of verification of agents’ evaluation will motivate agents to invest effort. This result is good news for online platforms that rely on voluntary, expert verification of quality of user-generated content.

Table 13. Average effort for the last generation for 24 scenarios without the participation strategy. 25% of pages with expert’s ratings

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Skill | Payoff | | | | | |  |
|  | **PH** | **O** | **P** | **H** | **M** | **N** |  |
| 0 | 0.98 | 0.003 | 0.98 | 0.98 | 0.99 | 0.99 |  |
| 0.3 | 0.99 | 0.003 | 0.99 | 0.98 | 0.99 | 0.99 |  |
| 0.5 | 0.99 | 0.003 | 0.99 | 0.99 | 0.99 | 0.99 |  |
| 1 | 0.99 | 0.003 | 0.99 | 0.99 | 0.99 | 0.99 |  |

Table 14. Average reputation scores for each scenario, 25% pages with expert’s rating

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Skill | Payoff | | | | | |  |
|  | **PH** | **O** | **P** | **H** | **M** | **N** |  |
| 0 | -90 | -210 | 39 | -18 | 76 | 128 |  |
| 0.3 | -5.8 | -206 | 90 | 34 | 116 | 154 |  |
| 0.5 | 87 | -224 | 140 | 96 | 141 | 186 |  |
| 1 | 250 | -196 | 266 | 198 | 171 | 206 |  |

Table 15. Average effort for the last generation for 24 scenarios without the participation strategy. 0% of pages with expert’s ratings

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Skill | Payoff | | | | | |  |
|  | **PH** | **O** | **P** | **H** | **M** | **N** |  |
| 0 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 |  |
| 0.3 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 |  |
| 0.5 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 |  |
| 1 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 |  |

* + - 1. **Impact of payoffs on agents’ effort, including the participation strategy**

Tables below show result for scenarios including the participation strategy for 25% of sites with expert’s rating. Results of scenarios including the participation strategy for 0% of sites with expert’s ratings are similar to results presented above for scenario including 100% pages without expert’s evaluation. The only difference is that the participation strategy is unstable, but in average is between 40%-60%.

Table 16. Average effort for the last generation for 24 scenarios including the participation strategy. 25% of pages with expert’s ratings

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Skill | Payoff | | | | | |  |
|  | **PH** | **O** | **P** | **H** | **M** | **N** |  |
| 0 | 0.03 | 0.02 | 0.98 | 0.02 | 0.98 | 0.98 |  |
| 0.3 | 0.03 | 0.02 | 0.98 | 0.98 | 0.98 | 0.99 |  |
| 0.5 | 0.98 | 0.02 | 0.98 | 0.98 | 0.98 | 0.99 |  |
| 1 | 0.99 | 0.02 | 0.99 | 0.99 | 0.98 | 0.99 |  |

Table 17. Average reputation scores for each scenario, 25% pages with expert’s rating

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Skill | Payoff | | | | | |  |
|  | **PH** | **O** | **P** | **H** | **M** | **N** |  |
| 0 | -2.5 | -1.4 | 30 | -0.9 | 73 | 139 |  |
| 0.3 | -2.7 | -1.4 | 93 | 33 | 106 | 157 |  |
| 0.5 | 73 | -1.25 | 150 | 83 | 135 | 155 |  |
| 1 | 238 | -1.28 | 233 | 208 | 195 | 210 |  |

Table 18. Average participation strategy for the last genaration, 25% of websites with expert’s rating

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Skill | Payoff | | | | | |  |
|  | **PH** | **O** | **P** | **H** | **M** | **N** |  |
| 0 | 0.007 | 0.007 | 0.98 | 0.008 | 0.98 | 0.98 |  |
| 0.3 | 0.008 | 0.007 | 0.98 | 0.98 | 0.98 | 0.99 |  |
| 0.5 | 0.98 | 0.007 | 0.98 | 0.98 | 0.98 | 0.99 |  |
| 1 | 0.99 | 0.007 | 0.99 | 0.99 | 0.98 | 0.99 |  |

1. **Conclusions**

The goal of our research was to evaluate the effect of applying penalties in reputation systems on the effort and quality of user input. This research question is of interest in many real-life systems, for example, different crowdsourcing-based platforms [35]. This kind of system was used in our experiment. The results indicate the positive impact of penalties on the time spent on content evaluation (a measure of effort for our homogenous group of users) and the accuracy of content evaluation.

We attempted to extend the experimental results using a simulation study that enabled us to consider more payoff schemes. Generally, the results show that increased penalties lead to increased effort. However, the simulation results also show the risk of increased penalties, as they have an adverse effect on motivation to participate. We have shown that when agents with low levels of skill have an opportunity to resign (having no internal motivation for participation), they prefer this option to investing effort.

The simulations also allowed to check the effect of reduced availability of expert’s ratings. Only if a Webpage had an expert’s ratings, the reputation system could modify an agent’s reputation. The results of these simulation show that as long a small number (25%) of Webpages have expert evaluations, penalties motivate agents to maintain a high effort. However, if the number of expert ratings drops to 0%, agents stop investing any effort in their evaluations.

Our simulations confirmed the hypothesis proposed by Witkowski [2] that agents with low skill levels would drop out of the system, while agents with higher skill levels would remain and invest greater effort. This is an important result, since in our simulation model the payoffs were much more realistic (they follow a simple, fixed scheme), and the model is strongly based on empirical experimental data. Therefore, we have not only confirmed the hypothesis that penalties lead to the selection of workers with higher levels of skills, but we have also extended the range of applicability of this hypothesis.

The conclusions from the simulations are limited by the simulation model, which assumes that agents value effort and reputation equally. However, our model has been validated in the sense that it reproduces the behavior observed in the experiment (that effort increases in the Medium penalties scenario as compared to the No penalties scenario). Also, the simulation model’s design and parameters have been based on empirical data from our experiment (such as the use of the bias distribution). Therefore, the results obtained from the simulation part of our study can be treated as a good basis for deriving hypotheses for future empirical research or reputation system design.

In the future, we plan to consider a situation where ground truth for the evaluation of reputation is not available to the platform, so that more sophisticated algorithms, such as peer prediction, must be used.

**Acknowledgements**

We thank all the participants who took part in our experiment, and all lecturers who allowed us to carry out the experiment and provided a reward for participants. This work is supported by Polish National Science Centre grant 2015/19/B/ST6/03179.

**References**

[1] Y. Zhang, M. Van der Schaar, Reputation-based incentive protocols in crowdsourcing applications. In: INFOCOM, Proceedings IEEE, (2012) 2140-2148.

[2] J. Witkowski, Y. Bachrach, |P. Key, D. Parkes, Dwelling on the negative: Incentivizing effort in peer prediction. In: First AAAI Conference on Human Computation and Crowdsourcing (2013)

[3] N. Miller, P. Resnick, R. Zeckhauser, Eliciting informative feedback: The peer-prediction method. Management Science 51(9) (2005) 1359-1373.

[4] P. Zhang, and Y. Chen, Elicitability and knowledge-free elicitation with peer prediction. Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems. International Foundation for Autonomous Agents and Multiagent Systems. (2014) 245-252.

[5] J. Zhang, M. S. Ackerman, L. Adamic, Expertise networks in online communities: structure and algorithms. In: WWW, ACM (2007) 221 – 230.

[6] W. C. Kao, D. R. Liu, S. W. Wang, Expert Finding in Question-Answering Websites: A Novel Hybrid Approach, ACM (2010) 867 – 871.

[7] Ç. Aslay, N. O’Hare, L. M. Aiello, Alejandro Jaimes, Competition - Based Networks for Expert Finding, ACM (2013) 1033-1036.

[8] L. Page, S. Brin, R. Motwani and T. Winograd, The Pagerank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies Project, 1999.

[9] J. M. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM (JACM), v.46 n.5 (1999), 604-632.

[10] J. Zhang, M. S. Ackerman, L. Adamic, K. K. Nam, Qume: a mechanism to support expertise finding in online help-seeking communities, in UIST ’07: Proceedings of the 20th annual ACM symposium on User interface software and technology. New York, NY, USA: ACM (2007) 111–114.

[11] M. Bouguessa, B. Dumoulin, S. Wang, Identifying authoritative actors in question-answering forums: the case of Yahoo! Answers. In KDD, ACM (2008) 866-874.

[12] J. Witkowski, D. C. Parkes, Peer prediction without a common prior, Proceedings of the 13th ACM Conference on Electronic Commerce. ACM (2012) 964-981.

[13] A. Dasgupta, and A. Ghosh, Crowdsourced Judgement Elicitation with Endogenous Proficiency. In Proceedings of the 22nd ACM International World Wide Web Conference (WWW’13) (2013) 319-330.

[14] N. Archak, A. Ghose, Learning-by-Doing and Project Choice: A Dynamic Structural Model of Crowdsourcing, ICIS 2010 Proceedings (2010).

[15] Y. Baba, H. Kashima, Statistical Quality Estimation for General Crowdsourcing Tasks. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2013) 554–562.

[16] D. Haas, J. Ansel, L. Gu, and A. Marcus, Argonaut: Macrotask Crowdsourcing for Complex Data Processing. In: Proceedings of the VLDB Endowment 8(12) (2015) 1642–1653.

[17] D. Kahneman, Thinking, Fast and Slow. Farrar, Straus and Giroux, New York, 2011.

[18] O. F. Zaidan, C. Callison-Burch, Crowdsourcing Translation: Professional Quality from Non-Professionals. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (2011) 1220–1229.

[19] X. Chen, P. Bennett, K. Collins-Thompson, E. Horvitz, Pairwise Ranking Aggregation in a Crowdsourced Setting. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (2013) 193–202.

[20] J. Ren, Y. Zhang, K. Zhang, and X. Shen, SACRM: Social Aware Crowdsourcing with Reputation Management in Mobile Sensing. Computer Communications 65 (2015) 55-65.

[21] O. Scekic, H. Truong, S. Dustdar, Incentives and Rewarding in Social Computing. Communications of the ACM 56(6) (2013) 72-82.

[22] A. Kittur, J. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, J. Horton, The Future of Crowd Work. In: Proceedings of the 2013 conference on Computer supported cooperative work (2013) 1301-1318.

[23] A. Kittur, E. Chi, B. Suh, Crowdsourcing User Studies with Mechanical Turk. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2008) 453-456.

[24] G. Kane, S. Ransbotham, A. Boynton. Is High Performance Contagious among Knowledge Workers? AIS Electronic Library (AISeL) (2012).

[25] Y. Huang, P. Vir Singh, K. Srinivasan, Crowdsourcing new product ideas under consumer learning. Management Science 60.9 (2014) 2138-2159.

[26] B. J. Fogg, Prominence-interpretation theory: Explaining how people assess credibility online. CHI'03 extended abstracts on human factors in computing systems, ACM (2003) 722-723.

[27] M. Rafalak, K. Abramczuk, A. Wierzbicki, Incredible: Is (Almost) All Web Content Trustworthy? Analysis of psychological factors related to website credibility evaluation. In Proc. WWW 2014 (2014) 1117-1122.

[28] M. Kąkol, M. Jankowski-Lorek, K. Abramczuk, A. Wierzbicki, M. Catasta, On the subjectivity and bias of web content credibility evaluations. In Proceedings of the 22nd International Conference on World Wide Web, ACM (2013) 1131-1136.

[29] V. Kostakos, Is the crowd's wisdom biased? A quantitative analysis of three online communities. In Computational Science and Engineering, International Conference on. IEEE (2009) 251-255.

[30] R. Nielek, A. Wawer, A. Wierzbicki, Spiral of hatred: social effects in Internet auctions. Between informativity and emotion, Electronic Commerce Research 10.3-4 (2010) 313-330.

[31] S. Jamieson, Likert scales: how to (ab) use them. Medical education 38(12) (2004) 1217-1218.

[32] A. Wierzbicki, Trust and fairness in open, distributed systems. Springer, 2010

[33] A. P. Wierzbicki, A mathematical basis for satisficing decision making. Mathematical Modelling, vol. 3 (1984) 391-405.

[34] L. Samuelson, Evolutionary games and equilibrium selection, MIT Press, 1998.

[35] R. Jurca, B. Faltings, Mechanisms for Making Crowds Truthful. Journal of Artificial Intelligence Research 34(1) (2009) 209-253.