A comparative assessment of answer quality on four question answering sites

Pnina Shachaf
Indiana University, Bloomington

Abstract
Question answering (Q&A) sites, where communities of volunteers answer questions, may provide faster, cheaper, and better services than traditional institutions. However, like other Web 2.0 platforms, user-created content raises concerns about information quality. At the same time, Q&A sites may provide answers of different quality because they have different communities and technological platforms. This paper compares answer quality on four Q&A sites: Askville, WikiAnswers, Wikipedia Reference Desk, and Yahoo! Answers. Findings indicate that: 1) the use of similar collaborative processes on these sites results in a wide range of outcomes. Significant differences in answer accuracy, completeness, and verifiability were found; 2) answer multiplication does not always result in better information. Answer multiplication yields more complete and verifiable answers but does not result in higher accuracy levels; and 3) a Q&A site's popularity does not correlate with its answer quality, on all three measures.

Keywords
Q&A Sites; Social Q&A; Community Question Answering; Social Reference; Information Quality; Crowd-sourcing

1. Introduction

Question answering (Q&A) sites, like other Web 2.0 social sites, take advantage of the wisdom of crowds [1], and rely on users' participation [2]. Online communities of volunteers on these sites are processing millions of questions and answers online. For example, Yahoo! Answers has over 100 million users [3] and more than 23 million archived questions [4]; it is among the most frequently consulted reference sites, second only to Wikipedia. The social Web, and Q&A sites in particular, forces researchers to raise questions about the reliability of user-created content; some provide evidence that the quality of Wikipedia, for example, is as good as traditional encyclopedias [e.g., 5]. Q&A sites, like Wikipedia, build on the idea that everyone knows something [6], and through collaborative knowledge production users provide answers to questions and create an archive of questions and answers. The potential of these Q&A sites in crowd-sourcing question answering work (social reference) is contested [7], and the benefits of these sites are questionable. Do they pose a threat to our culture and traditional institutions by supporting a culture of mediocrity where *everything is miscellaneous* [8] and by fostering a *cult of amateurs* [9]? Or do they provide an opportunity for further development of our traditional institutions and practices by providing high quality information to users? In this context, it is critical to evaluate the quality of the information produced by means of crowd-sourcing on Q&A sites. The possible benefits of cost reduction and user participation are attractive features of crowd-sourcing, but at the same time, a major risk includes the potential provision of inferior services. More specifically, researchers should address questions such as: What is the quality of answers that users receive on these Q&A sites? Are these answers accurate, reliable, and complete? Do some of these sites provide better services than others? How good are these Q&A sites compared with professional question answering services (i.e., libraries)? Are users satisfied with these Q&A sites? How do users determine and choose the best answer to their question? How many users should collaborate to provide a good answer? What are the information seeking and use behaviours of Q&A site users? How can good answers be identified and captured for re-use? One of the first steps in answering these questions relies on a rigorous quality assessment of information quality on these sites.

This paper aims to understand if the collaborative process of question answering, which involves massive user participation, provides answers of high quality; it aims to identify which Q&A sites provides better answers. The study examines and compares answer quality on four Q&A sites, two of which are very popular Q&A sites, Yahoo! Answers and Wiki Answers, and two of which are less popular Q&A sites, Askville and the Wikipedia Reference Desk.

Information quality is determined by answer reliability, focusing on three specific measures: accuracy (correct answer), completeness (thorough answer), and verifiability (reference or link to information source). Based on content analysis of 1,522 randomly selected transactions[1] from these four Q&A sites, the study reveals significant differences in answer quality between the sites. The Wikipedia Reference Desk outperformed all other Q&A sites and provides much more accurate, complete, and verifiable information, while the most popular Q&A site, Yahoo! Answers, provides the lowest quality in terms of answer verifiability. The results also show that collaborative question answering processes are advantageous in terms of verifiability and completeness of answers but not in terms of accuracy. The whole answer, an answer that combines multiple responses, provided better information quality than the "best answer," the answer chosen by the user who asked the question or by a community vote.

2. Background and related work

Researchers from various disciplines (information retrieval, human computer interaction, reference, information seeking behaviour and use) are trying to grasp the depth and range of impact that Q&A sites have on traditional conceptions of information creation, dissemination, intermediation (reference), and use [10]. Answer quality is a common concern among all of these disciplines; yet, there is no clear agreement as to what constitutes a high quality answer and what measures should be employed.

Prior research on answer quality and on information quality can shed some light on this challenge. Research on answer quality has been one of the foci in reference research over the last several decades. While "a lack of attention [has been] given to theory" in reference research [11, p. 3], much of this research includes studies that evaluate the quality of reference services, including over 1,000 works that focus on digital reference services. Shachaf [7] uses this body of knowledge to propose a model for understanding social reference on Q&A sites and proposes variables for assessment of answer quality; these include, for example, accurate, complete, and verifiable answers. Information quality, a broader area of interest when considering answer quality, has attracted much research attention in a wide variety of disciplines; among them are business and management information systems (MIS), computer science and human computer interaction (HCI), communication, history, philosophy, and library and information science (LIS). Within LIS, information quality was examined through the lenses of information seeking behaviour research [e.g., 12] and reference research [13, 14], for example. One approach focuses on users' judgment of information credibility [12] and the other approach focuses on objective measures of an information artefact (website, answer), such as accuracy and completeness [13]. In the context of Web 2.0 and Q&A sites, Kim and Oh's [15] study of users' relevance criteria for evaluating answers and Kim's [16] study of users' credibility judgments of answers follow the first approach, while Stvilia et al.'s [17] study of information quality on Wikipedia and Shachaf's [18] study of the quality of answers on the Wikipedia Reference Desk embrace the second approach. The utilization of the second approach to information quality in the study of answer quality on Q&A sites can be useful, but poses certain challenges as the artefact is dynamic and multifaceted and may require multiple levels of analysis.

As the popularity of Q&A sites increases among users, scholarly interest in answer quality increases as well. Researchers have mostly focused on Yahoo! Answers and have based their assessment on user rankings of "best answers" [e.g., 4, 19]. "Best answer" is a feature of Yahoo! Answers that indicates an answer that was chosen by the user who asked the question, or by a community vote, as the best answer to have been posted in response to the same question. A few studies of answer quality extend beyond Yahoo! Answers [20, 21] and report that answer quality on Google Answers is better than Yahoo! Answers, AllExperts, and Live QnA [20]. At the same time, O'Neill [21, p. 10] contends that "responders at Yahoo! Answers and Askville could find it difficult to handle questions that really require an old fashioned reference interview and/or some knowledge of resources not easily uncovered by a simple search". These earlier studies suggest that answer quality varies across Q&A sites, but because of the limitations of the studies it is unclear which of the existing Q&A sites is better than the others. Two of Harper et al.'s [20] Q&A sites are no longer active (Google Answers ceased operation in 2006 and Microsoft shut down its Live QnA site in 2009) and the sample size in O'Neill's [21] study was too small to draw conclusive findings. Extrapolation from these studies is problematic also because they do not rely on a natural corpus of questions from these Q&A sites, but use proxy users that submit carefully designed questions and analyze their answers. While it is quite possible that answer quality will

[1] A transaction refers to a question and all of its respective answers.

vary, it is unclear which of these communities will provide superior answer quality and which of them are less reliable. Hence, a comparative study of answer quality on multiple Q&A sites is much needed.

The first step in such an assessment across multiple Q&A sites should be to identify the quality measures that could be used. Prior research on Q&A answer quality has mostly focused on Yahoo! Answers and has primarily assessed quality based on user rankings of "best answers" [e.g., 4, 19]. However, user rankings are problematic because they provide a subjective measure of answer quality. Users are not always knowledgeable enough about the topic of their own question and cannot accurately ascertain whether the answer is accurate or complete [18]. Further, users may not be able to distinguish between their gratitude toward those who helped them and the quality of the answer they received [22]. For example, in 29.8% of cases where users chose "best answers" in Yahoo! Answers, their selection was based on socio-emotional criteria rather than on the content or utility of the answer [15]. Poston and Speier [23] argue that "rating validity, describes the degree to which the rating reflects the intrinsic quality of the content…[it] may be low for a variety of reasons… [it is] inherently subjective and voluntarily provided, resulting in mismatch between the true quality of the content and the rating given… [and] those submitting ratings may manipulate ratings…" (p. 223). Further, for a given question, the "best answer" is not the unique and single good answer, and therefore the utility of "best answers" as an indication of answer quality is limited. In other words, it was found that "there are also good answers not among the best answers identified by Yahoo! Answers users" [24, p. 142], and that "at least 78% of the Q&A best answers are reusable when similar questions are asked again, but no more than 48% of them are indeed the unique best answers" [25, p. 497]. Another limitation of the "best answer" measure is that it focuses on comparing multiple answers on a given question, but does not assess the quality of the answer at the transaction level (between transactions). However, scholars argue that the quality of an individual answer (or first) answer is lower than the quality of a whole answer (an answer that combines multiple responses), because users can add information and modify the answer in a manner that improves answer quality [18, 26]. For example, in Naver's Knowledge –iN, the last answer is more likely to be selected as "best answer" because it improves or corrects previous answers [26]. Another limitation of the "best answer" measure is that the choice of "best answer" is based on site-specific criteria; these criteria differ from one Q&A site to another. Not only can the "best answer" measure on Yahoo! Answers differ from the "best answer" measure on Answerbag or Askville, it is irrelevant on sites that have no "best answer" feature, such as WikiAnswers and the Wikipedia Reference Desk. Therefore, findings from studies that utilize this measure should be generalized with great caution, and in many cases, extrapolations from Yahoo! Answers studies to other Q&A sites are irrelevant. Consequently, the "best answer" measure cannot be utilized for comparison of answer quality across Q&A sites.

Another method to identify answer quality is by tracking user reputation. This method is based on the assumption that certain users are more likely to provide better answers than others [27]. Examples of this approach include the ranking of authoritative responders using link analysis [28, 29] or user expertise [24]. Other ranking methods measure a user's reputation based on their activity level (e.g., 27, 30, 31], their credibility (authority), or the number of "best answers" they have previously posted [3, 30]. However, this approach is also problematic because even users with good reputations do not always provide high quality answers. Besides, user reputation is determined by site-specific criteria, which differ from one site to another and is not an applicable measure on all Q&A sites. Therefore, the utilization of the user reputation measure, just like the "best answer" measure, cannot be effectively utilized in a comparative analysis of Q&A sites.

User satisfaction is another measure for evaluating answer quality [18, 25, 32]; yet again, like "best answer," it is a subjective measure of answer quality. An examination of answer quality from the point of view of the user includes four measures [32, pp. 398-399]:

> completeness represents the degree to which the system provides all necessary information; accuracy represents the user's perception that the information is correct; format represents the user's perception of how well the information is presented; and currency represents the user's perception of the degree to which the information is up-to-date.

Measuring answer reliability, as an indication of answer quality, is not as common in Q&A research as approaches focused on "best answers" or user reputation, yet it can help in quality assessment across sites. Continuing a tradition of assessment of answer quality provided by information professionals (reference research), quality on Q&A sites was examined [e.g., 18, 20] and frameworks have been developed [7, 30, 32]. Under this approach, high quality answers were determined based on content analysis of the answers [18, 20, 30, 33, 34]. Scholars that analyzed the content of answers have found, for example, that better answers are longer [4, 20, 30], or include references to external sources [34]. Interestingly, Blooma, Chua, and Goh [30] report that question category, answer accuracy and completeness, and

length of answer are significant predictors of answer quality, whereas asker's and answerer's authority and reputation are not.

Thus, the present study utilizes three reliability measures that are not confined by site-specific criteria: accuracy, completeness, and verifiability [18, 30, 32]. These measures are used here to compare Q&A sites and test whether there are variations in answer quality across them. Specifically the study will test if the most popular site, Yahoo! Answers, provide better answers than the less popular sites (Askville, WikiAnswers and the Wikipedia Reference Desk). One could expect that Q&A sites with high traffic (many questions and answers) may reflect a high level of user satisfaction, which in turn may indicate a high level of answer quality and may correlate with high quality of information [32]. However, it is likewise possible that the quality of answers on Yahoo! Answers and WikiAnswers will be inferior to the other less popular Q&A sites, and that, despite having fewer questions, the Wikipedia Reference Desk and Askville will provide better answer quality. Furthermore, because researchers suggest that the quality of an individual answer is inferior to the quality of a whole response [7, 18, 26] reliability levels of "best answers" have been assessed and compared with the reliability levels of the whole answer (an answer that combines multiple responses), using the same measures. Specifically, the study tests the following hypotheses:

- H1: Yahoo! Answers will provide better answers than the other Q&A sites.
    - H1a: Yahoo! Answers will provide more accurate answers.
    - H1b: Yahoo! Answers will provide more complete answers.
    - H1c: Yahoo! Answers will provide more verifiable answers.
- H2: A whole response will be more accurate than a "best answer."
    - H2a: A whole response will be more accurate than a "best answer."
    - H2b: A whole response will be more complete than a "best answer."
    - H2c: A whole response will be more verifiable than a "best answer."

3. Method

Four Q&A sites have been examined in this study: Askville, WikiAnswers, the Wikipedia Reference Desk, and Yahoo! Answers (Table 1). These sites have been chosen because of their relative longevity, their visibility, and the fact that each of their parent organizations has an established and popular reference site; two sites, Yahoo! Answers and WikiAnswers, are the most popular Q&A sites [35].

Table 1. The four Q&A sites.

Q&A site's name	Parent organization and launching date	URLs
Askville	Askville, an Amazon Q&A community, was created by Joseph Park, Fai Leong, and Christian Cabanero and was launched in December 2006.	http://askville.amazon.com/Index.do
WikiAnswers	WikiAnswers was founded by Chris Whitten in 2002 as FAQ Farm and was acquired by Answers Corporation (Answers.com) in November 2006.	http://wiki.answers.com
Wikipedia Reference Desk	The Wikipedia Reference Desk was launched in 2001 as part of the Wikipedia project, but was not heavily used during the first couple of months. Larry Sanger (co-founder of Wikipedia and founder of Citizendium) asked the first question.	http://en.wikipedia.org/wiki/Wikipedia:RD
Yahoo! Answers	Yahoo! Answers was launched in December 2005 by Yahoo Inc. and become the most popular Q&A site soon after (Hitwise, 2008).	http://answers.yahoo.com/

3.1. Data collection

A random sample of 1,522 transactions from these four Q&A sites was collected. The sample includes transactions from Yahoo! Answers (N=584), WikiAnswers (N=605), Wikipedia Reference Desk (N=77), and Askville (N=256). First, for a pilot study, data from the Wikipedia Reference Desk was collected and analyzed; these transactions included questions and answers posted in April 2007 taken from all seven topical categories [18]. The small number of transactions from the Wikipedia Reference Desk is an accurate representation of the average number of questions posted on the Wikipedia Reference Desk per day [18]. Next, for the follow up study, data was collected from Yahoo! Answers, WikiAnswers, and Askville. Three Perl programs were used to harvest transactions from each topical category on each of these three Q&A sites. The programs were set up to collect the most recent questions per category over a 24-hour period at a random minute of every hour (24 points of time) and to collect all the relevant answers that were posted over the following 18 days.[2] To control for the possible bias in category choice on answer quality, the samples from each of the Q&A sites include questions from each of their topical categories. Because this sampling method yields unequal sample size drawn from each of the Q&A sites, the quality rate was determined in percentages and not frequency counts; these percentages have then been then used in the comparison across Q&A sites.

The response rates vary across Q&A sites (Table 2); not all questions have been answered. The highest response rate was found on the Wikipedia Reference Desk (96% of the questions are answered) and the lowest response rate was found on WikiAnswers (16% of the questions are answered). Given that WikiAnswers is the second most popular Q&A site [35], the low response rate is surprising. In order to verify that the WikiAnswers data set had not been corrupted, additional examination of 50 randomly selected, unanswered questions was conducted a year after data collection. Using the history feature on each of these questions, an effort was made to determine if in fact there was no answer on the day of data collection, or if this is only a feature of the sample. Nearly all (94%) of the 50 questions either had not been answered at all (66%) or were answered after the date of retrieval (28%); three were merged with (or were called the same as) another question (6%). Thus, the analysis shows that the transactions that do not have answers in the sample also did not have answers on the WikiAnswers site on the date of data collection, but 28% of them received answers after more than 18 days from the day of data collection. Still, WikiAnswers has the lowest response rate of all four Q&A sites in this study.

Table 2. Q&A sites response rates.

Q&A site	Response Rate	
	Answers/Questions	**Percent**
Askville	189/256	74%
WikiAnswers	99/605	16%
Wikipedia Reference Desk	74/77	96%
Yahoo! Answers	497/584	85%

3.2. Data analysis

Content analysis of each of the 1,522 transactions was conducted to identify the reliable answers on each of the four Q&A sites. “Content analysis is an empirically grounded method, exploratory in process and predictive or

[2] The sample includes questions posted on Sept 30, 2008 and all of their respective answers, which were posted by October 18, 2008.

inferential in intent" [36, p. xvii]. Content analysis of answers is a widely used method to evaluate answer quality on Q&A sites [e.g., 20, 30, 33, 34]; it is used here to compare answer quality across multiple Q&A sites. Content analysis of answers enables the evaluation of answer quality based on quantifying the presence or absence of quality codes in the answer.

Content analysis was done at the transaction level, focusing on answer quality and using three codes to measure reliability (Table 3). Using these codes, two coders coded the data at the transaction level (whole answer). The coders assigned a value (yes/no) for each code (accuracy, completeness, and verifiability) to each of the transactions. For two Q&A sites (Askville and Yahoo! Answers) the coder also assigned a value for each code to each of the "best answers," an individual answer. Only these two Q&A sites had a "best answer" feature. A "best answer" is chosen by the user who asked the question or by a community vote; the coder does not define the "best answer" but determines the quality of this individual answer that was chosen, in some transactions, as the "best answer."

First, frequency tables were created for each Q&A site tallying the presence of codes (yes values) for all transactions on the specific Q&A site. Then, a comparative table was created where percentages of codes per Q&A site were marked (Table 4). Finally, based on these comparative tables, statistical analysis using SPSS 17.0 was done to determine if the differences are statistically significant (Tables 5, 6).

The coders, who were graduate students, were instructed to determine the accuracy, completeness and verifiability of the answers as if they were asking the question, to the best of their own knowledge. They were asked to code "on the surface" [37], and were not expected to conduct thorough research, or verify the information with authoritative sources, nor were they asked to follow the information provided on the links or cited sources. Neuendorf [38] emphasizes the importance of inter-coder reliability not only to strengthen the validity of the study results, but also to enable division of the workload among multiple coders with large data sets. Inter-coder reliability was determined using simple agreement, also called percent agreement, which is the percent of all codes that a pair of coders agreed on [37]. To determine inter-coder reliability a second coder coded ten percent of the data from each Q&A site, and the results indicate high inter-coder reliability (.92 per Q&A); as a rule of thumb, coefficient of .90 or greater would be acceptable to all [37, 38].

Table 3. Codes and their definitions.

Code	Definition	Values
Accuracy	Accuracy of an answer refers to a correct response.	Yes/No
Completeness	Completeness of an answer refers to an answer that is thorough, provides enough information, and answers all parts of a multi-part question.	Yes/No
Verifiability	Verifiability of an answer refers to a response that provides a link or a reference to another source where the information can be found.	Yes/No

3.3. Limitations

In an effort to use random samples of real transactions from each of the four sites for comparing answer quality, data collection was conducted using the same method. But due to the method of data collection, the random samples, which were retrieved from each Q&A site, differ in size (i.e., the number of transactions per site varies). This is due to the fact that the number of transactions that are posted on these sites differs. For example, on the Wikipedia Reference Desk the average number of questions posted on all topical desks is 70 [7], while on Yahoo! Answers more questions are posted per hour. Also, the data from the Wikipedia Reference Desk was collected earlier than the other Q&A sites; there might be variations in answer quality over time even for the same Q&A site. Furthermore, because response rate varies across sites the sample size of transactions that had answers that could be analyzed varies further. To overcome this challenge the frequency tables were transformed into percentage tables before any comparison across the four Q&A sites was done. Cross tabulations analysis used percentages of codes to control for different sample size; none of

the cells in the contingency tables (of percentage) had less than five percent, meeting the required assumptions for cross tabulation (at least five cases per cell), and therefore such statistical analysis is also valid.

Another limitation of the study is its sample of Q&A sites: Askville, WikiAnswers, Wikipedia Reference Desk, and Yahoo! Answers. While the choice is justified by aiming to sample two popular sites and two less popular sites, the sample involves other inherent differences between the sites. For example, while WikiAnswers and the Wikipedia Reference Desk use wikis, the others do not. In addition, even the use of wiki is different on each of these two sites; WikiAnswers presents one collaborative answer, while the Wikipedia reference desk presents all individual answers. The Wikipedia Reference Desk in this respect is similar to Yahoo! Answers and Askville, in that all three of these Q&A sites present multiple answers. Despite this limitation the fact that the difference does not parallel the size differences may not be a major confounding problem. The Q&A sites differ also, besides size differences, in their utilization of the "best answer" feature, and those that have a "best answer" feature differ in the way they determine "best answers." Askville (at the time of data collection) and Yahoo! Answers had this feature, while the other two sites did not. Still, because this difference does not parallel the size distinction, it is not a major confounding problem.

4. Findings

On each of these four sites, the level of accuracy, completeness, and verifiability was assessed at the transaction level (N=1522). In addition, reliability of the "best answer" was assessed for all the transactions from two of the four Q&A sites, Askville and Yahoo! Answers (N=1,356); the other two Q&A sites, WikiAnswers and the Wikipedia Reference Desk, do not have a "best answer" feature.

Table 4. Answer reliability

		Askville	Yaoo!Answers	Wikipedia	WikiAnswers
Whole Answer	Accuracy* (not accurate; undetermined)	47% (1%;52%)	32% (6%;62%)	56% (26%;18%)	53% (36%;11%)
	Completeness	84%	75%	63%	77%
	Verifiability	43%	25%	76%	6%
Best Answer	Accuracy	50%	37%	NA	NA
	Completeness	82%	60%	NA	NA
	Verifiability	39%	9%	NA	NA

* Because accuracy is a relevant measure only for informational questions, the cells for accuracy include also the percentage of inaccurate answers followed by the percentage of answers with undetermined accuracy rate.

In order to examine the first hypothesis, the level of the reliability measures – accuracy, completeness, and verifiability – was examined for each of the four sites and cross-tabulation analysis was then conducted. The four Q&A sites differ on all three reliability measures (Table 4), but only for two of the quality measures are these differences statistically significant (Table 5). Cross-tabulation results show that the difference in the level of accuracy between the four sites is statistically significant, $\chi^2(1, N=400)=3.25$, $p=.07$. The Wikipedia Reference Desk provides the most accurate level of answers and Yahoo! Answers provides the least accurate answers. The first part of the first hypothesis (H1a) was therefore not supported. Completeness levels vary across sites but cross-tabulation results indicate that these variations are not statistically significant, $\chi^2(1, N=400)=.10$, $p=.74$. Thus, the second part of the first hypothesis (H1b) was not supported and Yahoo! Answers provides answers that are as complete as the other Q&A sites. These four Q&A sites differ significantly in the level of answer verifiability, $\chi^2(1, N=400)=74.82$, $p=.00$, where the Wikipedia Reference Desk provides the most verifiable answers and WikiAnswers provides the least verifiable information, followed by Yahoo! Answers. The third part of the first hypothesis (H1c) was not supported as well. The first

hypothesis (H1) was not supported, Yahoo! Answers does not provide better answers than the other Q&A sites, and the Wikipedia Reference Desk outperformed it on all three measures.

Table 5. Answer reliability (whole answers) across four sites

Variable (N=400, df=1)	χ^2	Cramer's V	*p* level
Accuracy *	3.25	.09	.07
Completeness	.10	.01	.74
Verifiability ***	74.82	.42	.00

* sig. < .1 ** sig. < .05 ***sig. < .01

To examine the second hypothesis (H2), the level of each of the three reliability measures was assessed for the "best answer" on data from Askville and Yahoo! Answers (Table 4). Cross-tabulation analysis was conducted to examine if the quality of the "best answer" differs from the quality of a whole answer. Results indicate significant differences for two of the three reliability measures on Yahoo! Answers (Table 6). Completeness and verifiability levels were significantly better for whole answers compared with "best answers" on Yahoo! Answers (for completeness $\chi^2(1, N=200)=5.12$, $p=.02$; and for Verifiability $\chi^2(1, N=200)=9.07$, $p=.00$); this trend was observed on Askville as well, but the differences there were not statistically significant (for completeness $\chi^2(1, N=200)=1$, $p=.31$; and for Verifiability $\chi^2(1, N=200)=.33$, $p=.56$). Therefore, H2a and H2b were supported. It is important to note, however, that accuracy levels did not significantly differ between "best answers" and whole answers on Yahoo! Answers $\chi^2(1, N=200)=.55$, $p=.45$, or on Askville $\chi^2(1, N=200)=1.8$, $p=.67$. H2c was not supported. Interestingly, the trend was opposite for accuracy compared with the other two measures; "best answers" were more accurate than the whole answer on both of these Q&A sites. Thus, a whole answer on Yahoo! Answers is significantly better than a "best answer" in terms of completeness and verifiability, and the second hypothesis (H2) was partially supported.

Table 6. Answer reliability of "best answer" vs. whole answer

	Variable (N=200, df=1)	χ^2	Cramer's V	*p* level
Askville	Accuracy	1.8	.03	.67
	Completeness	1	.06	.31
	Verifiability	.33	.04	.56
Yahoo! Answers	Accuracy	.55	.05	.45
	Completeness**	5.12	.16	.02
	Verifiability***	9.07	.21	.00

* sig. < .1 ** sig. < .05 ***sig. < .01

Table 7 provides a summary of the results of the hypothesis tests. The findings support the argument that answer reliability varies across Q&A sites, and provides evidence that the Wikipedia Reference Desk gives answers that are more accurate and verifiable than the other three Q&A sites. The most popular Q&A site, Yahoo! Answers, is the least accurate and, along with WikiAnswers, provides answers with the lowest level of verifiable information. The findings also partially support the argument that a whole answer is better than the "best answer" (only completeness and verifiability on Yahoo! Answers differed significantly); accuracy levels did not differ significantly between the "best answer" and a whole answer, even though "best answers" seemed to be more accurate than whole answers.

Table 7. Summary results of hypothesis tests

H1: Yahoo! Answers will provide better answers than the other Q&A sites.	**Not Supported**
H1a: Yahoo! Answers will provide more accurate answers.	Not supported
H1b: Yahoo! Answers will provide more complete answers.	Not supported
H1c: Yahoo! Answers will provide more verifiable answers.	Not supported
H2: A whole response will be more accurate than a "best answer."	Partially supported
H2a: A whole response will be more accurate than a "best answer."	Not supported
H2b: A whole response will be more complete than a "best answer."	Supported
H2c: A whole response will be more verifiable than a "best answer."	Supported

5. Discussion

Three questions surfaced from the findings and will be discussed next: 1) What are some of the reasons for the significant variations in quality of answers across Q&A sites? 2) How can the mixed findings of differences between "best answers" and whole answers be explained? 3) Does crowd-sourcing reference services on these Q&A sites provide a threat or an opportunity for our traditional institutions?

5.1. What are some of the reasons for the significant variations in quality of answers across Q&A sites?

The fact that Q&A sites are formed around different communities and use information and communication technologies differently may explain the variations in answer quality [39]. Q&A sites vary in their community size, user demographics (e.g., age, gender, education level), policies and training mechanisms, motivators and technology infrastructure and use, and possibly in question types. For example, policies on Yahoo! Answers are set in a top-down process, but they are developed bottom-up on the Wikipedia Reference Desk. Wikipedia users are identified with the larger Wikipedia community and not specifically with the Wikipedia Reference Desk; the larger Wikipedia community is engaged in mass knowledge production where information accuracy, reliability, and verifiability are major concerns. Unlike the Wikipedia Reference Desk, user profiles on the other three Q&A sites indicate users' activities and ranks on the Q&A sites, and not their contributions to and affiliation with their respective parent organizations. These factors may have an impact on user participation and motivations. These users are driven by both intrinsic and extrinsic motivations [26, 40]. On Yahoo! Answers the use of extrinsic motivations, such as the reputation system, aims to increase user contributions and answer quality [40], and on Naver's Knowledge-iN users are motivated in part through a point system [26]. Raban and Harper [40] claim that monetary incentives are important to attract users to a particular site, but that social motivations lead to persistent participation. Extrinsic user motivations, such as ratings, monetary incentives, and social gratification significantly differ from one site to another resulting in various levels of answer quality across Q&A sites. Furthermore, demographic variations (e.g., age, gender, education level) between the Wikipedia Reference desk and the other Q&A sites may be significant. For example, while the Wikipedia Reference Desk is a male dominated environment, most Q&A sites are female dominated, as they attract mainly stay-at-home moms and teenagers.[3] Still, similar patterns of participation have been observed on various Q&A sites, such as Naver's Knowledge-iN, Google Answers, Wikipedia Reference Desk, and Yahoo! Answers (e.g., 4, 18]. On these sites certain users that handle most of the questions [4, 18, 41], and users who exhibit higher level of participation provide better answers [26]. Raban and Harper [40] claim that on Q&A sites there is a power distribution of user participation, where there are a few active users and a long tail of far less active users, and that this pattern resembles participation on other

[3] Yahoo! Answers gender distribution is 60% female and 40% male. This gender distribution is based on a sample of 459 users that took part in transactions answering 100 randomly selected Yahoo! Answers questions, that have been analyzed in this study. Only the 101 users that mentioned their gender on their Yahoo! Answers user profiles are included.

online communities. While similar patterns of participation are common to these Q&A sites, some of the sites may be more supportive of collaboration while others may enable only microcollaboration. Gazan [42] claims that on Q&A sites, such as Answerbag, microcollaboration occurs; microcollaboration is a form of collaboration that occurs on sites that have not been designed to support collaboration. The Wikipedia Reference Desk significantly differs from Answerbag in that it supports collaboration. Taking this idea into consideration, it is possible that the extent to which specific Q&A sites support collaboration vs. only microcollaboration may have a major impact on the quality of answers produced through each of these collaborative models. It is clear that a better understanding of these various online Q&A communities could shed light on some of the reasons for the variations found in answer quality and future research into these communities is much needed [39].

Another possible explanation of the variations in answer quality may be rooted in differences across sites in question types. Scholars have claimed that on any given Q&A site, answer quality could vary based on question type [20, 43, 44]. Some questions would require a higher level of expertise to answer, and answering these questions would be more time consuming [7]. Gray and Meister [45], for example, differentiate between dyadic knowledge sourcing, published knowledge sourcing, and group knowledge sourcing. In prior research on Q&A sites, researchers have differentiated between conversational and informational questions [43], between subjective and objective questions [44], or, between navigational, informational, transactional, and social questions [25]. Measures such as accuracy, verifiability, and completeness are better suited in evaluation of answers to questions that are "informational" and "objective," rather than for evaluating "opinion", "conversational", or "social" questions. Thus it is possible that Q&A sites that attract less "informational" questions have a lower level of answer quality when using reliability measures compared with sites that attract proportionally more informational questions. It is possible that the Wikipedia Reference Desk has more informational questions than Yahoo! Answers, and therefore has a higher accuracy rate. In fact, accuracy rate was undetermined more frequently for answers on Askville and Yahoo!Answers, than on the Wikipedia Reference Desk or WikiAnswers (Table 4). Interestingly, he Wikipedia Reference Desk and WikiAnswers, who exhibit higher accuracy rates, also have the highest inaccurate answers rates. Future research should examine this possible explanation by first categorizing questions on multiple sites according to their types, and by then examining the differences and similarities in the proportions of various question types on these Q&A sites. It would likewise be useful to evaluate answers to the same questions on various Q&A sites, similar to Harper et al.'s study (2008).

5.2. How can the mixed findings of quality differences between "best answers" and whole answers be explained?

This study aimed to examine the claim that mass collaboration and information sharing can result in the provision of cost effective high service quality for and by prosumers [46] that can be utilized for customer support [32]. More specifically, following the expectation that the quality of an individual answer will be inferior to the quality of a whole answer [18, 26], the study found that a whole answer improves the amount of verifiable information compared to a single answer, and addressed multi-part questions better than a single answer. However, while it was possible for a whole answer to be more accurate than a typical single answer, it was not more accurate than a single "best answer." This may be a result of the fact that a whole answer may also include, in addition to the "best answer," which is more accurate than all other answers, answers that are inaccurate, conflicting, or contradicting. These multiple answers that are part of a whole answer form a forest of mediocrity and may confuse users (and researchers alike) in their quality judgments [18]. Still, while a whole answer was not found to be better than a "best answer" in terms of accuracy, a whole answer may be more accurate than a typical answer or the first answer [26]. Future research should examine if a whole answer is better than another typical answer, such as the first answer. Future research can also examine how many answers are needed to reach a critical mass for a response to be accurate and at what point additional information may potentially reduce answer quality and try to identify how many answers are optimal for highest quality.

5.3. Crowd-sourcing reference services

Given these findings, one wonders if these Q&A sites pose a threat to our traditional cultural institutions (for example, libraries) by supporting a culture of mediocrity, or, do they provide an opportunity for further development of our traditional institutions and practices (for example, reference services)?

Crowd-sourcing could potentially result in cost reduction and the provision of better services and products through amateur user participation. Yet, at the same time, a major risk includes the potential provision of inferior services. Prior

research on Q&A sites, mostly report that answer quality is as good and even better than the quality provided by traditional institutions, such as libraries [20], and argues that the collaborative question answering process on Q&A sites enables the provision of better services [7]. However, this study does not support this claim. Despite the fact that all of the Q&A sites in this study are engaged in collaborative problem solving processes, the quality of answers that they produce significantly differs from one site to another. The use of similar collaborative processes results in a wide range of outcomes (the Wikipedia Reference Desk outperformed all other Q&A sites). The lack of consistent improvement in answer quality was found both at the transaction level (a whole answer was not more accurate than a "best answer," but was significantly more complete and verifiable than a "best answer"), and across sites (Yahoo! Answers, the most popular Q&A site, did not provide better answers than the other Q&A sites). Site popularity (higher user participation) did not clearly correlate with better answers. Future research should look into the collaborative question answering model further and examine the evolution of each transaction, from the first answer to the last one, to determine whether there is an improvement that can be attributed to the collaborative process or not. The collaborative process in question answering should be unpacked and a better understanding of answer multiplication should be gained.

6. Conclusion

While most research on Q&A sites to date has focused on Yahoo! Answers and has rarely examined more than one Q&A site at a time, this study reports on a large scale comparative analysis of answer quality on four Q&A sites. The study utilized a LIS approach to determine answer quality, focusing on collaborative and dynamic artefacts; the use of this approach was instrumental for the evaluation of answer quality between Q&A sites (assessing quality at the transaction level), and within transactions (comparing an individual "best answer" to a whole answer).

The findings indicate that a Q&A site's popularity does not necessarily correlate with the quality of information it provides to users. The Wikipedia Reference Desk seems to be an outlier among these Q&A sites and outperformed the others, including Yahoo! Answers, the most popular Q&A site. The potential benefits of whole answers compared to "best answers" are questionable, as quality significantly improved only in terms of answer completeness and verifiability, but not in terms of answer accuracy.

Future research may enhance our understanding of the reasons for the differences between Q&A sites, looking into 1) the nature of the online community and its dynamics; and 2) the relationship between answer quality and types of questions, as well as the distribution of questions by type on each of these sites and across multiple sites. Future research may look further into the collaborative process of question answering in relationship to answer quality, by looking at 1) the differences between first answers, best answers, and whole answers; as well as 2) the relationship between the number of users and the quality of whole answers.

8. References

[1] Surowiecki J. *The wisdom of crowds.* New York; Anchor Books, 2004.

[2] O'Reilly T. What is Web 2.0? http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html (accessed 20 August 2008).

[3] Dom B and Paranjpe D. A Bayesian technique for estimating the credibility of question answerers. In *Proceedings of the Society for Industrial and Applied Mathematics (SIAM) (2006).* http://www.siam.org/proceedings/datamining/2008/dm08_36_Dom.pdf (accessed 20 August 20, 2008).

[4] Adamic LA, Zhang J, Bakshy E and Ackerman MS. Knowledge sharing and Yahoo! Answers: Everyone knows something. In: *Proceedings of the International World Wide Web Conference*, (Beijing, ACM, 2008).

[5] Giles J. Internet encyclopedias go head to head, *Nature,* December 14, 2005. http://www.nature.com/news/2005/051212/full/438900a.html (accessed 19 August 2008).

[6] Noguchi Y. Web searches go low-tech: You ask, a person answers, *Washington Post*, August 16, 2006: A01, http://www.washingtonpost.com/wp-dyn/content/article/2006/08/15/AR2006081501142.html (accessed 20 August 2008).

[7] Shachaf P. Social reference: a unifying theory. *Library & Information Science Research* 2010; 32(1): 66-76.

[8] Weinberger D. *Everything is miscellaneous: The power of the new digital disorder.* New York: Henry Holt & Co, 2007.

[9] Keen E. *The cult of the amateur: how today's Internet is killing our culture.* New York: Doubleday/Currency, 2008.

[10] Shachaf P and Rosenbaum H. Online social reference: A research agenda through a STIN framework. In: *Proceedings of the iConference 2009*, Feb 8-11, 2009, Chapel Hill, NC.

[11] Saxton ML and Richardson JV. *Understanding reference transactions: transforming an art into a scienc*e. San Diego, CA: Academic Press, 2002.

[12] Rieh SY. Judgment of information quality and cognitive authority in the Web, *Journal of the American Society for Information Science and Technology* 2002; 53(2): 145–161.

[13] Fallis D. On verifying the accuracy of information: Philosophical perspectives, *Library Trends* 2004; 52(3): 463–487.

[14] Frické M and Fallis, D. Indicators of accuracy for answers to ready reference questions on the Internet, *Journal of the American Society for Information Science and Technology* 2004; 55(3): 238–245.

[15] Kim S and Oh S. Users' relevance criteria for evaluating answers in social Q&A site, *Journal of the American Society for Information Science and Technology* 2009; 60(4): 716-727.

[16] Kim S. Questioners' credibility judgments of answers in a social question and answer site, *Information Research*, 2010; 15(2): paper 432. http://InformationR.net/ir/15-2/paper432.html (accessed December 6 2010).

[17] Stvilia B, Twidale MD, Smith LC and Gasser L. Information quality work organization in Wikipedia, *Journal of the American Society for Information Science and Technology* 2008; 59(6): 983–1001.

[18] Shachaf P. The paradox of expertise: Is the Wikipedia Reference Desk as good as your library? *Journal of Documentation* 2009; 65(6): 977-963.

[19] Bian J, Liu Y, Agichtein E and Zha H. Finding the right facts in the crowd: Factoid question answering over social media. In: *Proceedings of the International World Wide Web Conference*, (Beijing, ACM, 2008).

[20] F.M. Harper, D. Raban, S. Rafaeli and J.A. Konstan, Predictors of answer quality in online Q&A sites. In: *Proceedings of the Conference on Human Factors in Computing Systems*, (Florence, ACM, 2008).

[21] O'Neill N. Chacha, Yahoo!, and Amazon, *Searcher* 2007; 15(4): 7-11.

[22] Howell BJ, Reeves HB and van Willigen J. Fleeting encounters: A role analysis of reference librarian-patron interaction, *Reference Quarterly* 1976; 16: 124-129.

[23] Poston RS and Speier C. Effective use of knowledge management systems: A process model of content ratings and credibility indicators. *MIS Quarterly* 2005; 29(2): 221-244.

[24] Suryanto MA, Sun A, Lim E and Chiang RHL. Quality-aware collaborative question answering: Methods and evaluation. In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. (Barcelona, Spain, ACM, 2009).

[25] Liu Y, Li S, Cao Y, Lin C, Han D and Yu Y. Understanding and summarizing answers in community-based question answering services. In: *Proceedings of the 22nd International Conference on Computational Linguistics,* (Manchester, UK, ACM, 2008).
[26] Nam KK, Ackerman MS and Adamic LA. Questions in, knowledge in?: A study of Naver's question answering community. In: *Proceedings of the 27th International Conference on Human Factors in Computing Systems* (Boston, MA, ACM, 2009).
[27] Bouguessa M, Dumoulin B and Wang S. Identifying authoritative actors in question-answering forums: The case of Yahoo! Answers. In: *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (Las Vegas, NA, ACM, 2009).
[28] Jurczyk P and Agichtein E. Discovering authorities in question answer communities by using link analysis. In: *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, (New York, ACM, 2007a).
[29] Jurczyk P and Agichtein E. Hits on question answer portals: Exploration of link analysis for author ranking, *Poster presented at the Annual ACM Conference on Research and Development in Information Retrieval*, (Amsterdam, ACM, 2007b).
[30] Blooma JM, Chua AYK and Goh DH. A predictive framework for retrieving the best answer. In: *Proceedings of the 2008 ACM Symposium on Applied Computing*, (Fortaleza, Ceara, Brazil, ACM, 2008).
[31] Chen W, Zeng Q and Wenyin L. A user reputation model for a user-interactive question answering system. In: *Proceedings of the Second International Conference on Semantics, Knowledge, and Gri*d, (Washington, DC, IEEE Computer Society, 2006).
[32] Ong Cc Day M and Hsu M. The measurement of user satisfaction with question answering systems, *Information & Management* 2009; 46(7): 397-403.
[33] Agichtein E, Castillo C, Donato D, Gionides A and Mishne G. Finding high-quality content in social media. *Proceedings of Web Search and Web Data Mining*, (Palo Alto, CA, ACM, 2008).
[34] Gazan R. Specialists and synthesists in a question answering community. In: *Proceedings of the American Society for Information Science & Technology Annual Meeting* 2006; 43(1): 1-10.
[35] Hitwise, *U.S. Visits to Question and Answer Websites Increased 118 Percent (2008).* http://www.hitwise.com/news/us200803.html (accessed 25 November 2009).
[36] Krippendorff K. *Content analysis: An introduction to its methodology*. 2nd ed. Thousand Oaks, CA: Sage, 2004.
[37] Lombard M, Snyder-Duch J and Bracken CC. Content analysis in mass communication: Assessment and reporting of intercoder reliability, *Human Communication Research* 2002; 28(4): 587–604.
[38] Neuendorf KA. *The content analysis guidebook.* Thousand Oaks, CA: Sage, 2002.
[39] Rosenbaum H and Shachaf P. A structuration approach to online communities of practice: The case of Q&A communities, *Journal of the American Society for Information Science and Technology* 2010; 61(10): 1933-1944.
[40] Raban D and Harper M. Motivations for answering questions online. In: D. Caspi and T Azran (eds), *New Media and Innovative Technologies.* Beer Sheva, Israel: Ben-Gurion University Press, Tzivonim Publications, 2007).
[41] JZhang J, Ackerman MS, Adamic L and Nam KK. QuME: A mechanism to support expertise finding in online help-seeking communities. In: *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology* (Newport, RI, ACM, 2007).
[42] Gazan R. Microcollaborations in a social Q&A community, *Information Processing & Management* 2010; 46(6): 693-702.
[43] Harper FM, Moy D and Konstan JA. Facts or friends?: Distinguishing informational and conversational questions in social Q&A sites. In: *Conference on Human Factors in Computing Systems*, (Boston, MA, ACM, 2009).
[44] Li B, Liu Y, Ram A, Garcia EV and Agichtein E. Exploring question subjectivity prediction in community QA. In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (Singapore, ACM, 2008).
[45] Gray PH and Meister DB. Knowledge sourcing effectiveness, *Management Science* 2004; 50(6): 821-834.
[46] Tapscott D and Williams AD. *Wikinomics: How mass collaboration changes ev*erything (Penguin Group Inc, New York, 2007).