Re-examining Price as a Predictor of Answer Quality in an Online Q&A Site

Grace YoungJoo Jeon, Yong-Mi Kim, Yan Chen
School of Information, University of Michigan, Ann Arbor, Michigan, USA
{yjeon, kimym, yanchen}@umich.edu

ABSTRACT
Online question-answering services provide mechanisms for knowledge exchange by allowing users to ask and answer questions on a wide range of topics. A key question for designing such services is whether charging a price has an effect on answer quality. Two field experiments using one such service, Google Answers, offer conflicting answers to this question. To resolve this inconsistency, we re-analyze data from Harper et al. [5] and Chen et al. [2] to study the price effect in greater depth. Decomposing the price effect into two different levels yields results that reconcile those of the two field experiments. Specifically, we find that: (1) a higher price significantly increases the likelihood that a question receives an answer and (2) for questions that receive an answer, there is no significant price effect on answer quality. Additionally, we find that the rater background makes a difference in evaluating answer quality.

Author Keywords
Question-answering, knowledge market, online community, information quality, information exchange

ACM Classification Keywords
H5.3. Information Interfaces and Presentation (e.g. HCI): Group and Organization Interfaces; J.0 Computer Applications: General

General Terms
Design, Economics, Experimentation, Measurement

INTRODUCTION

Online question-answering (Q&A) sites have emerged as a popular venue for both knowledge exchange and knowledge generation [1, 8]. On these sites, users ask and answer questions on a broad range of topics. The sites usually have features which enable users to ask and answer, to search, and to browse questions. Additionally, such sites have various reputation systems which enable users to evaluate both answers and answerers.

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.

These Q&A sites can be classified into price-based, community-based, or hybrid services [2]. In price-based services, such as Google Answers and Uclue, users post a question along with the price they are willing to pay for an answer, and answerers who are pre-selected by the site or other users answer the question. Community-based services, such as Yahoo! Answers [1], Answerbag [4], and Naver's Knowledge iN [8], allow any user to ask or answer questions for free. Given these different systems, a key question for designers of knowledge markets is the effect of price on the quantity and quality of answers.

Google Answers, which was in operation in the US from 2002 to 2006, is a price-based Q&A service where Google Researchers[1] answer questions posted by askers. Google Answers has attracted researcher attention as it also has characteristics of community-based services, principally its mechanism of allowing users to post free comments in addition to the official answer that can only be posted by a Google Researcher [3, 9, 10, 11]. Two independent controlled field experiments investigated whether the price component of Google Answers has any effect on answer quality. Interestingly, these studies find conflicting results. While Harper et al. [5] report that a higher price leads to higher answer quality on Google Answers, Chen et al. [2] find that price has no significant effect on answer quality.

Our study strives to understand the cause of these conflicting results. Since the two experiments differ in question selection, rater background, rating procedure and statistical analysis methods, we re-analyze the data from the two studies and provide a systematic evaluation of each component of each study. We employ a two-stage statistical method that accounts for sample selection bias [6]. Our re-analysis yields results that reconcile those of the two experiments.

TWO FIELD EXPERIMENTS

Harper et al. [5] conducted cross-site field experiments comparing answer quality from various Q&A sites. Their findings led them to suggest that Google Answers outperforms free services such as Yahoo! Answers, and that a higher price leads to a higher quality answer. They also

[1] Google Researchers were contractors approved by Google after being tested for their research and communication abilities, and were not Google employees.

found that, among community-based services, a larger community base leads to a larger volume of answers. Chen et al. [2] conducted a field experiment on Google Answers and found that the reputation of a Google Researcher has a significant effect on answer quality, while price has no significant effect on answer quality.

	Harper et al. [5]	Chen et al. [2]
Question Selection:	Make up own questions	Use real questions from the IPL
Price Range:	$3, $10, $30	$20, $30, $20 plus unconditional or conditional $10 tip
Rater Background:	Undergraduate English or Rhetoric majors	Masters students in Information Science
Rating protocol:	Rate the package of answers and comments on 8 dimensions	Rate only answers on 7 factors
Unanswered questions:	Included and coded as lowest quality	Excluded from analysis
Results:	Higher price leads to better answers	Higher price leads to longer, but not better, answers

Notes: IPL - Internet Public Library (http://www.ipl.org)

Table 1. Comparison between two field experiments

Table 1 summarizes the key differences between the two field experiments. We conjecture that any combination of the following four factors might have caused the conflicting results. First, [5] used undergraduate students majoring in English or Rhetoric to evaluate the answers, while [2] recruited graduate students trained in digital reference as raters. Second, the rating protocols used in the two experiments were different. Specifically, for each question-answer pair, raters were asked to rate along eight factors in [5] and along seven factors in [2]. Furthermore, the answer and comments were rated together for answer quality in [5], while only the official answer was rated in [2]. Lastly, 13% of the questions in [5] did not receive any answers or comments, while 25% of the questions in [2] did not receive an official answer. [5] included these questions in their analysis and coded them as having the lowest quality on a 1-5 star scale, while [2] excluded unanswered questions from their analysis. In what follows, we evaluate the effect of each of these components on the respective experiment results.

DATA RE-ANALYSIS PROCEDURE

Both the Harper et al. and Chen et al. data were re-analyzed to account for the different answer quality rating procedures and different methods for addressing the issue of unanswered questions. Chen et al.'s rating procedure was applied to Harper et al.'s data set, allowing for a comparison of the two data sets that controls for differences in rating procedures.[2] A two-stage analysis of price effects was applied to the original Chen et al. data, the original as well as re-rated Harper et al. data to correct for the sample selection bias from unanswered questions.

Design of Rating Method

The original 54 questions from [5] consist of (A) 12 questions receiving an official answer only, (C) 26 questions receiving comments only, (AC) 9 questions receiving both an answer and comments, and (N) 7 questions receiving neither an answer nor comments. We rated all 54 questions, using the Chen et al.'s protocol. For questions in A, C and N categories, we rated the answer, comments, and questions, respectively. For the 9 AC questions, to resolve a procedural difference between the two studies in handling comments, we separated them into (1) the answer (as in [2]), (2) answer and comments (as in [5]), and (3) the comments, and rated each component separately in this order. While we acknowledge possible order effects, we only used the rating of the official answer in the Heckman analysis which is not affected by the order effect. The reason for separately evaluating the answer and the comments is because an answer is a response to price, while comments are not.

To evaluate the effect of rater background, we used two separate groups of raters: undergraduate English majors to replicate the rater background in [5] and graduate students in the Master of Science in Information (MSI) program to replicate the rater background in [2]. A total of fifteen raters were recruited at the University of Michigan, with seven juniors and seniors from the English Department, and eight MSI students. These MSI students were recruited from students in the Library and Information Services specialization who had taken SI 647, Information Resources and Services, a course preparing students to perform reference services in settings such as libraries or other information centers. Of fifteen raters, thirteen were female. All were native English speakers, and the majority was in their 20's.

Rating Procedure

Our study asked raters to use Chen et al.'s [2] web-based rating system to rate question-answer pairs in terms of seven factors.[3] However, we changed the instructions for each session to accommodate the answer/comment combinations in our study (A, AC, C).[4] We used three separate rating sessions, each of which lasted approximately two hours, to prevent potential rater fatigue.

In the first session, raters were given training and then asked to rate the official answer for each of the 21 questions in A and AC. In the training session, the raters were asked to rate two Google Answers question-answer pairs outside our question set. A brief discussion regarding the rating

[2] Ideally, we should re-rate the Chen et al. data using the Harper et al. protocol as well. However, Harper et al. used oral instructions for rater training which were not archived. Thus, we were not able to replicate their rating protocol.

[3] Additionally, they also rated question difficulty and the overall quality of the answer.

[4] Rater instructions are posted at http://yanchen.people.si.umich.edu/.

activity followed the rating of each question-answer pair to ensure that all raters have a consistent interpretation of how to rate the pair. We emphasized that we do not expect the raters to achieve consensus, but to rely on their own judgment.

In the second session, participants rated each answer-comments package of the nine AC questions, and twelve questions with comments-only. In the third session, they rated the remaining comments, and filled out a background questionnaire at the end of the session. The interrater reliability for the ratings was greater than 0.8, using intraclass coefficient ICC [3, 15].

FINDINGS

In our re-analysis of the data from the previous experiments, we find that the price effect is two-fold. First, a higher price significantly increases the likelihood that a question receives an answer. However, for questions that receive an answer, price has no effect on answer quality. Additionally, we find that rater background affects the evaluation of answer quality.

Price Effects

We recognize that price has two levels of effect on answers provided through Google Answers. First, whether a question receives an answer might be affected by its price. Second, given that a question receives an answer, price may have an effect on the quality of the answer provided.

Central to the analysis of price effects is the handling of unanswered questions. Assigning the lowest quality score to unanswered questions (as in [5]) might over-estimate the price effects, while excluding those questions (as in [2]) might under-estimate its effects. The Heckman method [6] successfully accounts for such missing values and is widely used by social science researchers. Applying the Heckman method to our study involves estimating two equations: (1) a maximum likelihood estimate probit selection equation that models whether a question receives an answer; and (2) an ordered probit equation that models answer quality in terms of price, question difficulty, researcher experience, and researcher reputation, while controlling for whether a question receives an answer.

The first stage of the Heckman method predicts the likelihood that a question receives an answer as a function of price, question difficulty, and question length (i.e. word count).[5] Table 2 presents our probit regression analysis using Harper et al.'s original rating data, our re-rating of their data, and the Chen et al. data. We find that, in all three specifications, a higher price significantly increases the likelihood that a question receives an answer. Specifically, a $1 increase in price leads to a 1.8% increase in the likelihood that a question receives an answer in the $3-$30 price range of Harper et al. ($p<0.01$), and a 3.1% increase in the likelihood in the $20-$30 price range of Chen et al. ($p<0.05$).

	Harper et al. (original)	Harper et al. (re-rated)	Chen et al. (original)
Price	0.018 (0.006)***	0.018 (0.006)***	0.031 (0.012)**
Question Difficulty	-0.137 (0.128)	-0.145 (0.103)	-0.077 (0.067)
Question Length	-0.001 (0.002)	-0.001 (0.002)	-0.001 (0.002)
Observations	54	54	100

Notes:
a. Probit: standard errors are in parentheses.
b. Significant at: * 10 %, ** 5 %, *** 1 % level.

Table 2. Probit: Likelihood of receiving an answer

The second stage of the Heckman model predicts answer quality, while controlling for the likelihood of getting an answer. Table 3 presents the price effects on answer quality, using an ordered probit specification, again using Harper et al.'s original rating, our re-rating of their data, and the Chen et al. data. We find that price has no significant effect on answer quality in either experiment, while controlling for question difficulty, researcher experience (i.e. total number of questions answered), researcher reputation (i.e. past average rating), and a covariate calculated from the first stage (Inverse Mills Ratio). The last column identifies researcher reputation as the only significant predictor of answer quality, an effect identified in [2] which survives in the Heckman analysis.[6]

	Harper et al. (original)	Harper et al. (re-rated)	Chen et al. (original)	Chen et al. (original)
Price	0.123 (0.105)	0.052 (0.136)	-0.218 (0.155)	-0.207 (0.155)
Question Difficulty	-0.134 (0.933)	-1.074 (1.116)	0.450 (0.507)	0.551 (0.511)
Researcher Experience	-0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Researcher Reputation				1.362 (0.693)**
Inverse Mills Ratio	3.778 (3.213)	2.552 (4.470)	-4.786 (3.966)	-4.635 (3.978)
Observations	21	21	75	75

Notes:
a. Ordered probit: standard errors are in parentheses.
b. Significant at: * 10 %, ** 5 %, *** 1 % level.

Table 3. Ordered Probit: Determinant of answer quality

Rater Background and Rating Procedures

To analyze the effect of rater background on the evaluation of answer quality, we conducted Wilcoxon signed-rank tests, and found that rater background makes a difference in evaluating answer quality. Answer quality is evaluated along seven factors: relevance of answer, thoroughness of answer, source credibility, relevance of links,

[5] None of these three variables is correlated with answer quality, the dependent variable in the second stage of the estimation. The correlation table is available from the authors upon request.

[6] The research reputation measure in the Harper et al. data has no variation and thus is dropped.

summarization of cited information, pertinence of information, and understandability of answer. When including all questions, we found that English majors and MSI students evaluate three rating factors significantly differently: source credibility, pertinence of information, and understandability of answer ($p = 0.004$, 0.000, and 0.001, respectively). We also ran the signed-rank test by answer type, and found the two groups show differences in evaluating A questions only for source credibility ($p = 0.007$). In addition, the rater groups behave differently in evaluating AC questions for four rating factors: thoroughness of answer, relevance of links provided, pertinence of information, and understandability of answer ($p = 0.021$, 0.051, 0.008, and 0.044, respectively). For C questions, they display differences in evaluation for two factors: pertinence of information and understandability of answer ($p = 0.001$ and 0.026, respectively). Overall, we find that MSI students, who are considered semi-professionals, give lower answer quality ratings than do undergraduate English majors, although the average ratings of the two groups are highly correlated ($\rho = 0.85$, $p < 0.01$). Furthermore, MSI students achieve greater interrater reliability than English majors.

For the 9 AC questions, we compared answer quality ratings when an answer is presented without comments, and when it is presented as a package with comments. We found that the average rating of the answer alone is not significantly different from the corresponding average rating of the answer and comments as a package ($p = 0.435$, two-sided Wilcoxon rank-sum test).

CONCLUSION

Re-analyzing the Harper et al. [5] and Chen et al. [2] data, we find results that reconcile the findings of the two previous studies. Of the four factors which might have caused the conflicting results, we rule out rater background and rating protocols (answers vs. answer-comments package), and identify the statistical method for handling unanswered questions as the cause for the conflicting results. Unlike the previous studies, we decompose the price effect into the likelihood of receiving an answer and answer quality. Doing so, we find that: (1) a higher price significantly increases the likelihood that a question receives an answer and (2) for questions that receive an answer, there is no significant price effect on answer quality. Thus, price has an effect, but not on answer quality. These results are consistent with those of other studies examining the effect of pay on the quantity and quality of work [e.g., 7].

Additionally, we find that rater background makes a difference in evaluating answer quality, as MSI students, semi-professionals, give lower ratings and achieve greater interrater reliability than do undergraduate English majors.

Both answer quality and answer quantity are important for the success of a Q&A site. Answer quality is critical for attracting question askers. However, quantity, in the sense of the likelihood of a question being answered, cannot be ignored, as users are less likely to find a Q&A site useful if a large proportion of questions go unanswered. Thus, the provision of incentives promoting quantity and quality of answers is critical to the design of Q&A sites. Our study indicates that the price of an answer is an incentive for quantity but not quality of answers, while reputation systems provide incentives for answer quality. Future work is needed to evaluate the robustness of these findings in other online communities.

ACKNOWLEDGMENT

We would like to thank F. Maxwell Harper, Daphne Raban, Sheizaf Rafaeli, and Joseph Konstan for sharing their data with us, Tom Finholt, Sherry Xin Li, Mick McQuaid, Mark Newman, and Paul Resnick for helpful discussions and comments.

REFERENCES

1. Adamic, L. A., Zhang, J., Bakshy, E. and Ackerman, M. S. Knowledge Sharing and Yahoo Answers: Everyone Knows Something. In *Proc. WWW 2008*, 2008.
2. Chen, Y., Ho, T. and Kim, Y. Knowledge Market Design: A Field Experiment on Google Answers. *Journal of Public Economic Theory*, Forthcoming.
3. Edelman, B. Earnings and Ratings at Google Answers. Manuscript, 2004.
4. Gazan, R. Specialists and Synthesists in a Question Answering Community. In *Proc. American Society for Information Science and Technology,* 2006.
5. Harper, F. M., Raban, D., Rafaeli, S. and Konstan, J. A. Predictors of Answer Quality in Online Q&A Sites. In *Proc. CHI 2008*, ACM Press, 2008.
6. Heckman, J. Sample selection bias as a specification error. *Econometrica*, 47(1), 153-162, 1979.
7. Mason, W., Watts, D. J. Financial Incentives and the "Performance of Crowds". In *Proc. KDD-HCOMP '09*, ACM Press, 2009.
8. Nam, K. K., Ackerman, M. S. and Adamic, L. A. Questions in, Knowledge iN? A Study of Naver's Question Answering Community. In *Proc. CHI 2009*, ACM Press, 2009.
9. Raban, D., The Incentive Structure in an Online Information Market. *Journal of the American Society for Information Science and Technology*, 59(14), 2284-2295, 2008.
10. Rafaeli, S., Raban, D. and Ravid, G. How Social Motivation Enhances Economic Activity and Incentives in the Google Answers Knowledge Sharing Market, *International Journal of Knowledge and Learning*, 2007
11. Regner, T. Why Voluntary Contributions? Google Answers! Technical Report, University of Jena, 2009.