Why Pay?: Exploring How Financial Incentives are Used for Question & Answer

Gary Hsieh, Robert E. Kraut, Scott E. Hudson
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15213
{garyh, robert.kraut, scott.hudson}@cs.cmu.edu

ABSTRACT
Electronic commerce has enabled a number of online pay-for-answer services. However, despite commercial interest, we still lack a comprehensive understanding of how financial incentives support question asking and answering. Using 800 questions randomly selected from a pay-for-answer site, along with site usage statistics, we examined what factors impact askers' decisions to pay. We also explored how financial rewards affect answers, and if question pricing can help organize Q&A exchanges for archival purposes. We found that askers' decisions are two-part—whether or not to pay and how much to pay. Askers are more likely to pay when requesting facts and will pay more when questions are more difficult. On the answer side, our results support prior findings that paying more may elicit a higher number of answers and answers that are longer, but may not elicit higher quality answers (as rated by the askers). Finally, we present evidence that questions with higher rewards have higher archival value, which suggests that pricing can be used to support archival use.

Author Keywords
Question and answer, Q&A, pay-for-answer, market, social computing.

ACM Classification Keywords
H.5.3 Group and Organization Interfaces: Web-based interaction.

INTRODUCTION
Electronic commerce (e-commerce) payment systems have enabled many new types of monetary-based interactions, one of which is pay-for-answer question and answer (Q&A) services. Until fairly recently, when seeking help from others online, people would post their requests and questions on bulletin boards, forums and free Q&A sites. With e-commerce, people seeking help can offer a financial reward when they post their questions, which is awarded to answerers when the questions are answered.

Despite the number of e-commerce enabled pay-for-answer services that have been launched in the past few years (*e.g.*, Google Answers, Mahalo Answers, Just Answers, UClue, AskBright), we still do not have a good understanding of how financial incentives impact Q&A. Most existing research has explored whether or not financial rewards improve answer quality, but even this question has not been answered conclusively, with some studies concluding yes [14,16], and another no [4]. Many other important questions have been overlooked. For example, how do the question askers decide whether or not to offer rewards for answers? What factors affect how much they offer? Can question pricing be used to help other users find quality answers in Q&A archives? Without a more comprehensive evaluation of how pay-for-answer systems work, site designers will not be able to leverage the full power of the market system.

To improve our understanding of pay-for-answer Q&A services, we analyzed 800 questions randomly selected from Mahalo Answers, a pay-for-answer Q&A service that allows its users to ask both free and for-pay questions. Our analyses found that askers' decisions are more complex than simple cost-benefit analyses. Existing norms about how financial rewards should be used also seem to impact their decisions. Our results also support prior findings that paying more elicits a higher number of answers and longer answers. Additionally, we found that the offered financial reward correlates with the archival value of the questions, which suggests that the reward pricing may be used to help organize the generated knowledge repository.

Findings from this work offer multiple contributions. Most immediately, our findings can be applied to improve user experience on pay-for-answer Q&A systems. Our coding and analysis of Q&A questions also improves our understanding of the different characteristics of Q&A questions, which may be extremely valuable when targeting and suggesting questions to answerers. Finally, our findings on the use of financial rewards by askers indicate that financial rewards may be more appropriate for informational exchanges than for social interactions.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.
Copyright 2010 ACM 978-1-60558-929-9/10/04....$10.00.

QUESTIONING AND ANSWERING WITH MONEY

Question Asking

Research in help seeking suggests that people consider the costs and benefits when deciding whether or not to ask for help [*e.g.*,8]. On one hand, askers must determine how much they can gain from having the answers. The more useful the answers are to people's successful goal attainment, the more likely they will ask for help. On the other hand, there are costs associated with requesting help. While there are no monetary costs on free Q&A sites, there are psychological and social costs. Receiving help may reduce one's sense of competence and may also reduce one's external reputation [*e.g.*,28,29]. Prior work examined how aversion towards indebtedness impacts people's willingness to ask questions [*e.g.*,11,12] and showed that people who do not anticipate being able to return a favor are less willing to ask for help [*e.g.*,12]. But in general, the effects of these costs tend to be weaker for help-seeking on online Q&A sites because the questions are not targeted to any particular answerer, and because of anonymity.

How is the question asking decision affected by the availability of payment systems? Because of how these interfaces are typically designed, askers often have two decisions—whether or not to pay and how much to pay.

From a cost-benefit perspective, paying for help affects whether or not to ask a question in that it raises the cost of asking for help. Prior work has shown that askers are less likely to ask for help when getting help reduces their gains (*e.g.*, getting partial credit on an academic task) [8]. Similarly, recent related research on using financial payments to fight spam has shown that requiring a financial payment when sending help requests can make senders more selective in the messages they send [17]. We expect the same effects in the Q&A scenario – the additional cost will make askers more selective in the questions they ask. Specifically, question askers should be willing to pay more when they perceive the answers to be more valuable.

One way to infer the potential value of answers is from the characteristics of the questions. In general, we would expect factors such as importance (or sincerity, how much the asker really wanted an answer), urgency and difficulty (how hard it is to acquire the answers) to influence askers' valuation.

H1. People are more likely to pay for answers to questions that are more valuable—important, urgent and difficult to answer.

H2. People offer higher rewards for answers to questions that are more valuable—important, urgent and difficult to answer.

Aside from benefits and costs, decisions to pay or not pay may also be influenced by existing social norms. Prior work has suggested that there are two types of social interactions—exchange and communal [6]—and that monetary rewards may be perceived as violation of norms when used in communal interactions [6,1,15,30]. For example, paying for dinner is expected at a restaurant, but may be insulting when eating a home-cooked meal. This suggests that if askers' goals are to acquire commodities (*i.e.*, information), then they may be more willing to pay. However, if the goals are mainly social in nature (*e.g.*, conversational), then askers may think it is inappropriate to pay. While this has been suggested in a prior work on market-based Q&A [16], more research is needed to confirm this hypothesis.

Recent work has classified Q&A questions into categories: factual, advice, opinion, and non-questions [14, 16]. Factual questions seek objective data, or pointers to content; advice questions seek recommendations on an asker's individual situation; opinion questions seek others' thoughts on a topic of general interest; and non-questions are spam. Factual and advice questions are informational questions, while opinion questions are categorized as conversational questions—their goal is to stimulate a discussion [13].

Using this classification for questions, if users follow existing social norms when using the payment system, we would expect question askers to be more willing to pay for informational questions, which seek facts and advice, than for conversational questions.

H3. People are more likely to pay for questions seeking information than for questions initiating conversations.

Question Answering

Why people volunteer to help without any tangible rewards has been a subject of extensive research [*e.g.*, 2,3]. The motivators can be roughly classified as intrinsic and social. Experimental work and real world observations have indicated that people have numerous types of intrinsic motivations such as altruism, egoism, and self-rewards such as desire to learn, develop, expand and demonstrate abilities [*e.g.*,2,7]. People may also provide help for free to accomplish social goals such as making friends and seeking approval from others [*e.g.*,27]. Recent interviews with users of Korea's most popular Q&A service, Knowledge-iN, confirmed that many of these intrinsic and social-image motivations are attracting users to participate in the free service [22].

What happens when financial incentives are introduced? Research in both psychology and experimental economics has shown that small monetary incentives can crowd-out intrinsic and social motivators [*e.g.*,7,10]. According to self-determination theory [7], financial incentives may be perceived as controlling; people who are financially rewarded for working on a task may feel that they are doing so because of the tangible rewards, rather working on the task for its own sake.

But in the real world, individual motivations are not homogeneous. Those who use and continue to use pay-for-answer services may be the subset of population who are motivated by the small rewards offered due to self-

selection. Hence, the crowding-out effect may be minimal on these pay-for-answer sites.

Prior studies of the impact of financial incentives on answers have found contrasting results. Comparing answer quality on Google Answers, a pay-for-answer service, to popular free Q&A services, Harper *et al.* found that higher pay resulted in better answers [14]. A controlled field experiment by Hsieh *et al.* also showed that paying for answers increased average answer quality [16]. However, another study by Chen *et al.* showed that while paying more led to *longer* answers (*i.e.*, more effort), it did not result in *better* answers [4]. Chen *et al.* pointed out that a major difference in their study was the assessment of only the officially selected answer, while other studies assessed all answers. Here, we re-examine the effects of paying on answering by using data from Mahalo Answers.

H4. Higher rewards will elicit longer answers on average (more individual effort).

H5. Higher rewards will elicit more answers.

H6. Higher reward questions will elicit higher quality best/chosen answers.

Indirect Effects of Paying for Answers

Q&A sites often have a third purpose, aside from question asking and answering – to archive knowledge. In fact, Yahoo Answers is currently the second-most visited education/reference site after Wikipedia [20]. Unfortunately, the many frivolous, low-quality questions posted may make it difficult for users to find useful content. This has resulted in concerns about the value of Q&A services for reference purposes [20]. Prior research focused on how financial rewards affect the askers and answerers who engage in an exchange, but did not examine how paying for answers affects the majority of visitors to Q&A services, who are using the site for knowledge search.

One of the indirect benefits of using financial incentives is the publicly visible pricing information. The question rewards, while set by askers to support their own goals, may also act as useful indicators for those searching through the knowledge archives in the future. If questions with higher rewards also have higher archival values, then the offered price may be used to filter out the frivolous Q&A exchanges.

H7. Questions with higher rewards also have higher archival value.

MAHALO ANSWERS

To examine the impact of financial incentives on Q&A, we conducted a study of Mahalo Answers. Mahalo Answers is a pay-for-answers Q&A site, launched on December 15, 2008 [21]. According to Mahalo Answers, by mid-February 2009, the site had about 15,000-25,000 visitors per day.

Google Answers will be used as a comparison point in this paper since it has been the most often studied pay-for-answer Q&A site [9,14,23,24,25]. There are two major differences between these two sites. First, Google Answers uses a set of "researchers" chosen by Google to answer questions, whereas anyone who joins Mahalo Answers can answer questions and earn the financial rewards. Second, Mahalo Answers allows both free and paid[1] questions, while Google Answers only allows paid questions. This enables us to compare the usage differences when the questions are paid versus when they are not paid, which was not feasible in prior studies of Google Answers.

Figure 1. Screenshot of Mahalo Answers.

On Mahalo Answers, if askers choose to pay for answers, the minimum payment is one Mahalo Dollar (M$1), which costs one US Dollar to purchase. Questions paid for by the users are displayed separately from free questions on the site's home page. The paid questions are shown immediately above the screen fold, and the free questions are below the screen fold (Figure 1).

Within the first three days of posting a question, the asker can select any answer as the best answer (and consequently reward the answerer with the payment if one was offered). During the first 3-4 days, the question asker can also indicate that there was "no best answer" to receive a full refund on the question. Afterwards, if the best answer is not chosen, then the other members of the community can vote to select the best answer. Once answerers have earned more than 40 Mahalo Dollars, they can choose to cash out, at which point Mahalo Answers takes a 25% cut. In other words, cashing out M$40 will give the user $30.

[1] Mahalo Answers users can offer to pay when asking and also gratuitously tip any answerers post-hoc. They are both called "tipping" on the site. To differentiate the two, we use "paying" to refer to offering financial rewards when asking the question, and "tipping" to mean offering a bonus after the answer is received.

Each answerer can only give one answer per question, but there can be multiple answerers per question. If an answerer's answer is selected as the best one, the question asker can choose to rate it on a 5 point scale, where 3 is "good" and 5 is "above and beyond."

There were two major changes to Mahalo Answers during our data collection. First, around February 20, Mahalo Answers started sponsoring, or paying for, the free questions posted on the site as an effort to increase site traffic. These sponsored questions are still posted to the same place as the free questions. Mahalo sponsored M$0.25 per question initially, but the value varied over time. Even though Mahalo Answers automatically sponsored the questions when they are received, Mahalo Answers did remove the sponsored payments at their discretion. The second major change was that around February 24, Mahalo Answers started accepting questions asked through Twitter. Additionally, Mahalo Answers started actively pulling in questions from public Twitter accounts that were not intended for Mahalo Answers.

To account for these changes, in one of our analyses on question-asking decisions, we removed questions posted after February 20 so that we could focus on usage without the influence of company sponsorship, and ensure that questions were intentionally asked by askers.

DATA COLLECTION

We contacted Mahalo Answers for their data. While they stated their intent to offer a public API or make their data available, this did not occur. So instead, we wrote a Java program to gather the questions and answers posted on Mahalo Answers. Due to site moderation, we were unable to gather some posted questions that were later deleted. Also, while Mahalo Answers allows users to ask each other direct, private questions, we were constrained to only the public questions. We were able to gather a total of 22,205 public questions and 71,091 answers posted on Mahalo Answers between Dec. 04, 2008 and May 05, 2009. For our analyses, we removed all posts posted before December 15, 2008, before the site was launched. We also removed all posts after April 27, 2009 to ensure that we analyzed only questions that were closed to any new answers.

	Google Answers [23]	MA <Feb. 24	MA Full
Period of Study	6/2002-5/2006	12/2008-2/2009	12/2008-4/2009
Duration	48 months	2 months	4 months
# user-paid questions	~2,700	~1,600	~1,300
# free & sponsored questions	N/A	~2,500	~3,500
# answers provided	~1100	~17,000	~15,000
# comments sent	~3700	~5,600	~5,000
# users who joined	N/A	~6,900	~5,500
Avg. $ of question, user-paid only	$20.90	$2.52	$2.70
Rated answers	~680	~2000	~1700
Avg. answer rating (5 point scale)	4.63	3.73	3.85
System price range	$2-200	$1-100	$0.25-101

Table 1. Comparison of per month statistics of Mahalo Answers (both reduced dataset of usage prior to February 24 and our full dataset) to Google Answers.

General Site Statistics

Mahalo Answers had more answers per question than did Google Answers (Table 1). This was due to the fact that each question at Google Answers could only be answered by one Google Answers Researcher. This also resulted in comments being heavily used on Google Answers as an alternative mechanism to giving answers [23]. On average, the price of questions offered on Google Answers was

	Coding Category	Descriptive Text
Question Types	**Factual**	The question is asking for facts (objective data or pointers to content).
	Opinion	The question is asking for opinions (questions seek others' thoughts on a topic of general interest; these questions do not have a "correct" answer and may be answered without reference to the question asker's needs).
	Advice	The question is asking for personal advice (questions seek recommendations based on the asker's own situation; answerers must understand the question asker's situation to provide a good answer).
	Non-question	The question is spam / not a question.
Question Value	**Sincerity (Importance)**	How sincerely did the question asker want an answer to the question?
	Urgency	How urgently did the question asker want an answer to the question?
	Difficulty	How much work would it require an average high school educated person to answer this question? Keep in mind that work includes both getting the answer and also formulating the answer.
Question Politeness		How rude or polite is the question?
Question Archival Value		I think high-quality answers to this question will provide information of lasting/archival value to others.

Table 2. Rated characteristics of question.

much higher than the price offered on Mahalo Answers.

Rating Question Characteristics

We randomly selected 800 questions posted by non-Mahalo employees: 400 were user-paid, 400 were not. Questions were then rated by workers on Amazon's crowd-sourcing service, Mechanical Turk. Given only the question text from Q&A exchanges, workers rated each question on nine separate dimensions using Likert scales. Recent research has shown the feasibility of using Mechanical Turk to collect ratings and annotations [26]. Unpublished demographic studies have shown that >75% of the workers on Mechanical Turk are from the United States and that >70% of Turkers have a bachelor's degree [17], so most of the raters should be fairly proficient in English.

First, the question types were rated. Unlike prior research, instead of classifying questions into mutually exclusive question types, we asked coders to rate the extent to which the question asked for facts, asked for opinions, asked for advice, or did not ask for anything in particular (spam). (see Table 2). Question types are not mutually exclusive; a question may have both high opinion and advice ratings, indicating that it is asking for both opinions and advice.

Raters also rated questions on three dimensions that may possibly indicate the askers' valuations of having the questions answered. This included perceived sincerity, urgency and difficulty of the question. We defined sincerity as the extent to which question askers wanted answers to their questions, which we then used as a proxy for importance. Coders also rated the politeness of the question, which may affect responsiveness. These ratings were used as independent and control variables in our analyses. Finally, coders rated the questions on an outcome measure: archival value. Questions with higher archival values can improve the usefulness of the Q&A repository.

To improve the quality of the ratings, coders who gave ratings with very low variance, used noticeable patterns, or did not spend enough time determining ratings (<20 seconds per question) were removed. After this filtering process, there were 401 raters and each rated 19 questions on average. We standardized these ratings per rater (z-score) by subtracting the raters' mean ratings by the raters' standard deviation. When standardized, 0 means an average rating, +1 means a rating that is one standard deviation above the average, and -1 means a rating that is one standard deviation below the average. In our final dataset, we had on average 9.4 ratings per question. To check the reliability of the ratings, intraclass correlation reliabilities were calculated, which indicated what proportion of the variance was associated with questions and not with the judges. The general rule of thumb is that ICC = 0.40 to 0.59 is moderate, 0.60 to 0.79 is substantial and 0.80 is outstanding [19]. Most of our ratings had an intraclass correlation of around 0.60, except for the "not a question" and politeness ratings (see Table 3, last row).

ANALYSIS AND RESULTS

The analyses are broken up into question asking, answering, and archival use. In question asking, we explore what question characteristics predict whether the question is paid. In question answering, we analyze if paying improves

	Factual	Advice	Opin.	Not Q	Sincere	Urgent	Diff.	Polite	High Arch.	Is Paid	Avg. Answer Length	Ans. Count	Is Rated
Factual	1.00												
Advice	0.03 0.44	1.00											
Opinion	-0.45 0.00	0.35 0.00	1.00										
Not Q	-0.30 0.00	-0.21 0.00	-0.03 0.42	1.00									
Sincere	0.29 0.00	0.43 0.00	0.04 0.29	-0.51 0.00	1.00								
Urgent	0.29 0.00	0.47 0.00	-0.05 0.15	-0.33 0.00	0.63 0.00	1.00							
Difficult	0.39 0.00	0.24 0.00	-0.15 0.00	-0.33 0.00	0.43 0.00	0.48 0.00	1.00						
Polite	0.13 0.00	0.30 0.00	0.07 0.00	-0.37 0.00	0.59 0.00	0.35 0.00	0.20 0.00	1.00					
High Archival	0.47 0.00	0.44 0.00	-0.01 0.71	-0.37 0.00	0.56 0.00	0.43 0.00	0.45 0.00	0.37 0.00	1.00				
Is Paid	0.05 0.14	0.07 0.07	0.10 0.00	-0.02 0.40	-0.04 0.26	-0.00 0.95	-0.01 0.79	-0.05 0.20	0.09 0.01	1.00			
Avg. Ans Len	0.04 0.29	0.14 0.00	0.11 0.00	-0.03 0.40	0.11 0.00	0.11 0.00	0.17 0.00	0.03 0.33	0.12 0.00	0.21 0.00	1.00		
Ans. Count	-0.15 0.00	0.10 0.00	0.26 0.00	0.05 0.15	-0.05 0.16	-0.09 0.02	-0.18 0.00	-0.01 0.75	0.00 0.99	0.32 0.00	0.06 0.11	1.00	
Is Rated	-0.11 0.00	0.08 0.02	0.22 0.00	-0.00 0.91	0.04 0.32	-0.01 0.82	-0.05 0.12	0.04 0.24	0.04 0.24	0.25 0.00	0.11 0.00	0.22 0.00	1.00
ICC reliability	0.62	0.61	0.59	0.42	0.58	0.62	0.61	0.44	0.58				

Table 3. Correlation table (with significance) of rated characteristics and dependent variables for all 800 questions.

answer length, count and quality. Finally, we examine if higher payments predict higher archival value.

Question Asking

We found many significant correlations between the rated characteristics of questions posted on Mahalo Answers. First, results confirmed findings from prior work that the degree to which the question asks for facts and advice correlates with the degree to which the question seems sincere [16], and has higher archival value [13]. The results also showed that the degree to which the question asks for facts and advice correlates with degree of perceived question urgency and difficulty, but the opinion nature of the question did not. Furthermore, while factual and opinion ratings were negatively correlated, advice and opinion ratings were actually positively correlated (Table 3).

Is Paid 0=free, 1=paid	n=333		Model	
	Mean	SD	Odds Ratio†	SE
Factual	-0.04	0.47	5.19*	3.99
Opinion	0.03	0.44	1.17	0.88
Advice	-0.01	0.48	3.67	3.06
Not Question	0.02	0.38	0.57	0.46
Sincerity	-0.03	0.48	0.43	0.38
Urgency	-0.03	0.50	3.38	3.01
Difficulty	-0.03	0.50	0.24	0.20
Prior Q. (log)	1.48	1.40	2.07	0.78
Prior Earn (log)	0.84	1.56	2.07	0.85

* p < 0.05

Table 4. Random-effects logistic regression model predicting "is paid" (H1 & H3), using questions posted before February 20 (n=333).

†Odds ratio is a measure of effect size. It indicates how a unit increase in a variable affects the likelihood of the question being paid, holding all others factors constant.

To test which question characteristics predict whether a question is paid, a random-effects logistic-regression model was built (using STATA's xtlogit command). The dependent variable is whether the question is paid or not (binary). The question asker is modeled as a random effect and the independent variables are the four ratings of question types, and value-characteristics of sincerity, urgency and difficulty. Two control variables are included: the number of other questions the asker asked previously (log transformed[2]) and the amount of Mahalo Dollars the asker had earned through answering (log). These variables are included because askers' usage of the system may change over time, especially if they have learned how valuable help on the site may be, or if they have already earned credits that can be used to pay for answers.

[2] Logarithmic normalization used in our analyses is base 10, after adding a base value of 1

As mentioned previously, we used the set of questions posted before February 20 for this analysis (333 questions: 205 paid, 128 free). This way, we could examine askers' decisions without the influence of site sponsorship, and ensure that askers intentionally posted to Mahalo Answers.

Reward Level $1, $2-3, >$3	n=400		Model	
	Mean	SD	Odds Ratio	SE
Factual	0.03	0.44	1.28	0.43
Opinion	0.05	0.42	1.70	0.62
Advice	0.03	0.47	0.88	0.30
Not Question	0.01	0.40	0.99	0.38
Sincerity	-0.02	0.43	1.44	0.61
Urgency	-0.00	0.45	1.03	0.39
Difficulty	-0.00	0.46	4.78***	1.80
Prior Q (log)	1.93	1.42	0.52	0.07
Prior Earn (log)	1.57	2.00	1.24	0.12

*** p < 0.001

Table 5. Random-effects logistic regression model predicting reward value (H2), using user-paid half of the full dataset (n=400).

Table 4 shows the results for the logistic regression. In this analysis, the logistic regression estimates the probability that the question is paid (is paid=1). In logistic regression analyses, the probability is presented in odds, and the odds ratio (fourth column of Table 4) is the odds that a question is paid over the odds that a question is not paid. If the odds ratio is greater than 1, the presence of the predictor variable suggests higher odds that the question is paid, and the inverse. We found that when a question's factual rating is one standard deviation above the mean, the odds of it being paid were 5.19 times higher, which is equivalent to a 30% increase in the probability that the question is paid (based on post-estimation where we held other factors to be at their means and assumed the random effect is 0). Although advice ratings also had a fairly high effect size (24% probability increase), it was not significant in our model (p=0.11). The opinion and not a question ratings were not significant predictors in whether or not a question was paid. Similarly, none of the other characteristics were significant predictors, although urgency rating had a fairly high positive effect (+22% probability) and difficulty had a fairly high negative effect (-26% probability).

Given that all of the user-paid questions are intentionally posted by the Mahalo users, to examine what factors impacted askers' decision on how much to pay, we were able to use all (n=400) of the user-paid questions, instead of only the subset posted before February 20 used previously. We used the same set of independent, control, and random variables. The dependent variable here is the reward value. Because the dependent variable is not normally distributed, it is split up into three tiers: $1 questions (n=235), $2-3 questions (n=104) and $4-100 questions (n=61).

Table 5 shows that the only significant predictor of reward value is the question difficulty—the more difficult the question, the higher the pay. Recall that, interestingly, difficulty was not predictive of whether or not a question was paid. This supports our general intuition that the pay decision is two-staged, and that there are different factors influencing the decisions in each stage. We will discuss this in more detail in the discussion.

Average Answer Length (log)	n=399†		Model	
	Mean	SD	Coef.	SE
Reward ($2-3)		n=104	0.22*	0.09
Reward (>$3)		n=61	0.25*	0.11
Factual	0.00	0.47	0.17	0.10
Opinion	0.03	0.44	0.33**	0.10
Advice	0.01	0.47	0.04	0.10
Not Question	0.02	0.40	0.04	0.11
Sincerity	-0.02	0.44	0.31*	0.13
Difficulty	-0.01	0.46	0.14	0.10
Urgency	-0.00	0.45	-0.06	0.12
Politeness	-0.02	0.37	-0.19	0.12

** p < 0.01, *p < 0.05

Table 6. Random-effects regression model predicting logged average answer length (H4), using user-paid half of the full dataset (n=400).

†One of the questions did not have any answers, making it an outlier in the analysis. This question was removed.

Question Answering

In general, paid questions had more answers than free questions (4.2 to 2.2 answers), but this may be because paid questions were immediately visible on the homepage of Mahalo Answers, while free questions were not (see Figure 1). Therefore, instead of comparing answers between paid and free sections, we focus on how the increase in reward price affects answers. Since there were no significant interface changes to the paid section of the site during our data collection period, we used the full set of paid questions (400) and their answers for the following analyses.

All of the models in this section use the same set of independent and control variables. The independent variable is the question-reward, broken down into three tiers ($1, $2-3, $4-$100). The control variables include the types of questions asked, sincerity, difficulty, and politeness. All are question characteristics that may impact answerers' decision to respond. These models also all use question asker as a random effect.

First, we explore how question-reward affects answer length. Due to the non-normality of average answer length, a log transformation was applied. Table 6 shows that $2 or more questions elicit 22%-25% longer answers (on the logged length) than $1 questions. However, the improvement between middle and high tiers is not significant. Our model also showed that both opinion and sincerity ratings correlated with longer answers (Table 6).

Answer Count	n=400 All (μ=4.2)		n=400 Experts (μ=3.0)	
	IRR†	SE	IRR	SE
Reward ($2-3)	1.00	0.09	0.92	0.08
Reward (>$3)	1.25*	0.14	1.33*	0.15
Factual	0.93	0.09	0.88	0.08
Opinion	1.38**	0.14	1.50***	0.16
Advice	1.18	0.11	1.04	0.11
Not Question	1.08	0.11	1.08	0.12
Sincerity	1.02	0.13	0.98	0.13
Difficulty	0.81*	0.10	0.88	0.09
Urgency	0.08	0.10	0.98	0.11
Politeness	0.06	0.11	0.90	0.10

*** p < 0.001, ** p < 0.01, *p < 0.05

Table 7. Random-effects negative binominal models predicting answering count (H5), using user-paid half of the full dataset (n=400).

†IRR is the incidence rate ratio, which gives a relative measure of the effect of a given variable, like odds ratio.

We then explored the effects of financial rewards on answer count. Two dependent variables were used: the total number of answers to the question and the number of answers from "star" answerers (defined as those with average asker-rating above the median, *i.e.* above 3.66 out of 5). Negative binominal regressions were used for both of these models and the effects were similar (using STATA's xtnbreg command). Table 7 shows that while middle-tiered rewards did not increase the number of answers a question received compared to the baseline $1 rewards, high-tiered rewards (>$3) did. High reward questions had 1.25 time more answers than baseline questions and also 1.33 times more answers from "star" answerers. Control variables show that opinion questions in general got both more answers and more answers from "star" answerers, while more difficult questions got fewer answers (no significant impact on number of answers from "star" answerers).

The last model in this section explores the relationship between reward value and answer quality. In preliminary analysis, we tested two potential measures of quality: whether the asker chose a best answer and whether the best answer had at least a good rating (>=3). Because the results were similar, we will present only the model of whether the best answer had at least a good rating.

Results from our logistic regression model (table 8) show that compared to the baseline, the reward value may increase the likelihood that the answer is rated positively (+8-13% expect change in probability). However, because the effects were not significant, we cannot reject our null hypothesis that higher reward questions are more likely to be positively rated. We should also note there are two

significant variables in this model: opinion rating (+28% probability) and question difficulty (-24% probability).

Is Answer Rated Positively 0=no, 1=yes		n=400		Model
	Mean	SD	Odds Ratio	SE
Reward ($2-3)		n=104	1.76	0.74
Reward (>$3)		n=61	1.42	0.73
Factual	0.03	0.44	0.75	0.34
Opinion	0.05	0.42	3.09**	1.54
Advice	0.03	0.47	0.90	0.42
Not Question	0.01	0.40	1.36	0.69
Sincerity	-0.02	0.43	1.46	0.89
Difficulty	-0.00	0.46	0.38*	0.18
Urgency	-0.00	0.45	2.21	1.16
Politeness	-0.02	0.37	1.10	0.58

** p < 0.01, *p < 0.05

Table 8. Random-effects logistic regression model predicting is best answer rated positively (H6), using user-paid half of the full dataset (n=400).

Archival Use

Can the offered reward values indicate archival value? Since there were no major changes to the paid section of the site, we used the full 400 paid questions for this analysis. We built a random effects regression model with question asker as a random effect. We compared the archival value of the paid questions across the three-tiers of reward values. Our analysis found that high and medium levels of reward do predict higher archival values ($F(2,386)=8.16$, $p<0.001$), but the difference between medium and high levels was not significant ($F(1,391)=0.09$, $p=0.76$). This suggests that while payment values may be used as a way to filter questions for archival purposes, it may be more suitable as a threshold than a precise indicator (Table 9).

	$1	$2-3	>$3
Archival Value Rating	-0.04	0.14	0.16
std. error	0.03	0.04	0.06

Table 9. LS-Mean archival value across 3 payment tiers.

DISCUSSION

Our hypothesis that askers may be affected by cultural norms when deciding how to use the pay-for-answer system is supported by our results. Also, supporting prior findings, higher financial rewards on Mahalo Answers did elicit longer answers and a higher number of answers, but not necessarily better "best answers," as rated by the question askers. Finally, higher payment values indicate higher archival value. We discuss in detail the implications below.

Question Asking

How do question askers choose to pay for answers in a monetary-enabled Q&A system? Prior work has used expected utility to explain why there seem to be more serious questions when askers have to pay for answers—askers pay more if the questions are more valuable [16].

However, our results suggest that the decision of whether or not to pay is not made with a simple cost and benefit analysis, but rather, the decision may be impacted by perceptions about how financial rewards should be used. Past work points out that different types of incentives have different usages and meanings in different scenarios [6,1,15,30]. Financial rewards are used primarily in exchange relationships, but not in social, communal relationships. Applied in the Q&A domain, information seekers may feel that paying is appropriate when they are purchasing facts (and potentially advice) from the answerers. But when askers seek to initiate a conversation, which is the intent of the opinion questions, users may not think financial compensations are necessary, or appropriate.

The exchange and communal relationship does not seem to map directly to the distinction between informational and conversational questions. Part of the problem is that while factual and advice questions are both classified as informational, advice questions actually positively correlate with opinion questions (conversations), as shown by our correlation data. Perhaps a better way to classify the question is to simply rate how informational and how social the question is. Informational and social dimensions may also be a more useful breakdown when designing interaction support. While questions can be both informational and social, questions with high informational ratings will be more functional, with emphasis on getting high quality answers, whereas questions with high social ratings will be more oriented towards generating interesting discussions. For these social questions, perhaps other types of recognition, as opposed to a simple best answer selection, may be better employed (*e.g.* slash dot ratings – controversial, humorous, etc.).

In addition to deciding whether or not to pay, askers must also determine how much to pay. Our analysis of question rewards shows that question difficulty predicts how much askers pay. One reason why our measures of importance and urgency may not predict payment is that they are judged by outsourced raters rather than the asker. Despite that limitation, our result is still interesting because it shows that whether or not to pay and how to pay are two distinctive decisions in pay-for-answer systems. This is perhaps due to the two-stage process used in paying for answers—askers first select if they want to pay, then how much. More research is needed to explore how interface designs may be affecting the use of e-commerce services.

Question Answering

Our examination of Mahalo Answers also builds on prior research in answering research questions on how financial rewards impact answers. Our results support the hypotheses that paying more can result in longer answers and a greater number of answers, though not higher quality best answers (judged by askers, in our study).

As mentioned before, two prior experiments showed that increasing rewards improved answer quality [14,16], while one showed that it did not [4]. One explanation for these results is that the studies with positive effects used ratings from all of the answers, while the other study compared ratings from the single, best/official answer. Our findings support this explanation. While the reward value did not seem to affect best answer quality, we found that questions with higher pay did result in more answers from "star" answerers. Prior work has found that answerers with higher reputation scores, what we have referred to as "star" answerers, provide significantly higher quality answers [4]. This would suggest that while paying does not improve the quality of the individual best answers, higher payment can elicit more high quality answers. Perhaps when judged as a combined whole, the quality may be higher.

There are some subtleties in our results that need to be highlighted. First, our analysis of average answer length shows that there is a diminishing return of reward value on length, perhaps due to a ceiling effect. But at the same time, only high value payments increased the number of answers. One potential explanation is that questions with rewards in the high tier (>$3) are more salient, and more attractive, since majority of paid questions on Mahalo Answers (80%) fell within the $1 and $3 range.

One major difference between Google Answers and Mahalo Answers is that Google Answers only allowed for one answerer to provide the official answer. Our results suggest that the real benefits of a pay-for-answer may be realized when more than one answerer is allowed to contribute per question. Allowing multiple answerers to answer the same question is not only more natural (reducing the need for other answerers to answer using the comment option, as was done on Google Answers), but can also be more beneficial to question askers and the extended community of information seekers by providing them with more answers from "star" answerers on the site and potentially offer more tangential answers for archival readers.

Q&A for Archival Purposes

Existing studies of monetary-based Q&A systems tend to focus on how financial incentives affect the question askers and answerers, who are the direct benefactors of Q&A sites. This focus overlooks and neglects the thousands of indirect (and perhaps, the majority of) users of these sites, who rely on the Q&A archives in their knowledge search. Our findings provide evidence that pay-for-answer can support users searching through the Q&A knowledge archive. We found that questions with higher reward values also had higher archival value. The pricing mechanism, then, should be leveraged when organizing the site for archival use. For example, offered prices may be used as thresholds or weights in the searching algorithm for archival knowledge.

From a practical perspective, Q&A designers and developers should consider ways to leverage this additional information offered by the use of financial incentives. From a theoretical perspective, our results call out the need to provide a more comprehensive view on how financial incentives are affecting the larger Q&A community, the askers, the answerers and the knowledge searchers.

LIMITATIONS AND GENERALIZABILITY

In our analysis of answer quality, we used asker's rating as a proxy. However, it is a weak measure of best answers' actual answer quality because (1) not all questions are rated, (2) askers may use positive ratings to establish rapport with answerers for future interaction, and (3) answerers who pay more may over-rate (cognitive dissonance) or under-rate (have higher standards) the answers they receive. This should be considered when interpreting our analysis of payment's effect on answer quality.

Because our questions were randomly selected, our findings should generalize to other questions on Mahalo Answers. Also, because the questions were collected from a real-world, commercial site, various site-specific features need to be considered when comparing our findings to other pay-for-answer Q&A sites. We have already discussed some of the differences caused by interface differences between Google Answers and Mahalo Answers, but it is also important to keep in mind that Mahalo Answers is a fairly recently launched site. Despite that, most of the findings presented should generalize to other pay-for-answer services, especially to help inform how users behave during the early, and an important, stage of adoption.

CONCLUSION & FUTURE WORK

Using real data collected from Mahalo Answers, we were able to show ways in which financial incentives are used in question asking and answering. Findings show that askers are more likely to pay for factual questions, and askers pay more for more difficult questions. Furthermore, financial payments can result in longer answers and a greater number of answers. We also examined how pay-for-answer may affect the larger Q&A community and provide one example of a positive side-effect of using financial payments. How much a question is paid can be used as an indication of how valuable certain Q&A exchanges are for site visitors searching for references.

Most interestingly, our findings suggest that financial incentive systems may not be appropriate for social or conversational questions. More research is needed to further improve our understanding of how to leverage financial incentives in virtual communities. Part of our planned future work is to study the heterogeneity of user motivations on these pay-for-answer services. If we can differentiate between groups of users with varying goals, and incorporate that into the site design, then these sites will be able to provide a more rewarding experience.

ACKNOLWEGEMENTS

This work is funded in part by NSF Grants IIS-0713509 and IIS-0808711. We thank Scott Counts for pointing us to Mahalo Answers, Max Harper for input on question

classifying, Howard Seltman for help with data analyses, Ian Li, Rick Wash, other proofreaders and reviewers for their valuable feedback.

REFERENCES

1. Aggarwal, P. 2004. The Effects of Brand Relationship Norms on Consumer Attitudes and Behavior, *Journal of Consumer Research*, 31 (June), 87-101.
2. Anderson, J.C., & Moore, L. 1978. The motivation to volunteer. *Journal of Voluntary Action Research*, 51-60.
3. Batson, C. D., & Powell, A. A. 2003. Altruism and prosocial behavior. In T. Millon & M. J. Lerner (Eds.), Handbook of psychology, Volume 5: Personality and social psychology. 463-484.
4. Chen, Y., Ho, T., Kim, Y. 2008 Knowledge Market Design: A field Experiment at Google Answers. *Journal of Public Economics Theory*. To Appear.
5. Chiang, J. 1991. A Simultaneous Approach to the Whether, What, and How Much to Buy Questions. *Marketing Science*, Vol. 10, 4. 297-315.
6. Clark, M.S. and Mills, J. 1993, The Difference Between Communal and Exchange Relationships: What It Is and Is Not, *Personality and Social Psychology Bulletin*, 19, 684-691.
7. Deci, E. L., Koestner, R., & Ryan, R. M. 1999. A Meta-Analytic Review of Experiments Examining the Effects of *Extrinsic Rewards on Intr*insic Motivation. *Psychological Bulletin*, 125, 627-668.
8. DePaulo, B. M., & Fisher, J. D. 1980. The costs of asking for help. *Basic and Applied Social Psychology*, 1, 23-35.
9. Edelman, B. 2004. Earnings and Ratings at Google Answers. Unpublished Manuscript.
10. Fehr, E. and Gaechter, S., Do Incentive Contracts Crowd out Voluntary Cooperation?, *IEW - Working Papers iewwp034, Institute for Empirical Research in Economics - IEW*.
11. Fisher, J. D., DePaulo, B. M., & Nadler, A. Extending Altruism beyond the Altruistic Act: The effects of Aid on the Help Recipient. In *Altruism and Helping behavior*.
12. Greenberg, M. S. & Shapiro S. P. 1971. Indebtedness: An Adverse Aspect of Asking for and Receiving Help. *Sociometry*, Vol. 34, No. 2, 290-301.
13. Harper, F. M., Moy, D., and Konstan, J. A. 2009. Facts or Friends?: Distinguishing Informational and Conversational Questions in Social Q&A Sites. In *Proc. of the SIGCHI '09*, 759-768.
14. Harper, F. M., Raban, D., Rafaeli, S., and Konstan, J. A. 2008. Predictors of answer quality in online Q&A sites. In *Proc. of the SIGCHI '08*, 865-874.
15. Heyman, J. & Ariely, D. 2004. Effort for Payment: A Tale of Two markets, *Psychological Science*, 15 (11) 787-793.
16. Hsieh, G. and Counts, S. 2009. mimir: A Market-Based Real-Time Question and Answer Service. In *Proc.s of the SICHI '09*, 769-778.
17. Ipeirotis, P. 2008. Mechanical Turk: The Demographics. http://behind-the-enemy-lines.blogspot.com/2008/03/mechanical-turk-demographics.html
18. Kraut, R. E., Sunder, S., Telang, R., & Morris, J. 2005. Pricing electronic mail to solve the problem of spam. *Human Computer Interaction*, 20, 195-223.
19. Landis, J. R. and Koch, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*. Vol. 33, 159–174.
20. Leibenluft, J. 2007. A Librarian's Worst Nightmare. Yahoo! Answers, where 120 Million Users Can be Wrong. *Slate Magazine*, Dec. 7, 2007.
21. Mahalo Answers. www.mahalo.com/answers
22. Nam, K. K., Ackerman, M. S., and Adamic, L. A. 2009. Questions in, knowledge in?: a study of naver's question answering community. In *Proc. of the SIGCHI '09*. ACM, New York, NY, 779-788.
23. Raban, D. R. 2008. The Incentive Structure in an Online Information Market. *Journal of the American Society for Information Science and Technology* 59(14): 2284-2295.
24. Rafaeli, S., Raban, D., Ravid, G. 2007. How Social Motivation Enhances Economic Activity and Incentives in the Google Answers Knowledge Sharing Market. *International Journal of Knowledge and Learning* 3, 1, 1-11.
25. Regner, T. 2005. Why Voluntary Contributions? Google Answer! *CMPO Working Paper Series*.
26. Snow, R., O'Connor, B., Jurafsky, D., Ng, A. 2008. Cheap and Fast – But is it Good? Evaluation Non-Expert Annotations for Natural Language Tasks. In *Proc. of EMNLP*, 254-263.
27. Wentzel, K.R. 1991. Social Competence at School: Relation Between Social Responsibility and Academic Achievement. *Review of Educational Research*, 1991; 61: 1-24.
28. Wills, T.A. 1976. Perceptual Consequences of Helping Another Person. Paper presented at the meeting of the American Psychological Association, Washington, D.C., September 1976.
29. Wills T.A. 1978. Perceptions of Clients by Professional Helpers. *Psychological Bulletin*, 85, 968-1000.
30. Zelizer, V.A. 1994. The Social Meaning of Money. Basic Books, New York, NY.