Predictors of Answer Quality in Online Q&A Sites

F. Maxwell Harper[*], Daphne Raban[**], Sheizaf Rafaeli[**], Joseph A. Konstan[*]

[*] GroupLens Research
Department of Computer Science and Engineering
University of Minnesota
Minneapolis, Minnesota, USA
{harper, konstan}@cs.umn.edu

[**] Center for the Study of the Information Society
Graduate School of Management
University of Haifa
Mount Carmel, Haifa, Israel
draban@univ.haifa.ac.il; sheizaf@rafaeli.net

ABSTRACT

Question and answer (Q&A) sites such as Yahoo! Answers are places where users ask questions and others answer them. In this paper, we investigate predictors of answer quality through a comparative, controlled field study of responses provided across several online Q&A sites. Along with several quantitative results concerning the effects of factors such as question topic and rhetorical strategy, we present two high-level messages. First, you get what you pay for in Q&A sites. Answer quality was typically higher in Google Answers (a fee-based site) than in the free sites we studied, and paying more money for an answer led to better outcomes. Second, we find that a Q&A site's community of users contributes to its success. Yahoo! Answers, a Q&A site where anybody can answer questions, outperformed sites that depend on specific individuals to answer questions, such as library reference services.

Author Keywords

Q&A, online community, digital reference, information quality, expert services, knowledge networks, information exchanges.

ACM Classification Keywords

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

INTRODUCTION

User contributions continue to generate increasing amounts of rich online content. One relatively recent manifestation of this trend is question and answer (Q&A) sites – places where users ask questions and others answer them.

South Korean Internet portal Naver (http://naver.com), illustrates the potential of Q&A sites. As of July, 2007, Naver handled 77% of internet searches originating in South Korea, dwarfing worldwide leaders Yahoo! (4.4%) and Google (1.7%) [19]. One of the reasons for this disparity is the relatively small Korean language corpus available for crawling. To address this shortcoming, Naver built a Q&A site called Knowledge iN that encourages users to type questions for others to answer, rather than relying on search results [2]. Since their 2002 launch of Knowledge iN, Naver has accumulated 70 million questions and answers, and continues to receive over 40,000 questions and 110,000 answers per day [19].

Similar sites are now common worldwide. Yahoo! now offers Q&A sites localized to 26 countries.[1] As of December, 2007, Yahoo! Answers has attracted 120 million users worldwide, and has 400 million answers to questions [11]. Yahoo! incorporates their Q&A data into their search results. Although Google closed their U.S. Q&A site in 2006, they rejoined the trend with new services in Russia and China in 2007.

Given the increasing competition for users' questions and answers, different designs have emerged. Some sites allow anyone in the community to answer questions, while others have individual "experts" filling that role; some charge askers and pay answerers, while others use leaderboards, points, or stars to encourage answering. Design decisions such as these are likely to have a large impact on the type and volume of questions asked, as well as the quality and responsiveness of the answers (and corresponding value to "social search" system providers) [15], yet we still know little about the specific effects of these decisions.

From a user's perspective, it may be unclear which sites to turn to for high quality, friendly, or responsive answers. For users who have not yet joined a Q&A community, it is not obvious which questions to ask rather than search for. In Q&A sites that require payment, it is not obvious whether spending more money will earn a better answer. And it is possible that different types of questions and different rhetorical strategies will receive different responses, depending on the community. In general, it is unclear what approach to selecting and using a site will yield the most useful results.

CHI 2008, April 5–10, 2008, Florence, Italy.

[1] According to http://answers.yahoo.com, as of 1/7/2008

In this paper we seek to answer some of these questions through a comparative, controlled field study of responses provided across several online Q&A sites. We report both on quantitative data regarding site performance, and qualitative observations which illustrate several interesting and nuanced characteristics of these sites.

QUESTION AND ANSWER SITES

For the purposes of this study, a question and answer (Q&A) Web site is purposefully designed to allow people to ask and respond to questions on a broad range of topics. Fundamentally, all Q&A sites offer some interface designed for *asking* and *answering* questions. Commonly, users will be asked to categorize their question in some way, to route the question to answerers. Also, most Q&A sites offer an interface for *searching* and *browsing*, often organized by question status (e.g. "open", or "closed"). Yahoo! Answers presents an example of such an interface (see Figure 1).

Three Types of Q&A Sites

From the research literature and our observations and use, we identify three types of Q&A sites: "digital reference services", "ask an expert services", and "community Q&A sites." All three of these site types are in common use today. While they are all built to help users ask and answer questions, they achieve this goal in very different ways:

Digital reference services represent the online analogue to library reference services [13]. Traditional library reference services employ expert researchers[2] to help people find useful information. Today, many public libraries have added digital reference services, such as the New York Public Library's "Ask Librarians Online" (http://www.nypl.org/questions/). Digital reference services typically use basic tools for online communication: in one survey, 71.4% elicited questions via a web form, and 80% respond to questions via email [20]. Digital reference services rely on specific people performing specific tasks (such as question routing, researching, and answering) as well as highly-constructed workflow (e.g., by using issue tracking software for tracking open/closed questions). Thus, this type of reference service maintains the library's organized and structured model of question answering; some researchers have argued that it is these "clearly defined policies and procedures that are well understood by all the participants" that are a model for success [12].

Ask an expert services[3] represent a first step, technologically and socially, away from the structure and formality of digital reference, while retaining the overall goal of providing quality question answering service. These services are staffed by "experts" (of varying credentials), often in a relatively circumscribed topic area, such as science (e.g. at "MadSci Network", http://www.madsci.org) or oceanography (e.g. at "Ask Jake, the SeaDog", http://www.whaletimes.org). Ask an expert sites tend to have some organizational and procedural structure, though less so than digital reference services. For example, in some sites the category of the question asked may determine which expert will respond; other systems allow experts to declare which questions they will answer by locking the question, or by other means. Ask an expert services may be thought of as online communities, though member interactions tend to be very topic-oriented, "discussions" tend to read like FAQs, and askers and answerers do not interact as peers.

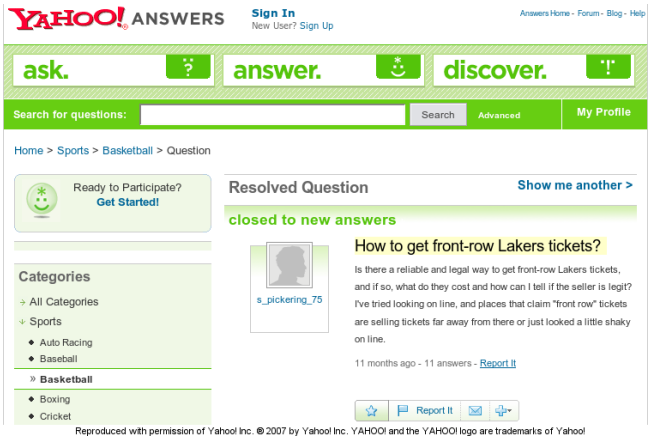

Figure 1. A question in Yahoo! Answers.

Community Q&A sites[4] leverage the time and effort of everyday users to answer questions – they represent Web 2.0's answer to more traditional online reference services. Established examples of community Q&A sites are Yahoo! Answers and Knowledge iN. Community Q&A sites have little structural or role-based organization, typically, although some sites have moderators, or users who have earned elevated privileges based on past contributions. These sites manifest strong online community features, including off-topic discussions and discussions where participants reply to one another, and the presence of repeat/regular users. Community Q&A sites also tend to embrace newer interaction designs than the other types of Q&A sites, by providing features like tagging and ratings interfaces, RSS feeds, and highly interactive browsing and searching capabilities.

Related Work on Q&A Sites

Q&A sites have been the subject of related work from researchers in such diverse communities as information science, economics, and information retrieval.

[2] In the U.S., reference services are typically staffed by librarians with a master's degree in library science.

[3] Also known as: expert services, knowledge networks, or information exchanges.

[4] Also known as: social Q&A or knowledge search; not to be confused with Question Answering (QA), a research area in natural language processing.

Recent work has examined question answering roles in the context of the community Q&A site AnswerBag (http://www.answerbag.com) [5]. This work distinguishes between two roles: *specialists* and *synthesists*. While specialists claim expertise in a given topic and answer questions based on their own knowledge, synthesists gather pointers to outside resources. The researchers found that answers from synthesists were slightly more useful than answers from specialists. We revisit this result by asking whether the number of links in an answer is correlated with that answer's quality.

Google Answers (http://answers.google.com) has attracted quite a bit of research attention, primarily because it created a monetized "information market", where users choose how much to pay for a question, and optionally tip answerers for work well done. Edelman found that answerers provide better answers as they gain experience, and that topic-specialists provide higher quality answers than generalists [4]. Rafaeli et al. found in observational study that levels of payment alone were not sufficient to explain variations in answer quality, but that non-monetary incentives such as ratings influenced answer quality [16, 17]. We extend this result by looking at additional outcomes, and through our use of a controlled field study instead of observation.

Other researchers have investigated differences among Q&A sites. A cross-site comparison of ask an expert services was conducted across 20 such sites [7]. In this study, the researchers injected questions into the sites and measured outcomes such as response rate, response time, and verifiable answers. This research led to modest findings: ask an expert services in general responded to 70% of all questions, and commercial sites were more likely to provide one or more answers than noncommercial sites. Perhaps more importantly, this study presents a useful methodology for studying these sites that includes "developing" questions by revising existing questions from the Internet Public Library (based on a small set of criteria) with verifiable answers. This methodology is less useful for studying community Q&A sites, as answers from one site quickly show up in search results, cross-contaminating conditions; our methodology addresses this problem.

More recently, Rousch conducted an informal comparison of six popular community Q&A sites by searching for and asking a small set of questions at each site [18]. Based on this informal study, the author concluded that Yahoo! Answers was the best performer, and ventured that the reason was their large base of users. This speculation led us to study the influence of a community of users in a larger and more formal study.

As Q&A technologies continue to evolve, qualitative observation has the potential to greatly increase our understanding of how questions are asked and answered online. One interesting example of qualitative observations is presented in [9], where the researchers (who are studying music information retrieval) present data on attributes of the questions asked in Knowledge iN and Google Answers. The researchers describe the questions asked on these sites as frequently "vague, incorrect, and incomplete". They also provide a breakdown of questions by the type of "information need", finding that in the domain of pop music queries, a large majority of users wish to identify a particular song or artist, or receive music recommendations. In this research, we report on qualitative observations to better understand the characteristics that emerge in different types of Q&A sites.

RESEARCH QUESTIONS

Both site designers and question askers would benefit from a better understanding of the predictors of Q&A quality, responsiveness, and effort. In this context, we present two main research questions that summarize the high level goals of this research.

Research Question 1: How do Q&A sites differ in the quality and characteristics of answers to questions?

We hypothesize that Q&A sites differ in how well they answer questions, and that there are a number of dimensions along which sites differ that influence the quality and characteristics of answers. Specifically, we investigate the following themes:

- How do different Q&A sites differ in answer quality, answerer effort, and responsiveness?
- How do community Q&A sites compare with sites that rely on individuals for answers?
- Do for-fee Q&A sites outperform free sites?
- How do Q&A sites with topic experts compare with those with research experts?

Research Question 2: What can question askers do to receive better answers from a Q&A site?

A first time visitor to a Q&A site might try any number of strategies to get useful answers to a question. We investigate some fundamental strategies to question asking:

- Does paying more in a for-fee Q&A site result in better answers?
- Do simple rhetorical strategies such as thanking the answerer or indicating prior effort affect answer quality?
- Does the topic of the question or the type of question affect answer quality?

METHODS

To determine how Q&A sites differ, and to determine strategies for question askers to receive better responses, we conducted a six week field study, using five sources of online answers. In this study, we asked real questions using made-up identities, then used a panel of blinded judges to rate the questions as well as the answers. In this section, we describe the sites and methodology used in this research.

Q&A Sites Used in This Research

We selected the following Q&A sites for our study:

Library Reference Services represent a traditional form of digital reference service. We consider these sites as a baseline for question answer quality and responsiveness. We divided our questions among eight brick-and-mortar libraries from the United States and Canada[5] that offer Web-based digital reference, plus the Internet Public Library's "Ask a Question". These services all operate using the same Q&A model and site interface – questions are submitted via a web form, responses come via email.

Google Answers is a hybrid service that combines a digital reference service with some community features. Google Answers employed approximately 500 paid researchers, who would answer questions and ask for question clarifications when necessary. To ask a question, users would declare how much they would pay for an answer, between $2 and $200. Researchers would earn 75% of the price of each question they answered. Optionally, users could "tip" researchers after they returned an answer. Both researchers and regular users could comment on questions; these comments often contained valuable information. Because Google Answers was our only paid site, we studied it at three difference price-points – $3, $10, and $30.[6] Google Answers closed for undisclosed reasons after our study was complete, on December 1, 2006.

AllExperts was among the largest and broadest ask an expert sites operating at the time of our study. To ask a question, users must first locate an appropriate "expert" by navigating a taxonomy of question categories such as "Style→Fashion→Hairstyling". If there are multiple experts for a given category of questions, users can look at experts' personal profiles and the ratings their past answers have received from other users. Questions are not publicly viewable until the "expert" has responded, and optionally stay private. After an answer has been provided, other users of the site may add comments. This design of this site is strongly oriented towards asking questions; it is a challenge to browse and comment on previously asked questions.

Yahoo! Answers is the most visited community Q&A site in the United States. As of December, 2007, Yahoo! Answers has provided over 400 million answers to questions [11]. It also represents what is quickly becoming the de facto standard community Q&A interface: questions are categorized and broadcast to the community, any user can answer any question, and users can rate questions and vote on "best answers". The design of this site encourages browsing questions by category, and emphasizes the newest content. Due to the heavy traffic of this site, questions often receive many replies very quickly after the question is asked. However, questions more than a few hours old often stop receiving further answers as they are lost in the flood of new information.

Live QnA is Microsoft's community Q&A site. While this site features a very similar interface to Yahoo! Answers, it does not have the same level of usage. Microsoft QnA launched on August 1, 2006, about two months before we began asking questions (Oct 10, 2006). We chose this site because it serves as an interesting contrast with Yahoo! Answers, and because it is interesting to study emerging communities. Today, Microsoft QnA has almost 100x less traffic than Yahoo! Answers.

Methodology Overview

We employed a comparative, controlled field study of responses provided across several online Q&A sites. Figure 2 summarizes our methodology for each of the 126 questions that we developed and asked – which includes a process for vetting questions according to our independent variables (described below), as well as a process for evaluating the "output" from the Q&A site. In the following sections we describe these steps in more detail.

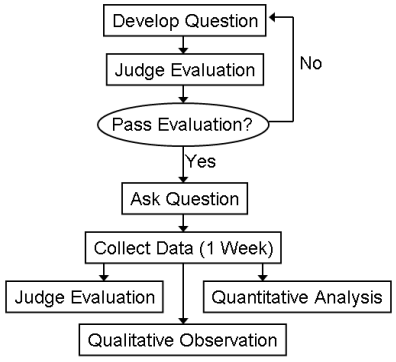

Figure 2. Methods: from question development to analysis.

Experimental Design

At the heart of our study is a set of questions that we developed and asked in Q&A sites. As mentioned earlier, there is a precedent in the Q&A research literature for the custom development of questions to enhance experimental control [7]. Given full control over the questions asked, we designed an experiment with several controlled and balanced independent variables, which we describe below. We chose these variables because they represent some of the high-level characteristics of questions that are

[5] Public libraries located in: Alberta, Kansas City, Memphis, Minneapolis, New York, Saskatchewan, Seattle, and Florida

[6] To determine monetary values to use, we observed one week of data (the week of August 20, 2006) to determine the price distribution of questions asked. During this time period, question prices ranged between the minimum and maximum payments possible ($2-$200); the median and mode prices were both $10. Using this distribution as a guide, we chose the 20th percentile ($3) as a "low" payment, the 50th percentile ($10) as a "medium" payment, and the 80th percentile ($30) as a "high" payment.

commonly asked, and because they plausibly would have an effect on answer quality or responsiveness.

Topic represents the information domain of the question. We examined three topics commonly seen in Q&A sites: technology, business, and entertainment.

Type represents the style of answer the question requires. This variable has three categories: factual, opinion, and personal advice. Factual questions seek objective data or pointers to content; these questions are geared towards researchers. Opinion questions seek others' thoughts on a topic of general interest; these questions do not have a "correct" answer and may be answered without reference to the question askers needs. Personal advice questions seek recommendations based on the asker's own situation; answerers must understand the question asker's situation to provide a good answer.

Prior Effort represents whether the question indicates explicitly that the asker had spent time trying to find the answer before turning to the Q&A site. This variable has two levels: prior effort and no prior effort. A question might indicate prior effort with statements such as "I did a Google search to figure out how to do this.." or "I've asked our office IT folks, and...".

Gratitude represents whether and how the question asker thanks the prospective answerer(s), and may be considered one component of a question's politeness. This variable has three levels: no thank you message, a short (3 words or less) thank you message, and a long (5 words or more) thank you message. We always placed the thank you statement (if any) as the last sentence of the question.

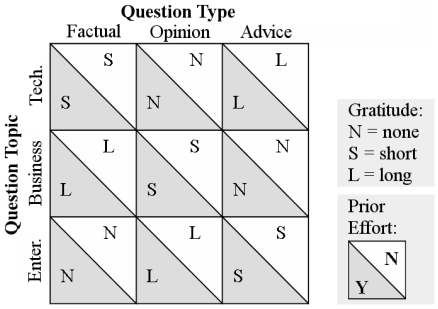

Figure 3. Our experimental design in a nutshell. We developed seven questions for each triangle.

In addition to these four independent variables, we distributed questions across different Q&A systems, as described below.

Destination collapses together Q&A site and cost. We make this combination because cost is only applicable to one of our chosen Q&A systems: Google Answers. This variable further collapses the nine library reference sites into a single conceptual destination to avoid overburdening any particular library, to avoid detection (these sites are very low volume), and to avoid the risk of using an outlier service. Our seven destinations are: Google Answers ($3), Google Answers ($10), Google Answers ($30), Library Reference, AllExperts, Yahoo! Answers, and Live QnA.

To trim the number of experimental groups necessary while preserving main effects and low order interactions, we employed a fractional factorial experimental design [1] to choose 18 out of 54 runs of a full factorial design (see Figure 3). Our four factors were: topic, type, prior effort, and gratitude. We distributed these 18 runs in a balanced fashion across the seven levels of the destination variable. Thus, for each combination of topic, type, prior effort, and gratitude, we developed seven questions. We randomly assigned each of these questions to one destination.

Developing Questions

Several criteria governed our process of question development. First, we required realistic questions that might be plausibly asked online. Second, we required that our questions would need research or expertise for high quality answers – for each question, we tried one or more intuitive Google searches to ensure there were no solutions (or existing answers on Q&A sites) in the immediate search results. Third, we required questions that adhered to the constraints of our experimental design.

To impartially evaluate questions (and later, answers), we employed a panel of six judges. These judges were all juniors and seniors in college, majoring in either English or Rhetoric. We selected these students because we believed their training in the use of language would enable them to understand and evaluate online questions and answers. The judges were each paid for their participation. They were blinded to the context of the question – they evaluated the unformatted text of the question and answer(s) – and they were not aware that we were the authors of the questions.

We wrote 18 question *templates*, representing the relevant combinations of the four independent variables in our fractional factorial design (as described above). From each of these templates, we wrote seven questions, one for each destination. The seven variants all contained the same number of sentences, and shared characteristics as governed by the experimental design. But they differed in the object of the question enough to avoid detection, and to avoid cross-contaminating other sites with search engine results. For example, two variants of the question defined by entertainment (topic), factual (type), no prior effort, and no thank you message were:

"Are there videos available of the Tonight show from the Steve Allen and Jack Paar years? For sale, online or even just in a museum, I'd like to find them."

"Where could I find unusual Gilligan's Island Memorabilia for a big fan's birthday? Not the usual videos and pictures, but preferably something like props from the original set."

To ensure that our questions reflected our desired independent variables, the judges evaluated each question across a series of criteria. For example, the judges were

asked to classify each question as "business", "technology", or "entertainment", and to rate each question in its difficulty. We discarded or rewrote any question that did not get a majority of judge agreement.

Pilot Study and Judge Training

About a month before the launch of our experiment, we conducted a pilot study to train our judges on non-essential data and to refine our experimental procedure. We developed seven factual questions and asked one in each destination. We collected responses for a week, then brought the six judges to a training session where we set expectations and answered questions.

Asking Questions

Beginning Oct 10, 2006, we spent six weeks asking our questions online. To avoid the potentially confounding effects of time of day or day of week, we asked all of our questions at approximately 1:00pm local time, on Tuesday, Wednesday, and Thursday of each week. We asked seven questions per day (one per destination) – twenty-one questions per week. For each question, we created a pseudonym using a random name generator, which we methodically turned into a Q&A account sign-in name and email address from a free web-based email provider.

Outcome Measures

Once we had asked a question, we recorded all responses that arrived for up to one week. Recall that all sites allow for multiple answers (or no answers at all). In our analysis, we analyze answers *in aggregate* unless otherwise noted (e.g. in reporting per answer metrics) – in this research study we are primarily interested in Q&A site usefulness from the perspective of the question asker, rather than in characteristics and variations between individual answerers. As such, our judges were trained to evaluate answers as a set, rather than individually, looking at the overall quality and feel of the answer(s) given. In our analysis of Google Answers, we consider both "comments" and formal answers as part of the answer set, as comments often contained valuable information to the asker.

We consider several primary outcome measures in our analysis. First, we report on several simple metrics about answers, such as the number of answers received to a question, the length of these answers (in characters), and the number of links in these answers. Second, we report on two index variables constructed from the judges' evaluations of answers. To construct each of these variables, we normalized and summed a set of Likert scale survey questions, then renormalized to a 0-1 scale. The first of these variables, *judged answer quality*, is constructed from five Likert scale survey questions that we think reflect the overall goodness or value of the answer(s) provided. These five survey questions measure: (1) answer correctness, (2) asker confidence in answer, (3) helpfulness of answer, (4) progress towards receiving an answer, and (5) monetary worth of the answer. The second of these variables, *judged answerer effort*, is constructed from four Likert scale survey questions that we think reflect the (perceivable) time and energy answerers spent answering the question. These survey questions measure: (1) degree of personalization in answer, (2) perceived answerer effort, (3) answerer friendliness, and (4) ease of use of answer. Both measures appear to be internally consistent (answer quality: Cronbach's $\alpha = 0.94$; answerer effort: Cronbach's $\alpha = 0.87$).

Analysis Methods

In this analysis we focus primarily on main effects. Except where noted, we use regression analysis to build predictive models of four main dependent measures: rated answer quality, rated answerer effort, number of answers, and answer length. We used a mixed model to measure the significance of the judged quality and effort metrics. As each question/answer was rated across many attributes by several different judges, we cannot assume that each judge-score is an independent observation. Thus, to control for judges' biases, we analyzed the data treating judge as a random effect. In our analysis we used a Restricted Maximum Likelihood method for fitting mixed models, as this method does not depend on balanced data.

QUANTITATIVE RESULTS

Table 1 summarizes several outcome measures. Across all destinations, 84.1% (106/126) of our questions received at least one answer. We received a total of 276 answers. On average, each question received 1.73 answers in the first day and 2.19 in the first week. The average judged answer quality was 0.48, and the average judged answerer effort was 0.51. Notably, the top 17 answers in terms of judged quality came from Google Answers, 8 of which were $30 questions, and 7 of which were $10 questions. The top overall answer scored 0.88 on our judged quality scale. As ranked by judged quality, the top AllExperts answer ranked #18, the top Yahoo! Answers answer ranked #21, the top library reference answer ranked #26, and the top Live QnA answer ranked #38.

In general, answers with high quality ratings were long and contained many links. We are able to explain 32% of the variance in judged answer quality using just two simple variables: the total length of the answers received, and the total number of hyperlinks in the answer. Our method here is to build a regression model predicting judged answer quality, controlling for the judge as a random effect, using answer length and number of links as our independent measures. We find that both length and number of links are statistically significant (answer length: $p<0.01$; number of links: $p<0.01$; total model: $R^2=0.32$) and positively correlated with judged quality.

Because we randomly assigned questions to destinations, we expect no difference in question difficulty across destinations. Indeed, as gauged by our judges, there was no significant difference between the seven destinations in predicting question difficulty ($p=0.76$). In addition, there is

Destination	% Ans.	# Ans.	Length	Quality	Effort
AllExperts	61%	0.61	629.45	0.33	0.37
Google A. ($3)	78%	2.39	571.47	0.41	0.44
Google A. ($10)	89%	2.78	815.60	0.59	0.60
Google A. ($30)	**100%**	2.83	**1393.92**	**0.68**	**0.71**
Library Ref.	78%	0.83	802.13	0.41	0.47
Live QnA	89%	1.89	257.76	0.40	0.43
Yahoo! A.	94%	**4.00**	319.24	0.51	0.53
Overall	84%	2.19	678.07	0.48	0.51

Table 1. Comparing destinations: % questions receiving >0 responses, avg. # answers/question, avg. answer length, avg. judged answer quality, and avg. judged answerer effort.

no evidence that our questions were perceived as out of the ordinary in any site: none of our 126 questions received comments or replies indicating that they were out of place.

Research Question 1: How do Q&A sites differ in the quality and characteristics of answers to questions?

Destination Characteristics

As a starting point in our cross-site comparison, we ask which destinations perform the best across a variety of metrics. See Table 1 for an overview of how the different destinations compare across several metrics. We find that that Google Answers at the $10 and $30 level outperformed all other destinations across most metrics, while Yahoo! Answers provided the most answers per question.. There is statistical evidence that the choice of Q&A site has strong effects on outcomes. According to our model, destination is a significant predictor of answer quality ($p<0.01$), answerer effort ($p<0.01$), total answer length ($p<0.01$), and number of answers ($p<0.01$).

Pairwise Wilcoxon tests, pairing questions from the same templates, allows assessment of the statistical significance of the observed differences between destinations.. We find that Google Answers at $30 provides significantly higher answer quality than Google Answers at $3 ($p<0.01$). There is marginal evidence that Yahoo! Answers has higher answer quality than library reference services ($p=0.08$), but no conclusive evidence that Yahoo! Answers outperforms Google Answers at the $3 level ($p=0.16$).

Yahoo! Answers outperforms Live QnA across all of our measures. Given the similar interfaces in these two sites, we see preliminary evidence that the community of users is an important factor in understanding Q&A outcomes. We return to this observation in Qualitative Observations.

Exploratory Analysis of Design Dimensions

We wish to make some claims regarding the effect of several design dimensions on answer quality in real Q&A sites. However, there are challenges in designing a field experiment that supports such an analysis, as there are many confounding factors inherent in all Q&A sites and it is difficult to isolate the effects of any single design feature. However, there is value in "real world" data, and this exploration may provide evidence to inspire future work in a different experimental setting.

For the following analysis, we rely on a similar statistical analysis as described in methods, but we treat Q&A site as a nested variable, as each site occurs at exactly one level of the dimensions that we test.

Community vs. Individual. We wish to understand whether answers from communities outperform answers from individuals. Of the five sites used in our study, two (Yahoo! Answers and Live QnA) give community answers, and two (AllExperts and Library Reference) give answers from individuals. We omit Google Answers data from this analysis, as it is the only fee-based site (which could dominate other factors), and because we consider this a hybrid system: while individuals "answer" questions, others may comment and add to the answer.

As we might expect, we find that community sites provide more answers than individual sites (means: 2.94 community vs. 0.72 individual; $p<0.01$; $R^2=0.39$). We find trends, but not strong evidence, of other effects. We find that community has marginally significant, positive effects on judged answer quality (means: 0.47 community vs. 0.42 individual; $p=0.09$; $R^2=0.09$), and on the total length of answers received (means: 881.92 community vs. 526.56 individual; $p=0.06$; $R^2=0.17$). However, we find no effect on judged answerer effort ($p=0.60$).

Paid vs. Free. We have already presented data that Google Answers outperforms all other Q&A sites across our measures, except when we paid a small amount ($3). If we consider Google Answers to represent "paid" answers, and the other sites to represent "free" answers, we find – not surprisingly – that paid answers appear to outperform free answers. Paid answers had higher judged answer quality (means 0.61 vs. 0.45) and higher judged answerer effort (means 0.62 vs. 0.50); they received more answers (means 2.67 vs. 1.83 answers) and longer answers (means 2527 vs. 704 characters). However, these differences could simply be a result of Google's system outperforming the other systems on average, irrespective of cost. In the absence of free questions in Google Answers (and paid answers in other sites), we can make no claims about paid vs. free answers. However, we later return to the effect of amount paid in Google Answers, and provide evidence that in this case, paying more improves answers.

Specialists vs. Synthesists. As we discussed earlier, Gazan showed that in the Q&A site AnswerBag, answers from *synthesists* were judged better than answers from *specialists* [5]. We are able to provide some corroborative evidence for Gazan's result, as we find number of links to be positively correlated with judged answer quality ($\rho=0.54$, $p<0.01$).

We can extend this analysis by making the assumption that Google Answers and library reference sites consist of a majority synthesists, given their reputation as employing professional or expert researchers. The other three sites we consider to consist of a majority specialists: AllExperts because they employ experts in particular subject areas, and the other sites because we speculate that users spend the majority of their time browsing and answering questions about favorite topics. There is some evidence that this assumption is plausible: taking number of links per answer to as a strength-of-synthesist metric, we find that Google Answers (3.23) and library reference sites (1.89) score highest, trailed by AllExperts (0.81), Yahoo! Answers (0.52) and Live QnA (0.36). Given this separation, our model supports Gazan's finding and shows a positive effect for synthesist sites in answer quality (means: 0.54 synthesist vs. 0.45 specialist, $p<0.01$).

Research Question 2: What can question askers do to receive better answers from a Q&A site?

Level of Payment

A question asker in Google Answers might wonder how much money to spend on a question to get the best answer. To help this person, we look at statistics that describe average answer characteristics across different payment levels. For this analysis we treat cost as a categorical variable; we are interested in whether cost is a significant factor in predicting outcomes, rather than modeling the relationship between actual price and outcome.

As gauged by our judges, when we paid more for an answer, we received higher quality answers and answerers spent more effort. $30 answers were rated 0.68 on judged answer quality, as compared with 0.62 for $10 answers and 0.47 for $3 answers; cost is a statistically significant factor in our model predicting quality ($p<0.01$). $30 answers were rated 0.71 on judged answerer effort, as compared with 0.63 for $10 answers and 0.49 for $3 answers. Cost is also a statistically significant predictor of effort in our regression analysis ($p<0.01$). Interestingly, paying more did not result in *more* answers, although it resulted in *longer* answers. The data show that higher payment leads to significantly longer answers (means: $3, 1365 characters; $10, 2266 characters; $30, 3949 characters) ($p=0.01$).

These statistical results perhaps belie more nuanced community behavior which we revisit in later sections with qualitative observations.

Rhetorical Strategy

In our methodology, we varied our expression of gratitude and our indication of prior effort. These are two simple rhetorical strategies that we speculated could be employed by question askers to affect Q&A outcomes.

It appears that neither gratitude nor prior effort had a statistically significant effect in predicting answer outcomes. On average, using the longest thank you message led to the highest judged answerer effort (means: 0.57 long vs. 0.54 short vs. 0.55 none), and the absence of a thank you message led to the lowest judged answer quality (means: 0.52 long vs. 0.52 short vs. 0.50 none). In our model, however, neither of these differences were statistically significant in predicting quality ($p=0.56$) or effort ($p=0.31$). On average, using an indication of prior effort actually decreased both judged answer quality (means: 0.53 no prior vs. 0.50 prior) and judged answerer effort (means: 0.56 no prior vs. 0.54 prior). In our model, prior effort was marginally significant in predicting quality ($p=0.07$) but not significant in predicting effort ($p=0.32$).

I.V.	One Interesting Result
Topic*	Entertainment questions received the most replies
Type*	Advice questions received highest quality responses
P. Effort	Indicating p. effort slightly decreased answer quality
Gratitude	No main effect, but interacts with destination
Dest.*	Yahoo! Ans. > Library Ref. and Google Ans. ($3)
Payment*	In Google Ans., paying more improves all outcomes

* Significant predictor of one or more outcome measures.

Table 2. There are many quantitative results in these data; this table shows one interesting finding per independent variable.

We might hypothesize that these variables could interact with the Q&A site. For example, perhaps being polite matters in one community with one set of norms, while it is considered strange in another community. We see some evidence of this: there is a significant interaction effect between gratitude and destination in predicting quality ($p<0.01$) and answerer effort ($p<0.01$). Every destination appears to respond to thank you messages differently. We can speculate that different Q&A sites have different cultures, a latent factor that is interacting with our different messages. Prior effort, on the other hand, does not interact with destination in a statistically significant way ($p=0.28$).

Type and Topic of Question

By varying two independent variables – type and topic – we investigate whether the informational goal of a question affects its resulting answer quality in Q&A sites.

Our data suggest that topic has a potentially large effect on the number of answers received (means: 3.07 ent. vs. 1.90 tech. vs. 1.59 bus.; $p<0.01$), a small and marginally significant effect on answer quality (means: 0.49 tech. vs. 0.48 bus. vs. 0.46 ent.; $p=0.06$), and no effect on answerer effort ($p=0.74$) or the length of answers received ($p=0.94$). In particular, it seems that entertainment-oriented questions received many replies, but those replies were poor in judged quality relative to other topics.

Asking different types of question appears to affect outcomes. Our data show that type has a statistically significant effect on quality (means: 0.55 advice vs. 0.46 opinion vs. 0.42 factual; $p<0.01$), effort (means: 0.57 advice vs. 0.51 opinion vs. 0.45 factual; $p<0.01$), and length (means: 2028 advice vs. 1408 opinion vs. 1020 factual;

p=0.02); type has no effect on the number of answers received (p=0.19). Thus, requests for personal advice appear to receive the most – and the best – attention of any question type in our study.

QUALITATIVE OBSERVATIONS

Through the course of our six week study, we observed interactions that illustrate the strengths and weaknesses of the different Q&A sites. In this section, we share some highlights of our study, to deepen our understanding of the dynamics that govern the Q&A process in different sites.

The data presented above strongly make the case that Google Answers, on average, provided the best answers of any of the sites studied. We believe that the reasons for this success go deeper than the financial incentives. Rather, the community of researchers *and regular users* was passionate about answering questions, and appeared to enjoy the "game" of answering challenging questions. In many cases, researchers and other users used the (unpaid) commenting feature to post lengthy replies to answers that other researchers had written. For example, we asked the following $30 question: "Which actress has the first female line in a talking movie? [...]" Within two hours, we received a long (3,800 character) answer that included information about the actress ("*Eugenie Besserer was the first female to speak in a full length talkie. She played Al Jolsen's mother, Sara Rabinowitz in the film the Jazz Singer*"), statistics about the first line, excerpts from the script, and six links to Web pages with further information. However, one community member disagreed with this answer, replying "*The actress with the first line in a talkie was Sarah Bernhardt in 'Le Duel d'Hamlet' around 1900 [...]*". All this led to a five user, passionate discussion concerning the subtleties of the question. The discussion led to two formal answer clarifications and a congratulatory post: "*Well done everybody! A Great Question has brought a Great Answer and some interesting Comments.*"

However, the pricing structure in Google Answers may lead to some awkwardness for new users. Google Answers researchers appear to have internalized a model of how much a question is "worth", while question askers (especially first-timers) may not understand how much money to offer for a question. One question we asked at the $3 level asked for "advice and pitfalls" concerning hiring a custodial service. In response, we received a single comment from a researcher: "*It'll cost you a lot more than $3 to hire a custodian and it would take a Google Answers Researcher a lot more time than $3 is worth to research this question.*" On the flip side, we appeared to overvalue other questions, such as one $30 question with the subject line "What e-mail system to use for mailing lists?" This question received just two short comments with brief recommendations, rather than any "authoritative" answer that synthesized outside data, or gave an expert opinion with background information. In fact, only 11/18 of our $30 questions received an official answer – the seven unanswered questions spanned all three topics and types. However, all 18 received at least one response in the form of a comment, underscoring the value of the community features that Google added to their reference service.

We found that our questions in community Q&A sites were more likely to get *some* response (92%, vs. 81% in the other sites). However, the benefits of high responsiveness are potentially offset by other, qualitative shortcomings. For example, in one Yahoo! Answers questions with the subject "How to get front-row Lakers tickets?", we received two separate responses that read "*ebay*", one that read "*buy them duh*", one that read "*You could try giving Jack Nicholson a call [...]*", and one that read "*Umm I can help ya... Sleep with Jack Nicholson*." On the other hand, we also received recommendations for two ticket resellers that both appear to be good options. Live QnA had a much less responsive feel, but this did not improve our perception of the signal-to-noise ratio. For example, we asked a question about reel film projectors, stating "All my searches for movie projectors point me to digital projectors [...], not a reel film projector", our only response was a link to a site that sells only digital projectors, probably indicating that the answerer did not read the full question.

Although library reference services and AllExperts are fundamentally different types of Q&A services, they shared several of the same advantages and drawbacks, perhaps because they both depend on individuals for answers. In both sites, the biggest problem was getting any answer – in AllExperts, only 61% of our questions received a response, while across libraries, only 78% of our questions received a response.[7] In both services, we typically received exactly one response per question, so there were no opportunities for collecting diverse opinions without re-asking the question elsewhere. In one respect, however, the two services differed – AllExperts responses reflected the answerer's interest in the topic, while the librarians' answers reflected an interest in the research process, or a lack of interest. For example, we asked AllExperts for "information on who might be the best baseball announcer of all-time", and the respondent enthusiastically wrote a lengthy response, stating "*The guy who I most enjoy listening to because he does all these things extremely well is Jon Miller [...] Listening to him makes the game much more enjoyable to me*". In contrast, a question directed at library reference services about "who is the most skilled celebrity chef?" received the dry response: "*I do not have any reliable source for this information. I did find a Website with award information [...]*"

DISCUSSION AND CONCLUSION

In this study, we found that (1) you get what you pay for in Q&A sites, and (2) a Q&A site's community of users

[7] Two of the nine libraries we chose did not reply to any questions; all other libraries had a 100% response rate.

contributes to its success. Across our answer quality and responsiveness metrics, Google Answers (a fee-based Q&A site) was superior to each of the free Q&A sites we studied. Further, when we paid more for an answer at Google Answers, we typically received longer, better answers. Qualitatively, we found that the volunteer efforts of the Google Answers community helped make answers better, and gave the site more diverse opinions. Among the free sites, Yahoo! Answers scored the best – its large community provided high answer diversity and responsiveness. Compared to Live QnA, a community with a very similar design but many fewer users, Yahoo! Answers typically yielded better responses, further underscoring the importance of a large, active community.

There is an ongoing debate concerning the benefits and drawbacks of information derived from open community participation. For example, the community-edited Wikipedia has been favorably compared with the professionally-edited Encyclopedia Britannica in terms of science article quality [6], but Wikipedia readers must cope with a small and increasing chance of viewing articles with intentional vandalism or misinformation [14]. In the Q&A domain, we find that the community in Yahoo! Answers provides surprisingly high-quality (aggregate) answers compared with the professionally-staffed library reference services, but Yahoo! Answers users must expect substantial variability in the quality of individual answers. On the other hand, the community discussion features in Google Answers appeared to add value to the system with no visible downside. Given the large body of literature that has shown extrinsic incentives "crowd out" intrinsic motivations [3], we might not expect the unpaid members of the Google Answers community to contribute as much value as they did. Future work might leverage this finding to better understand the properties of online community design that allow paid and free contributions to coexist.

In general, community Q&A sites are fertile ground for future work. They have integrated many compelling features that encourage work from their users, such as points, ratings, voting, and leaderboards. Research exploring the impact of these features (e.g. following the examples of [8] and [10]) on the quality and quantity of Q&A would help designers better understand when and where to deploy such features in their own online communities. Also, deep qualitative work holds much potential to better understand Q&A sites. There has been little research that seeks to understand what questions people ask, how they ask them, how they choose questions to answer, or how they respond to questions.

In conclusion, we leave you with a question: What change do you think will most improve Q&A sites of the future? $10 for the best answer.

ACKNOWLEDGMENTS

Thanks to Shilad Sen, Yan Chen, and John Riedl for their insightful comments. We gratefully acknowledge the support of the Israel Foundations Trustees (2006-2008) and the National Science Foundation, under grant IIS 03-24851.

REFERENCES

1. Box, G., Hunter, W, Hunter, S., Hunter, W. *Statistics for Experimenters*. John Wiley & Sons, New York (1978).
2. Chae, M., Lee, B. Transforming an Online Portal Site Into a Playground for Netizen. *J. of Internet Commerce 4*, 2 (2005).
3. Deci, E, Koestner, R., Ryan, R. A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation. *Psychological Bulletin 125*, 6 (1999).
4. Edelman, B. Earnings and Ratings at Google Answers. *Unpublished Manuscript* (2004).
5. Gazan, R. Specialists and Synthesists in a Question Answering Community. In *Proc. American Society for Information Science and Technology*, 43 (2006).
6. Giles, J. Internet Encyclopaedias Go Head to Head. *Nature 438* (2005), 900-901.
7. Janes, J., Hill, C., and Rolfe, A. Ask-an-Expert Services Analysis. *Journal of the American Soc. for Information Science Technology 52*, 13 (2001), 1106-1121.
8. Lampe, C., Resnick, P. Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proc CHI* (2004).
9. Lee, J. H., Downie, J. S., Cunningham, S. J. Challenges in Cross-Cultural/Multilingual Music Information Seeking. In *Proc ISMIR* (2005).
10. Ling, K., et al. Using social psychology to motivate contributions to online communities. *Journal of Computer Mediated Communication 10*, 4 (2005).
11. Leibenluft, J. A Librarian's Worst Nightmare: Yahoo! Answers, where 120 million users can be wrong. *Slate Magazine*, Dec. 7, 2007 (2007).
12. Mcclennen, M., Memmott, P. Roles in Digital Reference. *Information Technology and Libraries 20*, 3 (2001).
13. Pomerantz, J., Nicholson, S., Belanger, Y., Lankes, R. D. The Current State of Digital Reference. *Information Processing and Management 40*, 2 (2004), 347-363.
14. Priedhorsky, R., Chen, J., Lam, S., Panciera, K., Terveen, L., Riedl, J. Creating, Destroying, and Restoring Value in Wikipedia. In *Proc. GROUP* (2007).
15. Raban, D., Harper, F. Motivations for Answering Questions Online. In *New Media and Innovative Technologies* (2008).
16. Rafaeli, S., Raban, D., Ravid, G. How Social Motivation Enhances Economic Activity and Incentives in the Google Answers Knowledge Sharing Market. *Int. Journal of Knowledge and Learning 3*, 1 (2007), 1-11.
17. Rafaeli, S., Raban, D., Ravid, G. Social and economic incentives in Google Answers. *ACM Group 2005 Workshop: Sustaining Community: The role and design of incentive mechanisms in online systems* (2005).
18. Rousch, W. What's the Best Q&A Site? *Technology Review*, December 3, 2006 (2006).
19. Sang-hun, C. South Koreans Connect Through Search Engine. *New York Times*, July 5, 2007 (2007).
20. White, M. Diffusion of an Innovation: Digital Reference Service in Carnegie Foundation Master's (Comprehensive) Academic Institution Libraries. *Journal of Academic Librarianship 27*, 3 (2001), 173-187.