Knowledge Market Design: A Field Experiment at Google Answers*

Yan Chen Teck-Hua Ho Yong-Mi Kim

November 24, 2008

Abstract

We investigate the effects of various design features of online knowledge markets by conducting a field experiment at Google Answers. Specifically, we study the effects of price, tips and reputation systems on answer quality and answerers' effort, by posting real reference questions from the Internet Public Library to Google Answers under different pricing schemes. We find that posting a higher price leads to a significantly longer, but not better, answer, while an answerer with higher reputation provides significantly better answers. Our results highlight the importance of reputation systems for knowledge markets.

Keywords: knowledge market, field experiment, reputation, reciprocity
JEL Classification: C93, H41

*We thank Drago Radev, Paul Resnick, Soo Young Rieh, Yesim Orhun, Hal Varian, and seminar participants at Michigan, and the 2006 North America Regional Meetings of the Economic Science Association (Tucson, AZ) for helpful discussions, Maurita Holland for giving us access to the Internet Public Library database, Alexandra Achen, Tahereh Sadeghi, Xinzheng Shi and Benjamin Taylor for excellent research assistance. Chen gratefully acknowledges the financial support from the National Science Foundation through grant no. IIS-0325837. Any remaining errors are our own. Chen and Kim: School of Information, University of Michigan, 1075 Beal Avenue, Ann Arbor, MI 48109-2112. Email: yanchen@umich.edu, kimym@umich.edu. Ho: Hass School of Business, University of California, Berkeley, CA 94720-1900. Email: hoteck@haas.berkeley.edu.

1 Introduction

One of the most basic ways of finding information is by asking a question to another person. Traditionally libraries have provided this function through their reference services, where a reference librarian answers a patron's question or points them to resources for help. Traditional reference interactions have been one-to-one, with the interaction being transitory and also restricted to the patron and the reference librarian. With the advent of the Web, users have access to a variety of question-answering services on the Internet, ranging from ones based on the traditional library reference model to community-based models. Various terms have been used to refer to these services, such as knowledge market, question-and-answer service (Roush 2006), and question-answering community (Gazan 2006).

Knowledge markets can be categorized into price-based and community-based services. The former includes Google Answers, Uclue and BitWine, while the latter includes Yahoo! Answers, Answerbag and Naver's Knowledge-iN. Design features of these sites differ in who can provide answers and comments, whether to use price and tips, and properties of the reputation systems. From a designer's perspective, it is important to evaluate the effects of each of these design features on the behavior of the answerers and the quality of answers provided. In what follows, we first introduce Google Answers as a representative design of price-based knowledge market. We then introduce the main features of community-based services.

Google Answers was a service introduced by Google in April 2002 and remained in operation until late December 2006. Archived questions and answers are still publicly viewable. A user posted a question along with how much they would pay for an answer, where payment ranged from $2 to $200. The user also had to pay a non-refundable listing fee of $0.50. If the answer provided by a Google Researcher was satisfactory to the asker, the Researcher received 75% of the listed price and Google received 25%. Users could also tip the Researcher, although tipping was not required.

Google Researchers were selected by Google, who were initially recruited openly. An official answer could only be provided by a Google Researcher, but any Google Answers user could comment on the question. According to Google, Researchers had "expertise in online searching," with no claims being made for subject expertise. While Google Answers did not provide a mechanism for users to direct a question to a specific Researcher, users sometimes specified in the question title the Researcher they wanted to handle the question. Once a user posted a question, he or she could expect two types of responses: comments and an actual answer. There is no payment involved in comments, while payment is made for an answer supplied by a Google Researcher. If a satisfactory answer to the question was provided in the comments an answer may not be supplied. Incentive for commenting was that Google claimed Researchers would be recruited from commenters.

Google Answers had a transparent reputation system for Researchers. For each Researcher the following were visible: (1) Average answer rating (1 to 5 stars) of all this Researcher's answers; (2)

Total number of questions answered; (3) Number of refunds;[1] and (4) All the questions answered by the Researcher along with their ratings.

There are currently 53,087 questions available through Google Answers archives. After the closure of Google Answers, a number of Google Answers Researchers went on to establish Uclue, a fee-based question-answering system modeled on Google Answers.

In comparison to price-based systems such as Google Answers, community-based services, such as Yahoo! Answers, do not use money as a medium of exchange. Rather, answerers are rewarded through an elaborate system of points and levels based on the extent and quality of their participation in the question-answering community. A tension that emerges in free sites focusing on the community aspect is balancing community participation with the provision of high-quality answers. The value of these sites comes from the accumulated expertise and experiences of the community members being expressed through answers to posted questions. Points and levels are aimed at rewarding, and thus encouraging, users for providing high-quality answers and to reduce the motivation to participate in non-productive ways, such as carrying on unrelated conversations with other community members in an answer thread. In community-based sites, the extent of moderation varies, ranging from active and visible moderator presence in Ask MetaFilter to the more hands-off approach in Yahoo! Answers. Moderation takes the form of deletion of questions or comments that violate site policies. In general the ability to mark one or more answers as "Best Answer" is restricted to the original question asker, while any registered user may mark an answer as a favorite or award points to it. Free, community-based sites have no barrier to participation other than being a registered user of the site. Thus a question may be answered by both subject experts and novices alike. Mechanisms enabling the best answers and contributors to float to the top become essential in these circumstances.

Table 1: Features of Internet Knowledge Markets

Site	No. questions	Who answers	Price & Tip	Reputation system
Google Answers	53,087	Researchers selected by Google	$2 to $200	1 to 5 stars
Yahoo! Answers	10 million+	All registered users	No	Points, levels
Answerbag	365,000+	All registered users	No	Points, levels
Internet Public Library	50,000+	Librarians and LIS students	No	None

Notes:

1. Google Answers number of questions includes only those that can still be accessed through their archive. Some estimates have placed number of questions around 150,000.

2. According to Yahoo!'s blog, Yahoo! Answers had their 10 millionth answer posted on May 7, 2006.

[1] Askers who were unsatisfied with the answer could demand a refund.

Table 1 presents the basic features of four knowledge markets on the Internet, including the number of questions posted on the site, who answer the questions, whether price and tips are used, and a brief description of the reputation system on the site. In terms of investigating the effects of various design features, Google Answers provided a unique opportunity, as all important design features were used by the site.

In this paper, we investigate the effects of various design features of knowledge markets by conducting a field experiment at Google Answers. Specifically, we study the effects of price, tips and reputation systems on the quality of answers and the effort of the answerers, by posting real reference questions from the Internet Public Library to Google Answers under different pricing schemes. We find that posting a higher price leads to a significantly longer, but not better, answer, while an answerer with higher reputation provides significantly better answers. Our results highlight the importance of reputation systems for online knowledge markets.

The rest of the paper is organized as follows. In Section 2, we present the emergent literature on knowledge markets. Section 3 presents the experimental design. Section 4 describes our hypotheses. In Section 5, we present our analysis and results. Finally, in Section 6 we discuss our results and their implications for knowledge market design.

2 Literature Review

In recent years online question answering communities have drawn the attention of researchers. Question answering services have existed prior to the Web, most notably in the form of library reference services, in which librarians answered the public's questions on a variety of topics. Interaction was one-to-one and confidential, where the asker consulted with an expert (the librarian). Question asking and answering took place on Usenet, but this was not the main purpose of Usenet communities. Online communities devoted strictly to question answering are a fairly recent phenomenon. A number of studies have focused on characterizing Web-based question answering communities, such as Google Answers (GA), Yahoo! Answers (YA), Knowledge-iN and Taskcn.

To our knowledge, Edelman (2004) presents the earliest empirical study of Google Answers, focusing on labor market issues. Using more than 40,000 question-answer pairs collected between April 2002 to November 2003, he found that more experienced answerers received higher asker ratings, and higher rate of earnings. Relevant to our study, he found a positive and significant correlation between answer length and asker rating. In comparison, Regner (2005) studies the pricing and tipping behavior from a contract theoretic perspective, using 6,853 question-answer pairs from Google Answers. He extended and tested a principal-agent model with reciprocity, and found empirical support for the model. Specifically, frequent askers were more likely to tip, providing support for reputation concerns. However, 18% of one-time users also tipped, providing support for social preferences. Finally, answerers put more effort into the answer if the asker had

frequently tipped before.

In addition to economic analysis, knowledge markets have also been studied in Computer and Information Sciences. Rafaeli, Raban and Ravid (2005) provide the most comprehensive empirical study of Google Answers. They collected Google Answers site activity since inception in April 2002 through December 7th, 2004. Incomplete observations from the beginning and end of the sampling period were removed, resulting in a sample of 77,675 questions for the 29-month period of 06/2002 to 10/2004. Of these 77,675 questions, 37,971 were answered, and 21,828 had comments only. Of the answered questions, 23,869 were rated, and 7,504 were tipped. The average dollar value of a question was $19.37, while the average dollar value of an answer was $20.20.

Adamic, Zhang, Bakshy and Ackerman (2008) examined Yahoo! Answers from the perspective of a knowledge sharing community. In their analysis of YA they identified three categories of forums: (1) discussion forums, where questions are asking for opinion and conversation; (2) advice seeking, where the asker is seeking advice or personal experiences of other users, and the question asked may have multiple acceptable answers; and (3) factual or technical forums, where questions usually have a factual answer. Consistent with Edelman's (2004) finding on Google Answers, answer length emerged as a significant factor for predicting best answers across all categories, achieving about 62% prediction accuracy based on answer length alone. The "track record" of a user was found to be more predictive for best answers in technical categories such as Programming, compared to discussion or advice seeking categories such as Wrestling or Marriage. The track record is measured in terms of how many answers from that user within category were selected as best answers. In YA, the asker selects a best answer from the set of answers to his or her question.

Such communities are also popular in the non-English speaking world. For example, South Korea's most popular portal site and search engine, Naver (http://www.naver.com), runs a service called Knowledge-iN (KiN), in which users ask and answer questions (Nam, Ackerman and Adamic 2008). An estimated 4.5 million people use KiN every day. Nam et al. (2008) found that KiN users largely divided into askers and answerers. Users had a tendency to specialize in a small number of topic areas - that is, a user was likely to answer questions in one topic area or a small number of related ones, and not in a large number of topic areas. Through interviews with KiN users, they found varied motivations for participation in KiN, such as altruism, personal learning, and personal interest.

Yang, Adamic and Ackerman (2008) examined expertise sharing in Taskcn, one of a number of "Witkey" sites in China. "Witkey" is the term used in China to refer to a type of website in which "a user offers a monetary award for a question or task and other users compete for the award." In contrast to sites such as YA or KiN, the notion of expertise has been expanded to include the ability to perform a task, in addition to question answering. For example, a Taskcn user may ask for a new logo design. In Taskcn, a requester posts the task or question to the site, along with the monetary amount to be awarded, and deadline for submission. Users then submit their solutions. and upon

the deadline the requester chooses the winner. The winner gets the money, and the site gets a fee. Yang et al. (2008) found that on average a task had 65 contributors. While higher rewards attracted more views, the task reward was uncorrelated with the number of submissions. That is, money was not correlated with participation.

In their examination of user participation in question answering sites, Shah, Oh and Oh (2008) included a comparison of YA and Google Answers. One striking difference they found was in the number of answerers compared to the number of askers in the two sites. In Google Answers, the number of askers was more than one hundred times larger than the number of answerers, while in YA the number of askers and answers was more balanced. Google Answers had a limited number of answerers, or researchers, who were the only ones allowed to provide official answers to questions. In YA any user can answer any other user's question.

These studies relied on data gathered from large-scale crawling of the sites to examine motivations for participation in online question-answering communities. These motivations are examined to address the design question of how to encourage participation. There is an underlying assumption that any kind of participation is positive, with the more participation, the better. Raban and Harper (2008) posit that free riding may be preferable to negative contributions. In online communities, free riding can take the form of non-participation, as in neither asking nor answering questions. Negative contributions can be incorrect answers or references to poor-quality information sources. The quality of the contributions, then, is another aspect that needs to be examined in question answering communities. This leads to the question of what kinds of incentives lead to high quality contributions, such as high quality answers.

Harper, Raban, Rafaeli and Konstan (2008) investigate predictors of answer quality using a field experiment across several online question-answer sites. They find that answer quality was typically higher in Google Answers than in the free sites. In contrast to our finding, they find that paying more money leads to higher quality answers. Among free sites, Yahoo! Answers, where community users can answer or comment, outperforms sites that depend on specific individuals to answer questions. We will compare the results with ours and discuss the difference in Section 5.

Compared with other studies of knowledge markets, our study was the first field experiment conducted on a question-answer community to investigate the design features which might lead to higher answerer effort and answer quality. In addition to the empirical findings, we also develop a rigorous rating protocol for evaluating answer quality, which is a useful contribution to experimental methods.

3 Experimental Design

We design our experiment to investigate the effects of price, tip and reputation in inducing higher efforts and better quality. Specifically, we are interested in whether a higher price will lead to

higher effort and better quality, whether the promise of a tip will induce more effort and better quality, and whether, all else being equal, researchers with a higher reputation score will provide a higher quality answer.

3.1 Question Selection: The Internet Public Library Database

Our first design choice was whether to use real reference questions or to make up our own questions such as in Harper et al. (2008). To preserve the realism of the questions, we decided to use real reference questions from a password-protected database from the Internet Public Library.

The Internet Public Library (IPL) is a nonprofit organization founded at the University of Michigan School of Information in 1995. It provides two major services: a subject-classified and annotated collection of materials and links on a wide range of topics, and a question-answering reference service. Dozens of articles have been written about IPL, and awards received. Several observers conclude that the IPL is "perhaps the largest and most well known" of the free electronic reference and online library collection services (McCrea 2004).

IPL relies on unpaid volunteers to answer these questions. There are two sources of answers. Master students trained to be librarians from 15 universities use the IPL as part of their training. During a typical semester, there are roughly 6-7 classes taught using the IPL, e.g., SI 647 (Information Resources and Services) at the University of Michigan.[2] A second source is volunteers, many of whom used to be students trained on the IPL. Some volunteers want the experience at the IPL to get jobs at commercial question and answering services, such as Google, or 24/7.

Of the $50,000$ questions sent to the IPL, 1/3 to 1/2 were not answered, because they were out of scope, such as legal or medical questions, or because they were obviously questions from quizzes or exams, or because the volunteers ran out of time. On a typical day, the IPL receives 160-170 questions, of which an estimated 40% come from school children, 30% from outside the U.S., and many from reference librarians.

IPL keeps a complete data archive of the questions and answers. From fall 2003 on, each volunteer was asked to write down the actual amount of time it took to answer a question as well as why they pick a particular question. As time spent on a question was of particular interest to us, we selected our questions from the database from fall 2003 onwards in this study. When selecting questions, we use the following criteria. First, a question cannot be answered with a single link or a single piece of information, e.g., on Google, or Google Answers archives. Second, it should be open-ended so researchers can spend variable amount of time answering them. Part of what makes a "good" question is whether the researcher has to do work to find not just an answer, but the most authoritative or clear resources. We want questions where the quality of answers improve with the researcher efforts. For example, the question on women's rights in Afghanistan (GAID # 543515,

[2]In SI 647, each student is required to answer 12-15 questions for the IPL. Students are advised to pick a question in the area of their expertise.

Appendix B), fall into this category. A researcher can spend as little as 20 minutes to find some information, but can spend up to five hours to find and sort the information. Lastly, it should fit into one of the ten existing categories of the Google Answers Archive.[3]

We went through two rounds of selection. In the first round, we selected questions which were answered in about an hour by the IPL volunteers. We used time an approximation for the difficulty level of a question. In the second round, we asked our research assistants to independently search for an answer for each question. We discarded any question for which an answer could be found within ten minutes or for which an answer could not be found within thirty minutes. The discrepancy in time spent searching for an answer between the IPL volunteers and our research assistants were largely due to the exponential increase of information on the Web. For example, a question which took an hour to answer in 2003 might be answered in five minutes in 2005. At the end of two rounds, we selected 100 questions, each with two answers prepared by an IPL volunteer and one of our research assistants respectively.

3.2 Treatments

To prepare for the experiment, in June 2005, we downloaded 10,317 questions and answers from Google Answers public archive, uniformly distributed across the categories. Table 2 presents the summary statistics of these questions and answers.

Table 2: Summary Statistics from 10K Downloaded Questions from Google Answers

Price Range	% answered	% adding tip	mean price	median price	tip/price	OBS
[$0, $5]	38.2	13.2	3.3	2.6	1.21	4570
($5, $10]	36.6	19.6	7.2	7.0	0.52	2077
($10, $25]	36.0	17.0	17.8	20.0	0.42	2078
($25, $100]	39.0	19.5	46.0	50.0	0.29	1380
($100, $200]	45.8	19.6	180.2	200.0	0.20	212
[$20, $40]	34.9	18.2	24.4	23.6	0.35	1871
[$0, $200]	37.7	16.2	18.4	10.0	0.71	10317

Based on the statistics reported in Table 2 and the pricing tips offered by Google (Appendix A), we chose our price and tip parameters. We price our questions in the $20-$40 range based on the following considerations. First, questions in this range typically require at least 30 minutes of work, e.g., most of the questions from the IPL archive were answered between 30 minutes and 90

[3]The ten categories were (1) Arts and Entertainment, (2) Business and Money, (3) Computers, (4) Family and Home, (5) Health, (6) Reference, Education and News, (7) Relationships and Society, (8) Science, (9) Sports and Recreation, and (10) Miscellaneous.

minutes. Second, questions priced in this range get rapid attention, and therefore are more likely to get answered by GA Researchers. We designed the following four treatments for our experiment.

1. $20 fixed price: $20 per question, with no condition attached. Based on Google Answers' rule, 25-percent of the price was taxed by Google, while tips were not taxable. Based on statistics from our 10K downloaded questions, given a question was answered, there was a 16.2-percent chance that the researcher would receive a tip of an average $3.48. Therefore, if a researcher answered a question in this category, her expected earning was $15.56.

2. $30 fixed price: $30 per question, with no condition attached. Again, given that 16.2 percent of the answers were tipped in Table 2, taken into account the Google tax, the expected earning in this category was $23.06.

3. $20 plus an unconditional $10 tip: each question in this category is priced at $20, with a promise of a $10 tip. We used the IPL questions with a sentence added at the end promising a tip. We varied the sentences so that they sounded slightly different in each question (see Appendix B). We had a total of 25 questions in this category, 18 of them received an answer. All 18 answers received a tip. The expected earning was $25 in this category.

4. $20 plus a conditional $10 tip: when we sent out each IPL question, we again added a sentence at the end promising a $10 if the question was answered satisfactorily. In practice, if our research assistants judged the answer to be worthy of at least four stars, we added the tip. Seventeen out of 25 questions in this category were tipped $10 after receiving the answer. The expected earning was $21.80 in this category.

A comparison of Treatments 1 and 2 enables us to examine the price effect, while a comparison of Treatments 1 and 3, 1 and 4 enables us to evaluate the tip effect. Lastly, a comparison of 3 and 4 enables us to compare the effects of conditional and unconditional tips. We note that while the two fixed price conditions were common in Google Answers, to our knowledge, *ex ante* promised tips were non-existent prior to our experiment.

3.3 Experimental Procedure

We sent out 100 questions in July, October and November 2005. We space the entire sample of questions over a five-month period so as not to dramatically change the overall distribution of new questions on Google Answers. Each day, four questions were sent to Google Answers, one from each treatment.

To avoid asker side reputation effects,[4] we decided to use a different GA identity for each question. Therefore, our research assistants used a different GA user name for each question.

[4]In Google Answers, askers could develop a reputation in various dimensions, such as the types of questions she

Once a question was posted, if one of the GA researchers was interested in answering it, he locked the question so no other researcher could answer it simultaneously. A locked question must be answered within four hours for questions less than $100 or eight hours for questions $100 or more, beyond which the question was automatically released. Sometimes researchers asked clarification questions before posting the answer. Once an answer was posted, the asker could decide whether to pay for the answer or not. If she decided to pay, the posted price was automatically deducted from the her credit card. The total number of refunds for each researcher was recorded in the GA archive. We paid for all answers to our questions. After the answer was posted, the asker had the option to rate it from one to five stars, one being "poor" and five being "great" according to GA. Our research assistants rated every answer to our questions. However, in the analysis, we excluded their ratings. Instead, we used rater ratings for the quality of answers.

If a question was not answered within a month, it was automatically closed. By the end of November 2005, 55 out of 76 questions were answered, most of which within a week it was posted. The remaining questions were closed within a month it was posted. In December 2005, we posted the remaining 24 questions from our set of 100 and reposted 15 of the unanswered questions under new user IDs. Of these questions, 21 were answered. Therefore, of the 100 questions we posted, 76 were answered. Seventy-two were answered in the first posting, and four were answered in the second posting. Of these 76 answered questions, one was excluded from analysis because the formal submitted answer referred to the comments without providing content of its own. A list of all 75 questions, together with their GA ID number and categories, are provided in Appendix B.

Ideally we would like to observe the amount of time a researcher spent on an answer, however, such data were not available in the public archive. Therefore, we resort to informal survey of the researchers. In 14 out of 76 cases, as soon as an answer was posted, our research assistants asked the researcher directly how long it took them to answer the question.[5] All of the researchers who were asked gave us time estimates. In Section 5, we will correlate the reported time and the answer length for these 14 questions, and use the result as the basis for using answer length as a proxy for effort.

4 Hypotheses

In this section, we describe our hypotheses comparing outcomes from these four treatments, and comparing our experimental data with other GA data.

We use two outcome measures, effort and quality. Our measure of effort is the length of an

asked, how she rated answers, amount of tips if any, and number of refunds demanded. For example, in GA 777817, Researcher Tutuzdad-ga started the answer by "Thank you for allowing me to answer another one of your interesting questions ..."

[5]We thank Paul Resnick for suggesting this approach. We decided to survey only 14 GA researchers rather than everyone in order not to reveal the ongoing experiment.

answer using word count. Our quality measure is based on rater data, which we will describe in more detail in Section 5. In what follows, we will state the alternative hypotheses with the corresponding null hypotheses being no difference.

Based on social preference theories and the empirical support in Regner (2005), we expect that answerers will reciprocate a higher price with more effort and better quality.

Hypothesis 1 (Reciprocity: effort)**.** *A question with a higher price generates an answer involving more effort.*

Hypothesis 2 (Reciprocity: quality)**.** *A question with a higher price generates an answer with better quality.*

Similarly, an *ex ante* promised tip should induce higher effort and better quality.

Hypothesis 3 (Tip: effort)**.** *A conditional (or unconditional) tip generates an answer involving more effort compared to a fixed price $20 question.*

Hypothesis 4 (Tip: quality)**.** *A conditional (or unconditional) tip produces a higher quality answer than a fixed price $20 question.*

Comparing conditional and unconditional tips, we expect that researchers might put in more effort for unconditional tips because of the trust implied in such tips.

Hypothesis 5 (Unconditional vs. Conditional Tips)**.** *An unconditional tip produces a better answer than a conditional tip.*

Lastly, we examine the effect of reputation on the quality of answers. Past research shows that reputation plays an important role in the well-functioning of online transactions. Resnick, Zeckhauser, Swanson and Lockwood (2006) conducted a randomized experiment on eBay, which showed that the difference in buyers willingness-to-pay for identical vintage postcards for a high reputation seller versus a new seller was 8.1% of the selling price.

Hypothesis 6 (Reputation)**.** *Answerers with higher reputation will provide better answers.*

5 Analysis and Results

In this section, we present our data analysis and results. We use human raters for answer quality analysis, a commonly used procedure in Information Retrieval and Psychology, but less common in experimental economics.[6] Therefore, we describe the rating procedure in detail in subsection 5.1.

[6]Landry, Lange, List, Price and Rupp (2006) use raters for the attractiveness of solicitors in a fund-raising experiment, using techniques developed in Biddle and Hamermesh (1998). Zhang (2008) uses a content analysis procedure similar to ours.

To compare our raters' quality rating with that of the real users in Google Answers, we randomly selected 125 question-answer pairs from the 10K questions we downloaded from GA and had our raters evaluate them, in addition to the 75 question-answer pairs from our experiment. The additional 125 pairs also enable us to examine the robustness of our findings on a wider range of prices and tips. In subsection 5.2, we present our results on the 75, as well as those on the entire set of 200 question-answer pairs.

5.1 Rating Procedure

The literature examining peer review of manuscripts in a number of fields (see, e.g., Strayhorn, McDermott and Tanguay (1993) van Rooyen, Black and Godlee (1999) Wood, Roberts and Howell (2004)) was examined for guidance. In our study, raters were expected to provide objective assessments of the quality of the answers. In peer review, reviewers are expected to provide objective assessments of the quality of the manuscript under review. Often they are asked to rate the overall quality of the manuscript on an ordinal scale in addition to the recommendation to accept or reject the manuscript. A study on the reliability of manuscript reviews in psychology (Strayhorn et al. 1993) found that interrater reliability[7] for the overall quality rating improved when raters had to provide ratings for a number of aspects of the manuscript prior to providing the overall quality rating. Reliability could also be improved by training the raters in the rating procedure, and averaging the scores of more than one rater. These strategies for improving reliability were taken into account for the rating procedures used in our study.

Sixteen raters were recruited from graduate students at the University of Michigan who had taken the course, SI 647, Information Resources and Services, in the past two years. This course is usually taken by students in the Library and Information Services (LIS) specialization in the Master of Science in Information (MSI) program. The course prepares them for reference services in settings such as libraries or other information centers, requiring students to work with actual reference questions submitted to the Internet Public Library. Each student is required to answer 12 to 15 IPL questions. Students thus gain expertise in searching, evaluation of information resources, and how to answer questions submitted online.

Initially six raters took part in rating sessions in May 2006. Another group of four took part in September 2006, followed by a third group of six in March 2007. There were two sets of 100 Google Answers question-answer pairs each to be rated. Set A was composed of 75 questions selected from the IPL and 25 questions selected from the 10K downloaded from the GA archives, and set B was composed of 100 questions selected from the 10K downloaded from GA archives that

[7]Interrater reliability provides "an indication of the extent to which the variance in the ratings is attributable to differences among the objects rated"(Tinsley and Weiss 2000). Interrater reliability examines the relative ordering of the rated objects. In contrast, interrater agreement measures the extent to which raters assigned the exact same rating to objects, and is thus sensitive to rater characteristics.

did not overlap the questions in set A. Set A was rated by the May 2006 group and 2 of the March 2007 group, while Set B was rated by the September 2006 group and 4 of the March 2007 group. Thus each question was rated by eight separate raters. Of the sixteen raters, twelve are female. All of them are native English speakers. Their undergraduate major areas were predominantly in the humanities, with seven of the raters having majored in English. The majority of the raters were in the 21-30 age group.

The same procedure was followed for all three groups of raters. Raters were presented with 100 question-answer pairs from Set A or Set B. For each question-answer pair, raters had to provide nine ratings, as shown below:

1. Please rate the difficulty of the **question**. (1 = very easy . . . 5 = very difficult)

2. Please rate the **answer** for the following factors:
   (1=strongly disagree . . . 5 = strongly agree, NA = Not Applicable)

   (a) The question that was asked is answered.

   (b) The answer is thorough, addressing all question parts.

   (c) The sources cited are credible and authoritative.

   (d) The links provided are to relevant web sites or pages.

   (e) Information in the cited sources is summarized.

   (f) Only information pertinent to the question is presented.

   (g) The answer is well-organized and written clearly, avoiding jargon and/or inappropriate language.

3. Please rate the overall quality of the **answer**. (1=very low quality ... 5=very high quality)

Training session All sixteen raters took part in a training session, in which they were asked to rate two question-answer pairs from Google Answers (but not in the set of 200 pairs to be rated). Raters were also asked to fill out a background questionnaire. For the training session, the question-answer pairs were viewable online through a web browser, and raters were provided with paper rating forms. For each of the training question-answer pairs, there was a discussion regarding the rating activity after all raters had completed rating that question-answer pair. For each rating, all raters were asked for their ratings and their rationale for their ratings. Clarification was provided for the rating instructions if requested. The rating coordinators also presented their ratings and their rationale. At the end of the training session the rating sheets were collected from the raters. The purpose of the discussion was for the raters to make explicit to themselves their own individual rating scales, not to establish consensus among the raters. Raters were explicitly asked in the rating instructions to rely on their own judgment when rating. The training instructions are included in Appendix C.

Table 3: Inter-rater Reliabilities: Intraclass Correlation Coefficient

Question Set	Difficulty (Q1)	Overall Quality (Q3)	Summed (Q2a-g)
A (IPL: 75)	0.71	0.77	0.78
A (GA: 25)	0.86	0.77	0.73
A (All: 100)	0.77	0.77	0.77
B (GA: 100)	0.89	0.72	0.72

Rating session For each group of raters, five sessions of two hours each were scheduled. The training session took part during the first rating session. The rating sessions took place in a computer lab. Raters were asked to go at their own pace, although there was a daily limit of 25 question-answer pairs to be rated to avoid fatigue. All rating sessions are completed between one and two hours. Raters were paid a $15 per hour flat fee to compensate for their time, as is standard in the field of Information Retrieval.

Rating was done using a web-based system. Once the rater had provided all the requested responses and clicked Submit, the next question-answer pair to be rated were displayed along with a blank rating form. If a rater had not provided all ratings, a pop-up window informed them of this and the rater was not allowed to proceed to the next question-answer pair until the form had been completed. Once a rater had rated the daily limit, a goodbye screen was displayed.

Raters were provided with unique logins and passwords to allow separate logging of their responses. The order of question-answer pairs was randomized for each rater. All identifying information, such as GA question ID number and answerer identities, as well as the price and reputation scores, had been removed for the training and rating sessions.

Interrater reliability was assessed with the intraclass correlation coefficient (ICC[3,8],[8]) which is a multi-rater generalization of the more familiar Cohen's Kappa used for the two-rater case. Table 3 shows the reliability statistics for the two groups of raters, A and B. In general, values above 0.75 represent excellent reliability, values between 0.40 and 0.75 represent fair to good reliability, and values below 0.40 represent poor reliability. Good to excellent reliability was observed for the ratings. The internal consistency of the multi-item scale (Q2 parts a-g) was high, with Cronbach's alpha 0.84. The alpha value indicates that items in the multi-item scale are highly correlated, reflecting a single underlying construct. Pearson correlation of the summed ratings (Q2 a-g) and the overall ratings (Q3) ranged from 0.75 to 0.92 for Group A and from 0.74 to 0.95 for Group B. These results indicate that the multi-item scale and the overall rating are measuring something very

[8]There are six main cases of intraclass correlation coefficient (ICC), and are distinguished by the numbers in parentheses following the letters ICC. The first number indicates the statistical model assumed. Case 3 assumes that judges are fixed and not drawn from a random population. The second number indicates the number of raters. More details for computing ICC can be found in Shrout and Fleiss (1979)

similar. Based on these results, Q3 will be used to represent quality of the answer in subsequent analysis.

To compare rater and consumer evaluations of answer quality, we randomly selected 125 question-answer pairs from the 10,000 questions downloaded from GA archives. We then ran an OLS regression, with the median rating as the dependent variable, and consumer rating as the independent variable. The coefficient for consumer rating is 0.847 (standard error = 0.02, $p < 0.01$, $R^2 = 0.93$). This indicates that our raters on average gave lower ratings to the overall answer quality than the consumers who asked the questions, which might reflect the fact that our raters were trained semi-professionals who might be more critical of answer quality.

5.2 Results

We first examine the price effects on researcher effort and answer quality. Social preference theory, especially the theory of reciprocity, suggests that a higher price should induce more work from the answerer, which was also implied by the Google Answers Pricing Tips (Appendix A). As the amount of time a researcher spent on an answer was not observable, we use answer length as a proxy for researcher effort. Recall that, in 14 cases, our research assistants directly asked the researcher the amount of time she used to answer a question. Based on these 14 cases, we find that the correlation between the number of words in an answer and the reported time is 0.635 ($p = 0.015$). Therefore, we use word count as a proxy for effort.

Table 4 reports three Tobit specifications, with Word Count as the dependent variable. Each specification includes the following independent variables: the price of the question, the tip, the past reputation score of the researcher, and experience approximated by the total number of questions the researcher has answered in the past. Specification (1) includes the 75 IPL question-answer pairs, (2) includes the 125 GA question-answer pairs, while (3) includes all 200 question-answer pairs.

We find that, in (2) and (3), a higher price leads to significantly longer answers. Furthermore, in (1) and (3), an answerer with a higher reputation score provides significantly longer answers. This provides empirical support for reciprocity, i.e., when a user posts a higher price, the answerer is more likely to put in more effort and provide a longer answer, controlling for the experience of the researcher. By specification (3), we reject the null in favor of Hypothesis 1. Lastly, we note that the promise of tips has no significant effect on the length (or effort) of answers in (1), by which we fail to reject the null in favor of Hypothesis 3.

Next, we investigate the determinants of answer quality. Table 5 presents three ordered probit specifications, where the dependent variable is the median quality rating of an answer. Again, the three specifications correspond to the 75 IPL, 125 GA and all 200 question-answer pairs. While price is no longer significant, researcher reputation is still significant in specifications (1) and (3). This indicates that a higher price does not necessarily lead to a higher quality answer. However,

Table 4: Determinants of Answer Length (Effort)

	Dependent Variable: Word Count		
	(1) IPL	(2) GA	(3) All
Price	7.472	13.097	12.575
	(23.035)	(2.545)***	(2.128)***
Tip	16.862	25.519	27.115
	(21.164)	(19.357)	(13.796)*
Reputation	1,368.709	1,073.285	1,143.625
	(434.286)***	(613.795)*	(413.810)***
Experience	-0.244	-0.136	-0.168
	(0.128)*	(0.126)	(0.095)*
Constant	-5,083.371	-4,259.175	-4,448.396
	(2,011.381)**	(2,687.905)	(1,801.455)**
Observations	75	125	200

Notes:

a. Tobit: standard errors are in parentheses.

b. Significant at: * 10-percent; ** 5-percent; *** 1-percent level.

a research with a higher reputation score provides significantly better answers. Therefore, while we fail to reject the null in favor of Hypothesis 2, we can reject the null in favor of Hypothesis 6. The reputation effect in our experiment is consistent with a similar effect in YA in Adamic et al. (2008). Additionally, we note that the promise of tips has no significant effect on the quality of answers, by which we fail to reject the null in favor of Hypothesis 4.

In contrast with our finding that price has no significant effect on quality, Harper et al. (2008) find that a higher price leads to higher quality answers. We point to two major differences in our quality rating procedures and speculate that either or both might have caused the difference in our results. First, our raters were semi-professionals trained in Information Sciences, while theirs were undergraduate English majors. As discussed in Edelman (2004) and Adamic et al. (2008), users who are not information professionals are more likely to give higher ratings to longer answers. We conjecture that their quality ratings might be a judgment of effort rather than quality *per se*. Second, our raters only rated the official answer, while theirs rated the official answer and the comments by the community as a package. Since only the researcher providing the official answer received the money, correct estimation of the price effect on quality should not include the comments.

Finally, we compare the answer quality under conditional versus unconditional tips, and find no significant difference ($p = 0.633$, Wilcoxon ranksum test). Therefore, we fail to reject the null in favor of Hypothesis 5.

Table 5: Determinants of Answer Quality

	Dependent Variable: Quality Rating		
	(1) IPL	(2) GA	(3) All
Price	-0.035	-0.000	-0.001
	(0.035)	(0.002)	(0.002)
Tip	-0.009	0.001	0.005
	(0.032)	(0.016)	(0.013)
Reputation	1.358	0.742	0.996
	(0.670)**	(0.501)	(0.396)**
Experience	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Observations	75	125	200

Notes:

a. Ordered probit: standard errors are in parentheses.

b. Significant at: * 10-percent; ** 5-percent; *** 1-percent level.

In sum, we find that posting a higher price leads to significantly longer (hence involving more effort), but not necessarily better (higher quality) answers. The price effect provides some support for reciprocity in knowledge markets. A robust finding from the analysis is the effect of researcher reputation. A researcher with a higher reputation score provides significantly longer and higher quality answers. In contrast, the promise of tips do not seem to affect either researcher effort or answer quality.

6 Discussions

Thanks to the Internet, the question-and-answer knowledge exchange market has become more common. Such exchange markets are deeply interesting because they allow knowledge sharing on a global scale. In this paper, we have made an initial attempt to study a price-based knowledge exchange market and analyze how three features of the market affect answerers' effort and answer quality. Our ultimate goal is to understand which features work and which fail in order to better design such markets, which are still evolving rapidly. Using a field experiment, we have systematically varied the asker-posted price and the level of tips for providing high-quality answers in a market where there exists a system for reputation building by answerers.

Consistent with reciprocity, we find that the asker-posted price increases answerer effort. That is, answerers spend more time in a question by providing a longer answer when askers post a higher price. However this extra effort does not translate into a higher answer quality. The question is

why don't answerers invest the extra time into increasing the quality of answers? Our conjecture is that answerers believe that askers judge "fair" wage to be both a function of quality of answers and answers' effort. In the event that they do not know how to increase quality, they can increase their effort to justify their wage. Since both parties may be aware that answerers' effort is proportional to the answer length, answerers can convince askers about their effort by providing a longer answer.

Interestingly, the level and type of tip does not have an effect on both answerers' effort or answer quality. If tips were to be treated as part of the total compensation, one would have expected to see its effects. Our conjecture is that answerers may interpret tip as askers' self-signal for their "niceness". Put differently, if answers were allowed to rate askers, then the former would be more likely to give the latter a higher rating if a tip was used. Future research can investigate this question further.

Lastly, we find that an answerer's past reputation has an effect on both their efforts and answer quality. In a world of anonymous interactions, reputation becomes the most powerful way of signaling quality. In Google Answers, reputation can have two kinds of payoffs. First, answerers with higher reputation may be more sought after by askers. Indeed, we sometimes observed that an asker would request that a particular answerer took her question, which seemed to be always honored. Second, their outputs are more likely to be perceived favorably by askers in cases where there may be uncertainty in determining quality. Hence, answerers with high reputation spend more time and produce high quality answers. This result also suggests that having a system allowing exchange parties to build reputation is a crucial feature for achieving high efficiency in knowledge exchange markets.

APPENDIX A. Google's Pricing Tips

Your price will also impact the length and level of detail in your answer, the breadth of resources consulted, and the amount of analysis done by your Researcher.

The more you pay, the more time and effort a Researcher will likely spend on your answer. However, this depends somewhat on the nature of your question.

Above all - try to pay what the information is worth to you, not what you think you can get it for - that is the best way to get a good answer - but only you can know the value of the information you seek.

Sample Questions at Various Price Points

$2 - $5

- Can be answered with a single link or a single piece of information. Sometimes, if a researcher is personally interested in the question's subject, they may provide a longer answer.
- Not appropriate for multipart questions.
- Only 60% of the questions asked in this price range are answered.
- Examples:
  How do I rid my apartment of ants?
  Question about baseball player Ted Williams and his batting average.
  How can I find the best money market funds?

$10-$15

- Can be answered with 30 minutes of work
- Most questions priced at least at $10 get answered
- Examples:
  Where can I find this type of patio umbrella stand?
  What do the symbols on the World Cup Soccer hat cap mean?
  Please give me a brief history of the term "economic development."

$20-$50

- Typically require at least 30 minutes of work
- Questions are very likely to get answered by Researchers
- If time is a concern for you, questions priced in this range get rapid attention

- Examples:

  What's the difference between witches, warlocks, and sorcerers?

  What's the easiest way to a flat tummy?

  What was the first military use of hepatitis-b vaccine?

$50

- The minimum price appropriate for complex, multi-part questions
- Researchers will typically spend at least one hour on $50 questions and be very responsive to follow-up questions
- If time is a concern for you, questions priced in this range get very rapid attention
- Examples

  What are the mechanics of migraines?

  Market statistics for cookbooks and cooking magazines

  Find federal and state government IT contract documents

$100

- Researchers will typically spend between two-four hours and will do highly thorough research for questions at the $100 price level
- Please be sure to check questions priced at this level frequently as researchers are likely to have questions for you about the answer you want
- If time is a concern for you, questions priced in this range get very rapid attention
- Examples:

  Parking in New York City

  Searching for a medical practitioner

  How does infant-family bonding develop?

$200

- Researchers will typically spend extensive amounts of time (4 hours plus) and do highly thorough research for questions at the $200 price level
- Please be sure to check questions priced at this level frequently as researchers are likely to have questions for you about the answer you want.
- If time is a concern for you, questions priced in this range get extremely rapid attention

- Examples:

  Searching for Barrett's Ginger Beer

  Applications using databases

  What is the impact of a baby with Down's Syndrome on its family?

APPENDIX B. List of Questions

We provide the list of 75 IPL questions sent to Google Answers and received answers, together with each GAID number, the subject line and the category in square brackets. The answers, while not provided due to space limitations, are available from the authors upon request, or from the Google Answers online archive by searching the GAID.

Treatment: $20 Fixed Price

1. 542434: Women's rights in Indonesia [Reference, Education and News]

   Can you provide me with information about women's rights in Indonesia? I'm interested in learning more about both the current situation and the history of women's rights in that country.

2. 542730: Dividends & Stocks [Business and Money]

   If a company increases its cash dividends, what happens to the value of the stock of the company?

3. 543096: Reading list [Family and Home]

   My 10 year old son doesn't like to read. He needs to improve his reading skills though so I want to try to get him interested. Where can I find lists of books that he might like?

4. 584325: Suggested Reading Before College [Reference, Education and News]

   I was told that there is a list of suggested books that someone should have read before beginning college. Do you know where I could find such a list? The public library here told me that each school has its own list, and I wondered if there is a generic one.

5. 584803: First ATM [Miscellaneous]

   I am trying to find information on the first Automated Teller Machine. I believe that Bank One (City National Bank at the time) introduced the first ATM in Columbus, Ohio. I am trying to find out the date it was introduced and any facts about its operation.

6. 585260: Men's Reactions to Independent Women [Relationships and Society]

   Could you please tell me where I might find recent (last couple years) information on how men today are reacting and feeling about powerful, financially independent women? This information could include both when there is a backlash against these kinds of women as well as when there are positive reactions from men about these women. Thank you very much for your help.

7. 586131: Kids Starting Businesses [Business and Money]

   My son is 10 and my daughter is 11. Each of them has profitable goals of starting a small business. I have tried every search engine and bookstore to find a book of business start up basics for kids by kids. I thought about starting out with Whiz Teens in Business, but I am picky in the sense that I don't want anything that cramps their creativity.

8. 586765: Erik Erikson [Reference, Education and News]

   Where can I find in-depth information about Erik Erikson's stage of industry vs. inferiority?

9. 587163: Racecars [Sports and Recreation]

   Where can I find basic information about racecar knowledge to share with young people in our program?

10. 588133: First English cookbook [Family and Home]

    What was the first cookbook published in English? What kind of recipes did it have? Where can I find that kind of recipes online?

11. 589092: Info about death of saleman [Arts and Entertainment]

    The play death of a salesman by Arthur Miller is said to be a love story between a father and his son. Is this statement thought to be true? Why? I would appreciate someone giving me a little more understanding to this question. Thanks you.

12. 592879: Do SAT Scores Predict College Success? [Reference, Education and News]

    How well do high scoring SAT students do in their first year in college? More generally, how good a predictor of college success is the SAT?

13. 602776: Genetically Modified Crops [Health]

    I know there is great concern over how genetically modified crops could affect people. Does any scientific evidence exist that this has happened?

14. 603329: Patricia Polacco [Reference, Education and News]

    Can you give me information on Patricia Polacco?

15. 603798: Light and Bacteria [Science]

    How does light affect the way bacteria grows?

16. 604460: Contests in Vermont [Reference, Education and News]

    Why are residents of Vermont not required to include return postage for contests?

17. 606608: Does noise affect our ability to memorize facts and details? [Science]

    Does noise affect our ability to memorize facts and details?

Treatment: $20 Price with a Promised Unconditional $10 Tip

Please note that the tag line for tips (in italics) were added by our research assistants. They were not part of the original IPL questions.

1. 542436: Laugh Tracks on TV [Arts and Entertainment > Television]

   I heard somewhere that all (or at least most) of the laugh tracks used on television today came from one show several decades ago. Is this true? If so, what show? *$10 tip to the researcher who can answer this one for me.*

2. 542733: History of Easter rabbit [Reference, Education and News]

   I would like to know the history of why the Easter rabbit is associated with the Easter holiday, where this originated and how long ago. *Also, if you answer my question, there will be a $10 tip for you.*

3. 543099: Becoming a library paraprofessional [Reference, Education and News]

   I would like some information about becoming a library paraprofessional. I am applying for a job as a circulation clerk at a library and I would like to know as much as I can about what that means. Are there resources online for library paraprofessional? Maybe some sort of groups, or websites for professional development? Any info about salaries or work conditions, etc.? Really, any sort of info you have. *I'll tack on an extra $10 once I get an answer!*

4. 543643: Women guillotined in French revolution [Reference, Education and News]

   I need names and a little information about women who were guillotined in France at the time of the French Revolution, other than the obvious ones like Marie Antoinette. *For your efforts, I will give you ten more dollars as a tip if you answer this.*

5. 584329: Origin of Phrase "fox's wedding day" [Miscellaneous]

   Do you know the origin of the phrase "a fox's wedding day,?" when it is sunny and raining at the same time? My mother always told me this when I was a child, and I would dearly love to know where it comes from. *$10 tip to the researcher who answers!*

6. 584808: Psychology Careers [Reference, Education and News]

   I would like suggestions for resources on non-Ph.D. careers in psychology. My husband is interested in the field but does not want to pursue a full Ph.D. Areas such as guidance and

human development are of interest to him. *I'll give an additional $10 in tip once I receive an answer.*

7. 586139: Health Insurance Overseas [Business and Money]

   Two factors that seem to be slowing the jobs recovery in the US economy are overseas outsourcing and the rapid rise in health insurance costs. So, I'm curious, do employees in outsourced operations, say in India, have any health coverage? If so, is it employer or government provided? If not, do people just go without? *I'll send you a $10 tip if you can answer!*

8. 586767: Corky Gonzalez [Reference, Education and News]

   I need as much information on Rodolfo 'Corky' Gonzalez, especially about his young life. Lots of biographical information would help a lot. *I tip $10 for answers.*

9. 587167: Film Preservation [Arts and Entertainment]

   I'm looking for resources on how films have been collected and preserved during the 20th century. I'm especially interested in the problems of decaying celluloid and the transfer to digital format. *I'll tack on an extra $10 when I get an answer.*

10. 587678: Courses for being a psychologist [Reference, Education and News]

    What high school courses should one take if they are interested in being a psychologist in the future? *I'll pass on an extra $10 to the answerer.*

11. 589099: History of dried pasta [Reference, Education and News]

    When did dried pasta first become commercially available to mass markets (outside of Little Italy neighborhoods in urban centers)? Early Italian immigrants set up their own home-based pasta emporia (just as, for example, there are Jews in Brooklyn making matzo using traditional methods vs. the stuff available on supermarket shelves) but I want to know when packaged dried pasta became available to wider markets. Thank you! *Tip of $10 after I get an answer.*

12. 592298: Top Album selling music artists [Arts and Entertainment]

    Where can I get a list of the top album selling music artists of all time? I'd like a list that cuts across genres if possible (not just pop stars, country stars, etc.). I'm just interested in album sales, not concert revenues or anything. *I'll add a $10 tip to anyone who can answer this question for me.*

13. 592882: Epson Ink Fading [Computers]

I printed pictures 3 months ago on an Epson printer with their ink but now I see that they are fading. I thought the new inks are supposed to last a long time? Is it really because I didn't use their paper or did something go wrong? *The researcher who answers this for me will get a $10 tip.*

14. 593401: What Causes Fear? [Science]

    Why do people get scared when they watch scary movies? *Any researcher who can answer this for me earns a $10 tip.*

15. 603334: Pica [Family and Home]

    I think my cat might have Pica. She gnaws on a wood table and a metal lamp, and yesterday I saw her licking the tv screen! I would like to know more about this condition and what, if anything, I can do to stop this behavior. Is it a vitamin/mineral deficiency? *I really want to know the answer, so I'll add a $10 tip when I get a response from a researcher.*

16. 603801: Average College Enrollments [Reference, Education and News]

    I am trying to determine the average enrollment of colleges/universities in the US – if it is greater than 5000. (I am using that statistic as a comparison to a specific small town population.) If the avg college/university has less than 5000 students, I'd like to know what percentage is greater than 5000. (Such as, x% of US colleges/universities have over 5000 students). *I promise a $10 tip to the person who answers this for me.*

17. 604934: Historical Discrimination Against Indians [Reference, Education and News]

    How were the Indians discriminated against? What were some difficulties that they faced? What did they do to overcome discrimination? What did the government do to help the Cherokee Indians? (I am Cherokee Indian myself my grandmother was pure Cherokee Indian.) *I'll tip the answerer $10.*

18. 606895: Update of Ptolemy's Model [Science]

    I would like to know about a fairly accurate Ptolemaic model of the solar system. Even though we know that the Earth is not the center of our Solar System, I have read that Ptolemy's model was tweaked in such a way that it became fairly accurate for years with only minor adjustments to keep it such. I figure it would be an interesting contrast to the traditional solar system if there is some interesting information about it. I would also like to get some historical insight too. *I'm willing to give a $10 tip for an answer.*

Treatment: $30 Fixed Price

1. 542475: Retirement Savings Plans [Business and Money]

   I want to start saving for retirement. I've read an article that mentioned IRAs, Roth IRAs, 40l(k)s, 403(b)s, and 457 government plans, but didn't explain them. I would like to know what they are and see some brief comparisons of the different plans, if possible.

2. 542691: Oldest Religious Texts [Reference, Education and News]

   Hey, what religion has the oldest texts – Judaism or Hinduism?

3. 543122: Indian Summer Origin [Reference, Education and News]

   Hey, why is the warm period after the first frost in New England called Indian Summer?

4. 543515: Women in Afghanistan [Reference, Education and News]

   Hi there! I am interested in learning more about the status of women's rights in Afghanistan. Can you help me to find articles?

5. 584328: History of Child Protection Movement [Reference, Education and News]

   I am looking for a good source of information of the history of the Child Protection Movement in the US. I also need to know the key people who were involved in the movement.

6. 584804: Origins of Redheads [Miscellaneous]

   What are the origins of red-haired people?

7. 586134: Font Size and Reading Speed [Science]

   Does the size of font affect how fast you read? In addition to your answer I would also like to know where else I may be able to find this kind of information.

8. 586476: Women's Rights [Relationships and Society]

   Where can I find information on women's rights in the 1920s?

9. 586766: Communities Acting Against Gang Violence [Reference, Education and News]

   I need to find ways that communities have acted to combat proximate gang activity and violence.

10. 587165: Shakespeare's "fair youth" [Arts and Entertainment]

    I read that the "fair youth" of Shakespeare's sonnets was an Earl. Where can I find out more about this?

11. 587677: Clarinet in classical orchestra [Arts and Entertainment]

    What was the role of the clarinet in a classical orchestra of the early 20th century? What about in jazz music from the same time?

12. 588136: Home counties [Miscellaneous]

    Why are the home counties called the home counties, and where are they located exactly?

13. 588551: Feminist critic of Jonathan Swift [Arts and Entertainment]

    Here is my feminist outlook on Jonathan Swift's GULLIVER'S TRAVELS: Swift portrays women as inferior creatures in GT, comparing them to lusty, dirty, and ignorant animals, ultimately leading to Gulliver's disgust in women in general at the end of the novel. So basically, I am looking for any feminist criticism on GT proving that Swift was a misogynist who oppressed women in his novel (particularly relating to his comparison of women to animals in GT). I already have found Felicity A. Nussbaum's article "Gulliver's Malice: Gender and the Satiric Dance," so I am looking for other sources along the lines of that article. I am looking for scholarly articles, preferably available on the Internet.

14. 589097: George Herbert Walker Bush on war on drugs [Reference, Education and News]

    What aspect of the "War on Drugs" did President George Herbert Walker Bush feel was the most important? Is there something that makes his drug policy stand out from Reagan's?

15. 592292: LLP Info [Business and Money]

    I would like to know where I can find the LLP for the Virgin Islands. I am doing a Business Plan for a Nightclub.

16. 593399: Stronger Bones [Science]

    I am learning about the human anatomy and physiology and I have two questions. Why do extended periods of inactivity cause the skeleton bones to degenerate? Why do athletes have stronger bones?

17. 602778: Native American Religions [Reference, Education and News]

    I need information on Native American Religions. I know that they tend to revolve around stories, and I've found several Native American myths, but they aren't very informative as to the -for lack of a better word- structure of the religion. Another thing I have noticed is that the myths, (and from that I conclude any other religious beliefs as well) cannot be said to be consistent between different Native American tribes. Most likely, information that would be useful to me would focus on one particular tribe, and be more in depth, as opposed to less detailed information on many tribes.

18. 603331: Gas Prices [Business and Money]

    What has made the price of gas go up so much this year?

19. 603799: Rare Books [Reference, Education and News]

    What characterizes a "rare book" I am doing a workshop for librarians to help them identify books in their collections..donated, gifted etc. that have potential value as collector's items What to look for?... first editions, special binding etc.

20. 604165: Historical Children's Correspondence [Reference, Education and News]

    I would like children's correspondence, either fictional or non-fictional in nature, from previous centuries, just two or three would do, so that my students can compare them with twentieth century letters and twenty-first century MSN correspondence. Thank You.

21. 604463: Artificial Mulch [Family and Home]

    Researching artificial rubber mulch and if it is safe to use with plants, shrubs? I am getting conflicting answers. Just want to see if you get more answers than I can. Thank you. Where to research further? Take care!!

22. 604931: Anti-Pornography Studies [Reference, Education and News]

    I am looking for any studies used by the government to support anti-pornography legislation.

23. 607164: Differences Between Aristotle, Plato, and Socrates [Reference, Education and News]

    Please explain the differences between Aristotle, Plato, and Socrates, mainly on the issues of government and education.

Treatment: $20 Price with a Conditional $10 Tip

Please note that the tag line for tips (in italics) were added by our research assistants. They were not part of the original IPL questions.

1. 542478: Paper Inventor [Reference, Education and News]

   I've heard conflicting stories about who invented paper. Was it the Chinese or the Egyptians? *Just to let you know, if I get an answer I like, ten more dollars will be on the way.*

2. 542689: Help With Our Public Library [Reference, Education and News]

   Hi. Can you help me find information regarding public libraries? number of volumes, population served, budget allowances, etc.? I work for a library in New York and we are attempting to compare our library with other public libraries in the state or country. *If you do well with this, you will get another ten dollars as a tip.*

3. 543120: Ancient Egyptian Crops [Reference, Education and News]

   Hi! What crops were grown in the time of the reign of Tutankhamen? Were they exported? What equipment was used? *$10 tip for a strong answer!*

4. 543512: Magic for Kids [Family and Home]

   My son is very interested in magic. He is 10. Are there sites online for kids that will talk about how to do magic tricks? Books would be ok too. *I tip $10 for good answers.*

5. 586143: Closed Oil Wells [Science]

   After an oil well has completely run dry, do they just cap the well or do they fill the hole? If they fill, what material is used ? *If you do a good job I'll throw in a $10 tip.*

6. 586769 Chinese Laundries [Reference, Education and News]

   Where can I find out about Chinese laundries in the late 1800s and early 1900s? *If you do a good job I'll add $10 in tip.*

7. 587171: Basic Painting Information [Arts and Entertainment]

   I am a single mother who has recently become interested in painting, but I can't afford to take a formal class. I want to start painting on canvas but don't know about the different kinds of paint and what I would need to know to do that. Is there a website or a book that can talk about the different art supplies and what they are used for? Maybe some basic descriptions of painting on canvas? I just want to learn as much as I can. Someday I will pass it on to my kids too, but now I just want to learn more about my hobby. Thank you so much for your help in this matter. *I'm willing to tip $10 if your answer is good.*

8. 587680: About Muslim prayer [Relationships and Society]

   What are the main corners (foundations) of Muslim prayer ? *I'll add a $10 tip if your answer is good.*

9. 588141 Information about Tango [Arts and Entertainment]

   Is it true that the Tango started in bordellos? What do "lunfardo," "guapo" and "compadrito" mean in the context of tango history? What kinds of music influenced tango music? How did the bandoneon become popular? Besides your answer, where can I research about the details of all this? *I'm willing to tack on a $10 tip for an answer that's good.*

10. 588557: Nobel prize winner [Arts and Entertainment]

    Who has won the Nobel prize in English literature in the recent years? Who are the top English poets/writers nowadays? *$10 tip if the answer is good.*

11. 589105: Santa Ana winds [Reference, Education and News]

    What are the Santa Ana Winds and how did they get that name? *A researcher who gives me a good answer will get a $10 tip.*

12. 592884: General Information about Wine [Miscellaneous]

    I would like some general sites or books about wine. I would like to learn about the subject but don't know where to start. I would just like general information, about grapes, varieties, etc. *Do a good job and I'll reward you with a $10 tip.*

13. 602783: Massage Myth? [Science]

    I have heard that supposedly when after you receive a massage you are often told to drink lots of water (or fluids, I would guess) because receiving a massage releases toxins and you need to flush them out of your system. Frankly, this sounds like hippie stuff. Is there any actual hard scientific evidence to support or refute this claim? *If your answer is good, I'll add a $10 tip.*

14. 603803: Wedding Speeches [Relationships and Society]

    Who is supposed to give the first speech at a wedding reception traditionally? Is it the best man? Is there any order for who should go second? *$10 tip for a well-researched answer.*

15. 604468: Training Multiple Dogs [Family and Home]

    I have two 3 yr old labrador mixes. I want to train them, but am having trouble finding information specifically on how one person should go about training two dogs. Everything I have read online and in books seems designed for training just one dog. Am I supposed to separate them for training? *I'll add a $10 tip if your answer is good.*

16. 604935: Does Chicken Soup Work? [Health]

    Does eating chicken soup really lessen the symptoms of the common cold, the flu and other similar maladies? If so, how? If not, why do we think it does? *Researchers: I'll tip $10 if I like your answer.*

17. 606604: Online Well-Known Sermons [Reference, Education and News]

    I would like to find out where there might be online texts of famous sermons, specifically Protestant, preferably Puritan or Calvinist, referring to salvation, predestination, heaven and hell. Any suggestions would be immensely appreciated. Thanks! *I'll throw a $10 tip your way if you give me a good answer.*

APPENDIX C. Training and Rating Session Instructions

Training Session Instructions

You are now taking part in a study that seeks to characterize the quality of answers in digital reference or commercial question answering services. Your participation will take the form of rating questions and their corresponding answers on several factors. You will initially take part in a training session, followed by five rating sessions spaced over the course of the week. We ask you to not communicate with the other raters during the rating sessions, nor to discuss your rating activities outside of this room during the course of this week. Should you have any questions during your sessions please ask us.

The purpose of this training session is to familiarize you with the rating methodology to be employed, and to ensure a common understanding of the factors used. However, this does not mean that you should all give identical ratings. We want to emphasize there is no single correct way to rate any of these question-answer pairs. We are interested in eliciting objective ratings from impartial raters. We ask you to rely on your own judgment when rating.

In this training session you will be asked to rate two question-answer pairs. For each question-answer pair, you will be asked for nine ratings, as shown below:

1. Please rate the difficulty of the **question**. (1 = very easy . . . 5 = very difficult)

2. Please rate the **answer** for the following factors:
   (1=strongly disagree . . . 5 = strongly agree, NA = Not Applicable)

   (a) The question that was asked is answered.

   (b) The answer is thorough, addressing all question parts.

   (c) The sources cited are credible and authoritative.

   (d) The links provided are to relevant web sites or pages.

   (e) Information in the cited sources is summarized.

   (f) Only information pertinent to the question is presented.

   (g) The answer is well-organized and written clearly, avoiding jargon and/or inappropriate language.

3. Please rate the overall quality of the **answer**. (1=very low quality ... 5=very high quality)

Are there any questions?
The procedure we will follow in the training session is as follows:

1. You will each receive two rating sheets and your rater ID.

2. You will rate the first question-answer pair, working individually. Please write your rater ID on the rating sheet for Question 1, then open a browser window and go to the following web page:
   `http://www-personal.si.umich.edu/~kimym/training/Q1.html`
   Enter your ratings on the rating sheet.

3. Please let us know when you have finished rating the first question-answer pair. We will wait until all the raters have completed rating. Do not proceed to the second pair.

4. When all raters have completed rating the first question-answer pair, there will be a brief discussion, no longer than 15 minutes, regarding the rating activity. We will go over each rating, asking all of you for your ratings. We will also present our ratings and why we rated them so. You may ask us questions at any time.

5. When all questions have been addressed, we will move on to the second question answer-pair, and repeat the procedure used for the first pair. The second question-answer pair is on the following web page:
   `http://www-personal.si.umich.edu/~kimym/training/Q2.html`

6. Please return your completed rating sheets to us.

Are there any questions? Before we start, we would like to ask you to please take the time to read each question and answer pair carefully when rating. We have found that it takes between 7 and 10 minutes to rate each question when evaluating them carefully. If there are no further questions, let's begin.

Rating Session Instructions

The actual ratings will be done using a web-based system. The system has been programmed to show you 20 question-answer pairs for rating per login session. Once you have rated twenty pairs you will be automatically logged out. If you have to quit your session before answering all twenty, simply close the browser window.

Instructions for rating on the web:

To start your rating session, please go to the following web page:
`http://www-personal.si.umich.edu/~kimym/login.php`

Now follow these steps:

1. Login using the login and password given to you by the coordinator.

2. Provide the nine ratings requested for the question-answer pair.

3. When you are done, click 'Submit' - note that you will not be permitted to continue until you have entered all nine ratings.

4. The next question-answer pair will be presented.

5. When you have finished rating the session limit, you will be shown a 'Goodbye' screen.

6. Close the browser.

References

Adamic, Lada A., Jun Zhang, Eytan Bakshy, and Mark S. Ackerman, "Knowledge Sharing and Yahoo Answers: Everyone Knows Something," in "WWW 2008" 2008.

Biddle, Jeff E. and Daniel S. Hamermesh, "Beauty, Productivity, and Discrimination: Lawyers' Looks and Lucre," *Journal of Labor Economics*, 1998, *16* (1), 172–201.

Edelman, Benjamin, "Earnings and Ratings at Google Answers," 2004. Manuscript.

Gazan, Rich, "Specialists and synthesists in a question answering community," in "Proceedings of the 69th Annual Meeting of the American Society for Information Science and Technology" 2006.

Harper, F. Maxwell, Daphne Raban, Sheizaf Rafaeli, and Joseph A. Konstan, "Predictors of answer quality in online Q&A sites," in "CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems" ACM New York, NY, USA 2008, pp. 865–874.

Jr., Joseph Strayhorn, John F. McDermott, and Peter Tanguay, "An Intervention to Improve the Reliability of Manuscript Reviews for the *Journal of the American Academy of Child and Adolescent Psychiatry*," *The American Journal of Psychiatry*, 1993, *150* (6), 947–952.

Landry, Craig E., Andreas Lange, John A. List, Michael K. Price, and Nicholas G. Rupp, "Toward an Understanding of the Economics of Charity: Evidence from a Field Experiment," *Quarterly Journal of Economics*, May 2006, *121* (2), 747 – 782.

McCrea, Richard T., "Evaluation of two library-based and one expert reference service on the Web," *Library Review*, 2004, *53* (1), 11–16.

Nam, Kevin K., Mark S. Ackerman, and Lada A. Adamic, "Questions in, Knowledge iN? A Study of Naver's Question Answering Community," 2008.

Raban, Daphne R. and F. Maxwell Harper, "Motivations for Answering Questions Online," in "New Media and Innovative Technologies" 2008.

Rafaeli, Sheizaf, Daphne R. Raban, and Gilad Ravid, "Social and Economic Incentives in Google Answers," in "ACM Group 2005 Conference" ACM November 2005.

Regner, Tobias, "Why Voluntary Contributions? Google Answers!," Technical Report Working

Paper No. 05/115, Centre for Market and Public Organisation, University of Bristol January 2005.

Resnick, Paul, Richard Zeckhauser, John Swanson, and Kate Lockwood, "The Value of Reputation on eBay: A Controlled Experiment," *Experimental Economics*, June 2006, *9* (2), 79 – 101.

Roush, Wade, "What's the Best Q&A Site?," *MIT Technology Review*, December 2006.

Shah, Chirag, Jung Sun Oh, and Sanghee Oh, "Exploring characteristics and effects of user participation in online Q&A sites," *First Monday*, September 2008, *13* (9).

Tinsley, Howard E. A. and David J. Weiss, "Interrater Reliability and Agreement," in Howard E. A. Tinsley and Steven D. Brown, eds., *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, San Diego, CA: Academic Press, 2000.

van Rooyen, Susan, Nick Black, and Fiona Godlee, "Development of the Review Quality Instrument (RQI) for Assessing Peer Reviews of Manuscripts," *Journal of Clinical Epidemiology*, 1999, *52* (7), 625–629.

Wood, Michael, Martyn Roberts, and Barbara Howell, "The Reliability of Peer Reviews of Papers on Information Systems," *Journal of Information Science*, 2004, *30* (2), 2–11.

Yang, Jiang, Lada A. Adamic, and Mark S. Ackerman, "Competing to Share Expertise: the Taskcn Knowledge Sharing Community," in "ICWSM 2008" 2008.

Zhang, Jingjing, "A Laboratory Study of Communication in Asymmetric Group Contest over Public Goods," 2008.