Received 5 June 2016
Revised 16 August 2016
24 October 2016
23 November 2016
Accepted 23 November 2016

Two's not always company: collaborative information seeking across task types

Chirag Shah
School of Communication and Information, Rutgers University, New Brunswick, New Jersey, USA
Chathra Hendahewa
Department of Computer Science, Rutgers University, New Brunswick, New Jersey, USA, and
Roberto González-Ibáñez
Departamento de Ingeniería Informática, Universidad de Santiago de Chile, Santiago, Chile

Abstract

Purpose – The purpose of this paper is to investigate when and how people working collaboratively could be assisted in a fact-finding task, specifically focusing on team size and its effect on the outcomes of such a task. This is a follow-up to a previously published study that examined exploratory search tasks.

Design/methodology/approach – This research investigates the effects of team size on fact-finding tasks using a lab study involving 68 participants – 12 individuals, ten dyads, and 12 triads. The evaluation framework developed in the preceding work is used to compare the findings with respect to the earlier traditional exploratory task (Task 1) and the complex fact-finding task reported here (Task 2), with task type being the only difference.

Findings – The analyses of the user study data show that while adding more people to an exploratory search task could be beneficial in terms of efficiency and effectiveness, such findings do not apply in a complex fact-finding task. Indeed, results showed that the individuals were more efficient and effective doing Task 2 than they were in Task 1. Moreover, they outperformed the dyads and triads in Task 2 with respect to these two measures, which relate to the coverage of useful information and its relation to the expression of information needs. If the total time taken by each team is disregarded, the dyads and triads did better than the individuals in answering the fact-finding questions. But considering the time effect, this performance boost does not keep up with the increased group size.

Originality/value – The findings shed light not only on when, how, and why certain collaborations become successful, but also how team size affects specific aspects of information seeking, including information exposure, information relevancy, information search, and performance. This has implications for system designers, information managers, and educators. The presented work is novel in that it is the first empirical work to show the difference in individual and collaborative work (by dyads and triads) between exploratory and fact-finding tasks.

Keywords Evaluation, Tasks, Group work, Information seeking, Collaborative search, Fact-finding

Paper type Research paper

1. Introduction

Working collaboratively could be beneficial in many situations that present an extremely difficult or even impossible task for an individual. Many recent works have shown the effectiveness of collaborative work in information-intensive tasks. Examples of such works range from healthcare (Reddy and Jansen, 2008) and education to office work (Hansen and Järvelin, 2005) and design (Olson *et al.*, 1992). However, these works often ignore potential downsides to collaboration. It may seem that most informational problems could be addressed better with multiple people working together. The question is – when is it not true?

The current paper investigates this question with a user study designed with two independent variables: task type (exploratory search and problem solving), and team size

© Emerald Publishing Limited
2050-3806
DOI 10.1108/AJIM-06-2016-0075

(individuals, dyads, and triads). The experimental results point out clear differences in the nature and outcomes of collaborations among the different team sizes with respect to the two tasks.

The rest of the paper is organized as follows. In the next section, relevant works from the literature concerning collaborative information seeking (CIS) and complex fact-finding tasks with informational queries are reviewed. The need to identify conditions that call for CIS is expressed and linked to the design of a new study described in the method section. This is followed by a description of data and the results. The discussion section provides further interpretations of the results, along with the limitations of this work. The paper concludes with a summary of the findings and directions toward future research.

2. Background

This section serves two purposes: providing an overview of CIS, and summarizing the evaluation framework used to investigate CIS within an exploratory search task, which is a precursor to the more recent work reported in this paper.

2.1 CIS

While the argument that information seeking is a social activity that warrants support for collaboration among information seekers is not new (Twidale *et al.*, 1997; Morris, 2007), recent years have seen a significant uptick in research and development related to CIS. A brief review of CIS can be found in this study's preceding article (Shah *et al.*, 2015). A more comprehensive review is provided by Shah (2014b).

Scholars have often argued that for CIS to be meaningful and successful, certain conditions must exist. For examples, Shah (2008, 2014a) presented the following conditions for a CIS project to succeed:

(1) the participants of a team possess different backgrounds and expertise;

(2) the participants have opportunities to explore information on their own without being influenced by other team members, at least during a portion of the entire information seeking process;

(3) the participants should be able to evaluate the discovered information without always consulting others in the group; and

(4) there has to be a way to aggregate individual contributions to arrive at the collective goal.

In other words, collaboration in information seeking activities is not always desired and/or fruitful. The question is – where do we draw the line? González-Ibáñez, Haseki and Shah (2012a) hypothesized that, as long as there are more aspects of a collaborative project than there are collaborators in a synchronous collaboration, said project should benefit from collaboration.

To answer the question and to test the hypothesis presented above, we did an experimental study. This study is a follow-up of one we previously reported (Shah *et al.*, 2015). In the earlier work, we showed that for an exploratory search task with a multi-faceted topic, two searchers are more successful than one, and three searchers are more successful than two. In this follow-up study, we asked the same participants to work on a different type of information seeking task focused on fact-finding with the same collaborative conditions (individuals, dyads, and triads). Section 3 describes this new user study. Before proceeding, we need to summarize how we previously measured various quantities relating to information seeking, as we will use the same evaluation framework here. This consistency allows us to compare CIS across different task types.

2.2 Information search and fact-finding tasks

Since a major part of the work described here concerns with information search task types, search behavior, task difficulty, types of web search queries, and fact-finding task types, this section provides a background of how such tasks are used in studies related to information seeking.

Different types of search tasks have been described in the literature. For example, Marchionini (2006) identifies three major types of search activities, namely, lookup, learn, and investigate. The former (lookup) is described as the "most basic kind of search task" (p. 42) and it involves tasks such as fact retrieval, known item search, and navigation. On the other hand, learn and investigate activities are associated to exploratory search, which are often described as complex search tasks.

Some researchers have investigated differences among different types of search tasks. For instance, Kellar *et al.* (2007) compared users' behaviors in different types of tasks. In their study, users were asked to classify themselves the different tasks they had to solve (i.e. fact finding, information gathering, transactions, and browsing). In particular, fact-finding tasks were characterized by their short duration, small number of pages visited, little use of browser functions, and relatively long queries. In turning our attention to the fact-finding task, which is the focus of the type of task used in the study described in this paper (in Section 3), the authors defined it as "a task in which you are looking for specific facts or pieces of information. These are usually short lived tasks that are completed over a single session because either you find the answer or you do not. Examples include looking for tomorrow's weather, a pizza dough recipe, or printer drivers for your printer" (p. 10).

With a focus on fact-finding search tasks, Aula *et al.* (2010) conducted two studies – one qualitative and the other quantitative – to investigate the effects of task difficulty on search behaviors. The type of questions they used was similar to a Google-a-Day questions (e.g. "You once heard that the Dave Matthews Band owns a studio in Virginia but you don't know the name of it. The studio is located outside of Charlottesville and it's in the mountains. What is the name of the studio? (difficult)," "I was watching the movie *Stand by Me* the other day. I know it is based on a Stephen King story with a different name. What is the name of the story? (easy)". The authors found that when users experience more difficulty in finding information, they end up using more diverse queries, advanced search operators, and spending more time in pages. Unlike previous work by White and Drucker (2007), Aula *et al.* (2010) found that even in well-defined search tasks such as fact finding, searchers can behave as explorer and not necessarily as navigators.

In the taxonomy of web search as defined by Broder (2002), the type of web search queries falls under informational, navigational, and transactional queries. In the case of the multiple-step fact-finding task that is being addressed in this paper, informational queries play a major role since in order to find the final answer to the task, the searchers have to issue various queries that lead to finding information in a sequential manner. Further, since the fact-finding task spans across multiple steps, the searchers might issue navigational queries to go to specific webpages in order to find the final answer based on what they find in the intermediate steps of their search.

In a study conducted by Bilal (2000), she extensively studied the search behaviors of young students in fact-finding tasks. From her study, she concluded that the success level of students achieving the task goals varied according to the task type and students were less successful in fact-finding tasks as opposed to research tasks in finding relevant information.

González-Ibáñez and Shah (2014a) used the "A-Google-A-Day[1]" question bank for the multiple-step fact-finding questions in their study of CIS and the effects of affective dimensions. They argued that by using the same level of difficulty (level 2) from the "A-Google-A-Day" questions, which corresponds to the ideal number of search steps required to find the answer to each question, ensured that the users of the study were exposed to the same levels of perceived difficulty, response precision, topic familiarity, and response time. In order to ensure that the study was conducted in a controlled setting, the difficulty level of the questions (level 2) was maintained in their study across experimental conditions.

In the area of information search, some of the key concepts that are being researched are user intent in search, learning as part of search process, to name a few. Rieh *et al.* (2016) provided an extensive review on the relationship between information search and learning and also proposed the concept of "comprehensive search" in which a searcher goes beyond receptive learning and moved in to more critical and constructive learning. Jansen *et al.* (2008) presented an analysis of users' intent in web search, which can be classified into informational, navigational, and transactional intents. Through their analysis and classification system, they concluded that more than 80 percent of the online web search queries were informational queries that usually relate to finding some information about a certain topic. Based on this finding, it can be seen that most of the searchers use online search as a medium to find information about topics that are of interest to the, hence, user intent mining and providing relevant and reliable information for web search queries have become important.

2.3 Evaluation framework

In this paper, we will use an evaluation framework presented in Shah *et al.* (2015), although the task at hand is different from the exploratory-type tasks that are used in the aforementioned work. The main purpose of using such existing metrics and evaluations is the ability to provide a comparative evaluation for CIS across different task types. Following is a brief summary of various measures described in Shah *et al.*'s (2015) evaluation framework. The reader is referred to Shah *et al.* (2015) for complete details.

Information exposure. As specified in Shah *et al.* (2015), the aspect of information exposure refers to the amount and quality of information being discovered in the process of online information searching. Since the mode of information searching in this paper focuses on online web search, this aspect is mainly evaluated around the webpages being visited by a user/team, as outlined in Table I.

Information relevancy. In addition to the amount of information being found, it is important to identify the amount of relevant information gathered by user/teams to complete their task at hand. The metrics used to evaluate the information relevancy are shown in Table II.

Information search. Information search attempts to capture the ways in which information searchers seek information. In online web search, most of the information searching occurs through search engine queries. Therefore, the metrics used to evaluate this aspect consider search queries and direct query artefacts – such as search engine results pages – as specified in Table III.

Search performance. In evaluating the overall success of a search process, traditional metrics such as *F*-score that incorporate precision and recall can be considered important. In addition, measures of effectiveness and efficiency in CIS as proposed by González-Ibáñez, Shah and White (2012b) can be utilized, as shown in Table IV.

Evaluation measure	Acronym	Definition
Coverage	$C(i)$	$\{p_1, p_2, \ldots, p_{\mathrm{n}}\}$ p_{n} is the distinct webpage visited by user/team
Universe of distinct pages	$U(i)$	$\bigcup_i C(i)$
Unique coverage	UC(i)	$C(i)$ only visited by user/team (i)
Likelihood of discovery for webpage (p_n)	LD(p_n)	$\left(-1 \times \lvert p_n \rvert\right)/(\lvert U \rvert)$
Likelihood of discovery for user/team (i)	LD(i)	$\sum_{n=1}^{\lvert C(i) \rvert} \mathrm{LD}(p_n)/\lvert C(i) \rvert$

Table I. Evaluation metrics for information exposure

3. User study

Since we replicated the study described in Shah *et al.* (2015) here with the task being the only difference, the reader is advised to refer to the original paper for the descriptions of the participants, the experimental system Coagmento (González-Ibáñez and Shah, 2011), the methods and tools for logging the data, and the other instruments and procedures used in these experiments.

3.1 Participants

In the experiments reported here, we used the original exploratory search task's same 68 participants in their originally assigned conditions, as listed in Table V.

Table II. Evaluation metrics for information relevancy

Evaluation measure	Acronym	Definition
Relevant coverage	$RC(i)$	The set of distinct webpages that user/team found and marked as relevant based on collected snippets
Universe of relevant pages	$UR(i)$	$\bigcup_i \mathrm{RC}(i)$
Unique relevant coverage	URC(i)	UC(i)∩UR(i)
Number of snippets saved	Snip(i)	$\lvert\bigcup_i \text{Snippets collected}_i\rvert$
Precision	Precision(i)	$\lvert\mathrm{RC}(i)\rvert/\lvert C(i)\rvert$
Recall	Recall(i)	$\lvert\mathrm{RC}(i)\rvert/\lvert UR(i)\rvert$

Table III. Evaluation metrics for information search

Evaluation measure	Acronym	Definition
Distinct queries	$Q(i)$	$\{q_1,q_2,\ldots,q_m\}$ q_m is the distinct query issued by user/team
Search engine results pages	SERP(i)	$\{s_1,s_2,\ldots,s_x\}$ s_x is the distinct SERP clicked by user/team
Query diversity	QD(i)	Mean$\{$levenshtein distance $\{q_1,q_2\}\}$, $q_1 \neq q_2 \wedge \{q_1, q_2\} \in Q(i)$
Query entropy	$E(q_m)$	$-\sum_{u=1}^{\lvert\mathrm{unigrams}_{q_m}\rvert} p_u \log_2 p_u$
Information content	IC(i)	$\sum_{m=1}^{\lvert Q(i)\rvert} E(q_m)/\lvert Q(i)\rvert$

Table IV. Evaluation metrics for search performance

Evaluation measure	Acronym	Definition
F-score	$F(i)$	$\frac{2\times\mathrm{Precision}(i)\times\mathrm{Recall}(i)}{\mathrm{Precision}(i)+\mathrm{Recall}(i)}$
Effectiveness	Effectiveness (i)	$\frac{\left\lvert\cup_i\left\{p_n\left(\text{dwell time}_{p_n} \geqslant 30 \text{ secs}\right)\right\}\right\rvert}{\lvert C(i)\rvert}$
Efficiency	Efficiency (i)	$\frac{\mathrm{Effectiveness}(i)}{\lvert Q(i)\rvert}$
Response precision	Response precision (i)	$\frac{\lvert\text{Correct answers}(i)\rvert}{\lvert\mathrm{Answers}(i)\rvert}$

Table V. Participants in different experimental conditions

Condition	Description	No. of units	Total participants
C1	Individuals	12	12
C2	Dyads	10	20
C3	Triads	12	36

3.2 Session workflow

Coagmento was adapted to guide users through various stages in the session workflow of this study. The system led users from stage 2 up to stage 5 via the automatic progression of stages upon completion of each stage either based on user input or allocated time elapsed as shown in Table VI. A researcher conducting the study session guided stages 1 and 6.

3.3 Task

The participants were asked to use online search to answer as many questions as possible during a finite period of time. The set of questions was obtained from "A-Google-A-Day." Each question had a unique short answer; however, different search paths could be taken to find it. Here is an example.

Question: how long is the river bordering the two countries that once were home to the Hamangia?

How to find the answer (Google's strategy): search (Hamangia). You will find that the Hamangia culture is a late Neolithic culture that once existed in what is now Romania and Bulgaria. Search (river bordering Romania and Bulgaria). You will find the answer is the Danube River. Then search for the (length of Danube River), and discover that it is about 1,777 miles long.

Answer: 1,777 miles.

Note that questions used in this study were published on A-Google-A-Day on August 2011. Moreover, questions were classified in different levels of difficulty as described in González-Ibáñez and Shah (2012, 2014b). Specifically, the level of difficulty was determined based on the number of steps/queries required to find the answer, which was implicitly specified in the solutions provided by A Google Day. In this study, we only used level-3 questions, where number 3 represents the ideal number of steps necessary to find the answer to each question. As a result, we fixed this variable so that participants in the different conditions of the study and regardless their experimental condition, were exposed to the same level of difficulty.

The users were allowed to access Google search engine as a part of the search system to find answers to the questions. We ensured that the history of the browser was deleted prior to each study session and the users were not allowed to login to their Google profile before starting the search to ensure that every user was exposed to the same conditions while searching without introducing bias due to prior browsing history. We did not restrict the aspect of whether users used Google Knowledge Graph[2] or only organic search listings since Google Knowledge Graph is an underlying component of Google search engine that has been introduced since 2012. Since all the users were exposed to the same search conditions, we can conclude that the benefits of Google Knowledge Graph or any other algorithms underlying the search engine were common to all users in the study.

Stage	Description	Time (min)
1	Because this was a continuation of the same user study during a second session, users were given instructions and reminders	2
2	Participants watched a brief tutorial in order to remind themselves of the basic functionalities required during the task	2
3	Participants worked on a simple practice task to get accustomed to the system	2
4	Each user/team worked on the fact-finding task by answering a series of "A-Google-A-Day" questions by searching and collecting relevant information and submitting short answers based on their findings	35
5	Participants filled out post-task questionnaires	3
6	Participants were briefly interviewed to get their feedback on the task and their experience	5

Table VI. Summary of session stages

However, given the nature of the questions (finding obscure piece of information that is not clearly connected with known concepts), it was highly unlikely that the users were able to get much, if anything, specifically from the Knowledge Graph.

Overall, user/teams were given 35 minutes to complete the entire task. The order in which the fact-finding questions were presented to the users was pre-determined at the system level so that every individual/team received the questions in the same order to avoid any bias introduced by question order. Individuals or teams were given the freedom to use as much or as little time as they wished to answer each question with a maximum time limit of 35 minutes. However, they were allowed to skip questions at their discretion at any time. They were provided with a text box to write their short answer for each question upon clicking the "Answer" button. Note that for the particular case of dyads and triads, all the participants in the group were required to agree in order to either submit their answer or skip the question (until all members of the team clicked the "ok" button, the next question was not presented to them by the system). In the rare event that an individual or team ran out of all the questions in the system's question bank within the specified time, the questions they had skipped were presented to them. All other settings and variables were kept the same as reported in Shah *et al.* (2015).

4. Results

The analysis of the data focused on four main research measures:

(1) The first is effectiveness: effectiveness measures the ratio of the number of pages with a dwell time greater than 30 seconds to the number of distinct webpages the groups visited. Dwell time, the length of time an individual lingered on a page, was set at 30 seconds based on existing research literature (Fox *et al.*, 2005; White and Huang, 2010). The number of distinct webpages visited by a group is also known as the coverage.

(2) The next measure taken into account was efficiency: efficiency is the ratio of effectiveness and the number of distinct queries run by the participants.

(3) Relevant coverage is the set of webpages that the individuals or teams found relevant by collecting snippets.

(4) Query diversity was found by averaging over the Levenshtein Distance or Edit Distance (Levenshtein, 1966) for all pairs of distinct queries that a given individual or team ran.

This paper focuses only on the analysis of the fact-finding task (Task 2). The analysis of Task 1 (exploratory task) and the corresponding results were presented in the paper by Shah *et al.* (2015).

For data analysis, each group of individuals, dyads, or triads was automatically assigned a project ID as they engaged in the task. The data for each metric were then aggregated for each project ID and each question completed. Missing values for questions skipped or left unanswered were simply coded "NA" and omitted from the analysis. Then the mean of each feature was calculated for a given study condition on a question-by-question basis. Statistical analysis was performed at the study level, so the performance of individuals, dyads, and triads could be compared across all questions for a given feature.

The Shapiro-Wilk test revealed that the data did not follow a normal distribution for each study condition. Therefore, the Kruskal-Wallis analysis of variance was performed to check for initial statistical significance. The data were grouped according to study condition. When the results of the Kruskal-Wallis test were significant, the Wilcoxon rank-sum test was used to test pairwise significance between study conditions. Project IDs were sorted by their study condition, and then the mean values of all the measures for each question were compared. For analysis between Tasks 1 and 2, feature values were compared for each project ID given that Task 1 only contained one question. For example, all project IDs with Condition 1 from Task 1 were compared to all project IDs with Condition 1 from Task 2.

Table VII presents the mean, median, and standard deviation for effectiveness, efficiency, relevant coverage, and query diversity across all questions (within the fact-findings task) for each study condition. For effectiveness and efficiency, Condition 1 (individuals) consistently had the highest mean, median, and standard deviation followed by Condition 2 (dyads) and then Condition 3 (triads), indicating that individuals were more effective and efficient in fact-finding tasks than groups. This might be indicative of the extra time that a group had to spend coming to a consensus before making the final decision at each question/answer, whereas individuals could proceed on their own pace. However, this trend does not hold for relevant coverage and query diversity. For relevant coverage, triads had the highest mean, median, and standard deviation followed by dyads and then individuals. Dyads also had the greatest standard deviation for query diversity, followed by triads and then individuals. This may indicate differences of ideas, opinions, and implementations (through the use of queries) among collaborators.

Table VIII contains the results from the statistical tests performed across all questions asked for each study condition. For effectiveness, the Kruskal-Wallis test yielded a χ^2 value of 14.223 with a $p < 0.01$. The Wilcoxon rank-sum test showed a $p < 0.01$ across all pairs of study conditions. When compared with the medians from Table VIII, it is clear that individuals are more effective than dyads and triads. For efficiency, Kruskal-Wallis had a χ^2 value of 26.664 with a $p < 0.01$. The Wilcoxon test returned a $p < 0.01$ across all pairs of study conditions. The medians in Table VII indicate that Condition 1 is greater than Conditions 2 and 3 for efficiency. While the Kruskal-Wallis test was significant for relevant

Measure	Condition 1 (individuals)	Condition 2 (dyads)	Condition 3 (triads)
Effectiveness	Mean: 0.681 Std: 0.221 Median: 0.689	Mean: 0.466 Std: 0.210 Median: 0.421	Mean: 0.421 Std: 0.167 Median: 0.367
Efficiency	Mean: 0.332 Std: 0.249 Median: 0.250	Mean: 0.102 Std: 0.085 Median: 0.062	Mean: 0.068 Std: 0.061 Median: 0.044
Relevant coverage	Mean: 1.462 Std: 0.376 Median: 1.500	Mean: 2.148 Std: 0.820 Median: 2.165	Mean: 2.362 Std: 0.877 Median: 2.598
Query diversity	Mean: 12.714 Std: 8.442 Median: 12.918	Mean: 19.102 Std: 9.305 Median: 17.845	Mean: 24.730 Std: 8.668 Median: 27.513

Table VII. Statistical measures for fact-finding task (Task 2)

Measure	Kruskal-Wallis	Wilcoxon C1, C2	Wilcoxon C1, C3	Wilcoxon C2, C3	Interpretation
Effectiveness	$\chi^2 = 14.223$**	$W = 57$** C1 > C2	$W = 47.5$** C1 > C3	$W = 0$** C2 > C3	C1 > C2, C3
Efficiency	$\chi^2 = 26.664$**	$W = 19$** C1 > C2	$W = 19$** C1 > C2	$W = 0$** C2 > C3	C1 > C2, C3
Relevant coverage	$\chi^2 = 12.227$**	$W = 792$	$W = 679$	$W = 486.5$	
Query diversity	$\chi^2 = 5.289$				

Notes: *$p < 0.05$; **$p < 0.01$

Table VIII. Statistical test results for fact-finding task (Task 2)

coverage with a p-value below 0.01 and a χ^2 value of 12.227, the Wilcoxon results were not significant across any pair of study conditions and consequently no interpretation can be made. The Kruskal-Wallis test was not significant for query diversity and so the Wilcoxon test was not performed.

Table IX represents a comparison between Tasks 1 and 2 for effectiveness, efficiency, relevant coverage, and query diversity. The evaluation was performed using Wilcoxon Signed Rank test to identify if the distributions of these metrics differ within subjects (in this case, within dyads, triads, and individuals who performed both Tasks 1 and 2). For all metrics, other than efficiency for dyads (C2) and triads (C3), the Wilcoxon rank-sum test yielded significant results. This indicates that the metrics for Tasks 1 and 2 for the individuals, dyads and triads were non-identical with statistically significant differences. Individuals were more effective and efficient in Task 2 compared to Task 1, which may be indicative of the fact that individuals are highly effective in fact-finding tasks compared to exploratory tasks. Furthermore, dyads and triads were also more effective in Task 1 than in Task 2, yet not significantly different in terms of efficiency, showing that when working as a team, they were not able to benefit much from synergic effects in the fact-finding task as opposed to the exploratory task. In both relevant coverage and query diversity, Tasks 1 and 2 populations showed statistically significant results where Task 1 values were always higher than Task 2.

These findings on relevant coverage show that in the fact-finding task (Task 2), users spent less time reading and collecting snippets because they needed to find a specific answer, whereas in the exploratory task (Task 1), collecting snippets that could be used in the final report and keeping track of the areas covered was quite important. Query diversity is significantly higher in Task 1 compared to Task 2. This indicates that for exploratory tasks, users had to explore more unknown paths using varied queries when compared to fact-finding tasks.

Another way to calculate the performance of individuals, dyads and triads in Task 2 is based on their response precision. This measure was operationalized as the ratio between correct answers and the total number of answers provided (González-Ibáñez and Shah, 2014a). Unlike Task 1, in Task 2 it was possible to accurately determine whether answers provided by the individuals or groups were correct by comparing user answers with those provided by “A-Google-A-Day” along with the questions.

To provide a complete view of performance using response precision, we also compared the three conditions in terms of correct answers, wrong answers, and the total number of answers. An exploration of the three variables showed that all three had a normal distribution according to the Shapiro-Wilk test. Additionally, the variance in the three experimental conditions was found to be homogeneous according to the Levene test. As a result, between-group comparisons were performed with one-way ANOVA. Results for

Table IX.
Comparison of Tasks 1 and 2

Measures	C1	C2	C3
Effectiveness	p-value = 0.01611* Task2 > Task1	p-value = 0.01367* Task2 > Task1	p-value = 0.002441* Task2 > Task1
Efficiency	p-value = 0.02686* Task2 > Task1	p-value = 0.1055 Not significant	p-value = 0.3013 Not significant
Relevant coverage	p-value = 0.0004883** Task2 < Task1	p-value = 0.001953** Task2 < Task1	p-value = 0.0004883** Task2 < Task1
Query diversity	p-value = 0.02113* Task2 < Task1	p-value = 0.0156* Task2 < Task1	p-value = 0.02842* Task2 < Task1

Notes: *$p < 0.05$; **$p < 0.01$

this analysis showed no significant differences between the three conditions in terms of correct answers and total number of answers. However, significant differences were reported in terms of incorrect answers and response precision. Specifically, our results showed that triads (C3) made significantly fewer mistakes than individuals (C1) ($p < 0.05$). Consistently, triads (C3) were able to achieve significantly higher response precision than the individuals (C1) ($p < 0.05$). Differences between C1 and C2 were marginally significant ($p < 0.075$) with response precision being higher for C2 than C1. Finally, no significant differences were found between C2 and C3. Table X provides a summary of the results for each measure and the corresponding results.

5. Discussion

The analysis for Task 1, as previously reported (Shah *et al.*, 2015), showed that adding extra users was beneficial. However, this trend did not hold for Task 2 with respect to effectiveness and efficiency. Indeed, people working alone were the most effective and efficient. A possible explanation for these results is the nature of Task 2. The task was geared toward answering the greatest number of questions in the least possible amount of time and therefore did not encourage users to linger on a page searching for information. Traditionally, the minimum dwell time threshold for a page to be considered useful is 30 seconds, which may be quite high under time-constraints scenarios or when the goal is to find as many answers as possible. It is also plausible that such a threshold is high for two or three people working together to find a fact. Individuals, on the other hand, may not benefit from the help of their peers and may subsequently need to spend longer on any given page. This may explain the boost in effectiveness and efficiency from Tasks 1 to 2 as reported in the previous section.

Despite favorable results for single users in Task 2 in terms of effectiveness and efficiency; being effective and efficient – measures that are operationalized in terms number of queries and useful coverage – in this particular task do not necessarily translate into high performance. In particular, high effectiveness occurs as a result of a favorable proportion between useful coverage and total coverage, whereas efficiency is improved when effectiveness is high and the number of queries is reduced. For the case of dyads and triad in this particular type of task (multiple-step fact finding), these two measures are negatively affected due to the individual contributions of group members in terms of useful coverage and the total number of queries issued by groups. As presented in Tables VIII and X, while

Measures	C1	C2	C3	ANOVA
Correct answers	Mean: 6.416 Std: 3.752	Mean: 8.363 Std: 2.766	Mean: 9.083 Std: 2.712	F value = 4.467 $p < 0.05$ However, no significant differences were revealed through post-hoc tests, in particular Tukey HSD
Incorrect answers	Mean: 9.583 Std: 2.609	Mean: 7.363 Std: 4.056	Mean: 5.333 Std: 2.806	F value = 10.960 $p < 0.01$ Tukey HSD C3 < C1 ($p < 0.01$)
Number of answers	Mean: 16.000 Std: 4.000	Mean: 15.727 Std: 3.977	Mean: 14.416 Std: 3.287	F value = 1.091 $p = 0.304$
Response precision	Mean: 0.384 Std: 0.164	Mean: 0.547 Std: 0.193	Mean: 0.634 Std: 0.159	F value = 12.850 $p < 0.01$ Tukey HSD C3 > C1 ($p < 0.01$) C2 > C1 (marginal $p = 0.075$)

Table X. Response precision-based performance for Task 2

dyads and triads were not as efficient and effective as single users, they were able to better succeed in Task 2 in terms of response precision. One possible explanation for this outcome is the nature of Task 2, which unlike Task 1 was non-dividable and moreover each question in Task 2 had a unique answer. As a result, group members in dyads and triads were on the same track, which promoted within-group evaluation about the relevance of information to answer each question properly. Our results for response precision did not show significant differences between C2 and C3, and while differences were found in terms of means (with C3 greater than C2); it is hard to say that increasing group size would also improve response precision. Moreover, besides group size there are additional factors (both internal and external) in group configuration that could also affect performance (González-Ibáñez, Haseki and Shah, 2012a; González-Ibáñez and Shah, 2014a).

The lack of significance for relevant coverage and query diversity indicates that no one study condition can be said to be superior to another. Task 2 allowed the option of skipping questions and each group worked on a slightly different set of questions. Therefore, it is logical that study condition would not be a factor in relevant coverage, because a diverse range of search topics would limit page overlap across groups. Additionally, fact-finding type questions would not encourage users to collect snippets. As for query diversity, each group member was working toward answering the same question. Because every group was searching for the same concrete answer, the diversity of the queries would be limited to a specific range directly related to the question. The analysis has shown that adding more people to a group does not present any specific advantage.

In comparison to Task 1, individuals were more effective in Task 2. The results for dyads and triads were not statistically significant. Given that Task 2 was more focused and the information more readily available, users would have needed to visit fewer pages/sources to complete the task. The effect of lower coverage would boost effectiveness. One possible reason why there was no statistical significance between Tasks 1 and 2 for Conditions 2 and 3 is that dyads and triads had very low numbers for useful pages, negating the effects of lower coverage. Effectiveness, however, is higher in Task 2 across all study conditions. One explanation is that with all users searching for the same information, the number of distinct queries made would be much lower, raising effectiveness. Interestingly, Task 1 was superior in terms of relevant coverage. Even though all the users in Task 2 were pulling from the same list of questions, the option to skip meant not all groups answered every question, meaning they were not visiting the same webpages. Additionally, users from Task 1 would be collecting more snippets, since the exploratory search task would require more information to generate a more complex answer. The Kruskal-Wallis test was significant for Condition 1 in query diversity, but no other condition was statistically significant. Additionally, the pairwise testing was not significant across any study condition. Further testing is needed to determine why query diversity does not seem to be affected by task type.

Finally, it is important to note that Task 2 is also linear with time, which means the more time one spends on it, the more questions one could answer and the more one answers, the more correct answers one could provide. Therefore, while comparing C2 to C1, we should technically half C2's numbers in Table X. For C3, we should divide them by 3. This linearity is not the case with Task 1. Spending more time does not guarantee better effectiveness or efficiency. This is shown in earlier works (e.g. Pickens *et al.*, 2008; Hendahewa and Shah, 2015). In effect, adding more people for Task 2 does not really help, even for response precision related measures if we really consider the nature of Task 2.

Fact finding could be more challenging task than it may seem, at least for the experiments reported here. For instance, it is quite likely that when the participants come across the same answer repeatedly for related queries they execute, they might consider it as the correct answer although it might not be since it does not come from reliable sources. Given that the

"A-Google-A-Day" questions used in the study were at level-3 difficulty level (ideal case requiring three steps to get the correct answer), the participants had to be very critical in finding the correct answer without being misled by frequent or obvious search results since the answers were not straightforward but one has to go from one step to the other to arrive at the final answer by issuing the most applicable queries. Such phenomena most likely made some influence in the dyads and triads when deciding the queries to execute at each step, making it difficult for them to come to a consensus as to what was the correct answer.

6. Conclusion

In the work reported here, we attempted to provide a follow-up on a previously published article that addressed the issue of assessing the effect of the number of collaborators in CIS. Specifically, in the earlier work it was shown that as the number of project workers go from one to two and then three, an exploratory search task benefits. This study focused on a different kind of information seeking task (fact finding), and found that individuals performed better than the dyads or triads in some respects, or at least there were hardly any advantages to adding more people to the task when focusing on effectiveness and efficiency. Yet, in fact-finding search tasks effectiveness and efficiency (as operationalized in this study) do not necessarily translate into high performance. When focusing on actual performance in this particular task type, our results indicate that increasing group size had a positive effect in response precision. On the other hand, considering the time effect, this performance boost did not keep up with the increased group size.

The cumulative findings from the past and the current paper provide us with interesting insights about when collaboration could be useful. They show that a clearly divisible task (Task 1) could lead to improved performance with the help of added project members, but the same is not true for a non-divisible task (Task 2).

While heuristics are often presented in the literature (e.g. London, 2012), this is the first clear empirical work to demonstrate where that line could be drawn in an information seeking situation. Combining the findings from these two experiments, we could also extend the implications to decision-making about the structure of collaboration. For instance, by doing task and domain analyses, we could identify parts of a project that are multi-faceted and those that are streamlined singletons. We could then create an appropriate collaborative condition (e.g. dyads, triads) for the former and let the individuals work independently on the latter.

Notes

1. www.agoogleaday.com
2. www.google.com/intl/es419/insidesearch/features/search/knowledge.html

References

Aula, A., Khan, R.M. and Guan, Z. (2010), "How does search behavior change as search becomes more difficult?", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY*, pp. 35-44.

Bilal, D. (2000), "Children's use of the yahooligans! Web search engine: I. Cognitive, physical, and affective behaviors on fact-based search tasks", *Journal of the American Society for Information Science*, Vol. 51 No. 7, pp. 646-665.

Broder, A. (2002), "A taxonomy of web search", *SIGIR Forum*, Vol. 32 No. 2, pp. 3-10.

Fox, S., Karnawat, K., Mydland, M., Dumais, S. and White, T. (2005), "Evaluating implicit measures to improve web search", *ACM Transactions on Information Systems*, Vol. 23 No. 2, pp. 147-168.

González-Ibáñez, R. and Shah, C. (2011), "Coagmento: a system for supporting collaborative information seeking", *Proceedings of the American Society for Information Science and Technology*, Vol. 48 No. 1, pp. 1-4, available at: http://doi.wiley.com/10.1002/meet.2011.14504801336

González-Ibáñez, R. and Shah, C. (2012), "Investigating positive and negative affects in collaborative information seeking: a pilot study report", *Proceedings of the American Society for Information Science and Technology*, Vol. 49 No. 1, pp. 1-4.

González-Ibáñez, R. and Shah, C. (2014a), "Performance effects of positive and negative affective states in a collaborative information seeking task", in Baloian, N., Burstein, F., Ogata, H., Santoro, F. and Zurita, G. (Eds), *Collaboration and Technology*, Springer International Publishing, Santiago, CA, pp. 153-168.

González-Ibáñez, R. and Shah, C. (2014b), "Performance effects of positive and negative affective states in a collaborative information seeking task", *CYTED-RITOS International Workshop on Groupware, Springer International Publishing*, pp. 153-168.

González-Ibáñez, R., Haseki, M. and Shah, C. (2012a), "Time and space in collaborative information seeking: the clash of effectiveness and uniqueness", *Proceedings of the American Society for Information Science and Technology*, Vol. 49 No. 1, pp. 1-10, available at: http://dx.doi.org/10.1002/meet.14504901080

González-Ibáñez, R., Shah, C. and White, R.W. (2012b), "Pseudo-collaboration as a method to perform selective algorithmic mediation in collaborative IR systems", *Proceedings of the Association of Information Science & Technology (ASIST) Annual Meeting, Baltimore, MD, October 26-30.*

Hansen, P. and Järvelin, K. (2005), "Collaborative information retrieval in an information-intensive domain", *Information Processing and Management*, Vol. 41 No. 5, pp. 1101-1119.

Hendahewa, C. and Shah, C. (2015), "Implicit search feature based approach to assist users in exploratory search tasks", *Information Processing & Management*, Vol. 51 No. 5, pp. 643-661, available at: www.sciencedirect.com/science/article/pii/S0306457315000795

Jansen, B.J., Booth, D.L. and Spink, A. (2008), "Determining the informational, navigational, and transactional intent of web queries", *Information Processing & Management*, Vol. 44 No. 3, pp. 1251-1266.

Kellar, M., Watters, C. and Shepherd, M. (2007), "A field study characterizing web-based information-seeking tasks", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 7, pp. 999-1018.

Levenshtein, V.I. (1966), "Binary codes capable of correcting deletions, insertions and reversals", *Soviet Physics Doklady*, Vol. 10 No. 8, p. 707.

London, S. (2012), "Building collaborative communities", in Mortensen, M.B. and Nesbitt, J. (Eds), *On Collaboration*, Tate, pp. 75-83.

Marchionini, G. (2006), "Exploratory search: from finding to understanding", *Communications of the ACM*, Vol. 49 No. 4, pp. 41-46.

Morris, M.R. (2007), "Collaborating alone and together: investigating persistent and multi-user web search activities", available at: http://research.microsoft.com/apps/pubs/default.aspx?id=70402 (accessed November 21, 2016).

Olson, G.M., Olson, J.S., Carter, M.R. and Storrøsten, M. (1992), "Small group design meetings: an analysis of collaboration", *Human-Computer Interaction*, Vol. 7 No. 4, pp. 347-374.

Pickens, J., Golovchinsky, G., Shah, C., Qvarfordt, P. and Back, M. (2008), "Algorithmic mediation for collaborative exploratory search", *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR), Singapore*, pp. 315-322.

Reddy, M.C. and Jansen, B.J. (2008), "A model for understanding collaborative information behavior in context: a study of two healthcare teams", *Information Processing and Management*, Vol. 44 No. 1, pp. 256-273.

Rieh, S.Y., Collins-Thompson, K., Hansen, P. and Lee, H.J. (2016), "Towards searching as a learning process: a review of current perspectives and future directions", *Journal of Information Science*, Vol. 42 No. 1, pp. 19-34.

Shah, C. (2008), "Toward collaborative information seeking (CIS)", *Proceedings of Collaborative Exploratory Search Workshop at JCDL, Pittsburgh, PA, June 20*, available at: http://arxiv.org/abs/0908.0709

Shah, C. (2014a), "Collaborative information seeking", *Journal of the Association for Information Science and Technology*, Vol. 65 No. 2, pp. 215-236.

Shah, C. (2014b), "Evaluating collaborative information seeking – synthesis, suggestions, and structure", *Journal of Information Science*, Vol. 40 No. 4, pp. 460-475, available at: http://jis.sagepub.com/content/40/4/460.abstract

Shah, C., Hendahewa, C. and González-Ibáñez, R. (2015), "Two's company, but three's no crowd: evaluating exploratory web search for individuals and teams", *Aslib Journal of Information Management*, Vol. 67 No. 6, pp. 636-662.

Twidale, M.B.T., Nichols, D.M.N. and Paice, C.D. (1997), "Browsing is a collaborative process", *Information Processing and Management*, Vol. 33 No. 6, pp. 761-783.

White, R.W. and Drucker, S.M. (2007), "Investigating behavioral variability in web search", *Proceedings of the 16th International Conference on World Wide Web (WWW '07), ACM, New York, NY*, pp. 21-30, doi: http://dx.doi.org/10.1145/1242572.1242576.

White, R.W. and Huang, J. (2010), "Assessing the scenic route: measuring the value of search trails in web logs", *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10), ACM, New York, NY*, pp. 587-594, available at: http://dx.doi.org/10.1145/1835449.1835548

Corresponding author
Chirag Shah can be contacted at: chirags@rutgers.edu