

Questionnaire mode effects in interactive information retrieval experiments

Diane Kelly^{a,*}, David J. Harper^{b,1}, Brian Landau^a

^a School of Information and Library Science, University of North Carolina, Chapel Hill, NC 27599-3360, USA

^b School of Computing, Robert Gordon University, Aberdeen, Scotland, UK

Received 10 October 2006; received in revised form 31 January 2007; accepted 19 February 2007

Available online 5 April 2007

Abstract

The questionnaire is an important technique for gathering data from subjects during interactive information retrieval (IR) experiments. Research in survey methodology, public opinion polling and psychology has demonstrated a number of response biases and behaviors that subjects exhibit when responding to questionnaires. Furthermore, research in human–computer interaction has demonstrated that subjects tend to inflate their ratings of systems when completing usability questionnaires. In this study we investigate the relationship between questionnaire mode and subjects' responses to a usability questionnaire comprised of closed and open questions administered during an interactive IR experiment. Three questionnaire modes (pen-and-paper, electronic and interview) were explored with 51 subjects who used one of two information retrieval systems. Results showed that subjects' quantitative evaluations of systems were significantly lower in the interview mode than in the electronic mode. With respect to open questions, subjects in the interview mode used significantly more words than subjects in the pen-and-paper or electronic modes to communicate their responses, and communicated a significantly higher number of response units, even though the total number of unique response units was roughly the same across condition. Finally, results showed that subjects in the pen-and-paper mode were the most efficient in communicating their responses to open questions. These results suggest that researchers should use the interview mode to elicit responses to closed questions from subjects and either pen-and-paper or electronic modes to elicit responses to open questions.

© 2007 Published by Elsevier Ltd.

Keywords: Interactive information retrieval experiment; Research methods; Questionnaires; Mode effects; Social desirability responding

1. Introduction

Bulmer (2004) defines a questionnaire as, “any structured research instrument which is used to collect social research data in a face-to-face interview, self-completion survey, telephone interview or Web survey. It consists

* Corresponding author. Tel.: +1 919 962 8065.

E-mail addresses: dianek@email.unc.edu (D. Kelly), d.harper@rgu.ac.uk (D.J. Harper), blandau@email.unc.edu (B. Landau).

¹ Tel.: +44 (0) 1224 262706.

of a series of questions set out in a schedule, which may be on a form, on an interview schedule on paper, or on a Web page” (p. XIV). Questionnaires in interactive information retrieval (IR) studies are typically self-administered via electronic or pen-and-paper mode, or via interview, and have been used to elicit a variety of types of information from system users including factual, behavioral and attitudinal.

Questionnaires can be comprised of closed questions, open questions or a mixture of both. Closed questions are questions that provide a fixed set of responses with which subjects must respond. It is common practice for usability questionnaires to include closed questions in the form of statements such as, *the system was easy to learn to use*. Subjects are typically provided with 5–7-point Likert-type scales for responding, where one scale end-point represents strong agreement and the other represents strong disagreement. The use of semantic differentials is another common way to respond to closed questions. Open questions, on the other hand, do not provide a response set and subjects are able to provide any type of response they feel is appropriate. For instance, open questions included on usability questionnaires in interactive IR experiments often identify a particular feature of the system and ask subjects for their impressions of the feature. It is also common to have open questions that ask subjects to identify the most positive and negative things about the system. Each type of question has its own set of merits and demerits, which is why a combination of question types is often used.

The questionnaire is a vital part of interactive IR studies since it is one of the primary vehicles for eliciting data from subjects. However, it is generally believed that subjects have a tendency to inflate their ratings of systems during usability evaluations (Czerwinski, Horvitz, & Cutrell, 2001; Nielsen & Levy, 1994). For instance, Nielsen and Levy (1994) conducted a meta-analysis of 57 human–computer interaction (HCI) studies with the goal of analyzing the relationship between performance and preference. Nielsen and Levy (1994) demonstrated that the performance of 101 of the 127 systems studied was rated higher than the neutral point of the scale used to make the ratings, and that users assigned a majority of systems a score of four on a 5-point scale. This meta-analysis provides some evidence for the notion that subjects inflate their ratings of systems or that usability questionnaires are not particularly sensitive, which makes obtaining valid data from subjects a challenge. Moreover, in experiments with two or more systems, this may create problems since there is a good chance that the systems will be rated the same by subjects. Indeed, it is conjectured that subjects in most studies of interactive IR rarely rate systems poorly even when they clearly violate basic usability principles. Since researchers are interested in obtaining as valid data as possible from subjects, identifying better ways of collecting usability data from subjects is important. Doing so will further allow researchers to better understand interactive IR processes and differences in systems’ support for such processes.

There have been a number of studies about interactive IR evaluation, including those presenting general frameworks and guidelines (cf., Borlund, 2003; Dumais & Belkin, 2005; Tague-Sutcliffe, 1992; Thomas & Hawking, 2006; Toms, Freund, & Li, 2004). However, studies investigating the impact of various choices of experimental components on study results are rare, with few exceptions (cf., Borlund, 2000). In contrast with other disciplines, where studies of methods and experimental design comprise an important portion of the literature, the impact of different experimental designs decisions, such as batch-mode vs. individual session, interview vs. electronic questionnaire, or open vs. closed questions, is virtually ignored in the context of interactive IR experimentation. In this study we explore the extent to which questionnaire mode impacts subjects’ usability ratings of two systems and their responses to open questions.

2. Literature review

There is a great deal of literature in psychology, public opinion polling and public health which has investigated the impact of mode effects and question type on subjects’ responses to questionnaires (see Groves, 1989; Richman, Kiesler, Weisband, & Drasgow, 1999; Tourangeau, Rips, & Rasinski, 2000 for reviews). Mode effects occur as a result of using a particular technique for collecting questionnaire data. Studies of mode effects have compared and contrasted responses elicited via a number of different questionnaire modes, including pen-and-paper, electronic, telephone and face-to-face interviews. These studies have demonstrated very complex (and often varying) relationships between questionnaire mode, question content, question type, and subjects’ responses. However, one of the most common results is that subjects are more willing to report sensitive information, and socially disapproved or illegal behaviors in self-administered questionnaires than in questionnaires administered via interviews (Tourangeau et al., 2000). One of the most popular explanations of

why differences have been observed between modes is social desirability responding, or “the tendency by respondents, under some circumstances and modes of administration, to answer questions in a more socially desirable direction that they would under other conditions or modes of administration” (Richman et al., 1999, p. 755).

2.1. *Social desirability responding*

Social desirability responding has been conceptualized as topic- and personality-based (Brewer, Hallman, Fiedler, & Kipen, 2004). Topic-based conceptualizations claim that the topic or content of particular questions prevent subjects from responding truthfully. For instance, questions about behaviors like drug use will be underreported, while behaviors like helping others or voting will be over-reported. Personality-based conceptualizations “link socially desirable responding to a personality trait characterized by lower self-reports of unflattering behaviors” (Brewer et al., 2004, p. 876). In such explanations it is not the topic of the question that inhibits or encourages socially acceptable responses, but an inherent trait that causes people to portray themselves in the most flattering way.

Results of research about social desirability responding and questionnaire mode have been mixed, although it is generally accepted that different modes are associated with different levels of social desirability responding, and specifically that social desirability responding will occur most often in the interview mode. Despite this, interviews are still used in many situations because the benefits (interviews are generally thought to be more motivating, reduce non-response and encourage longer, more elaborated responses) out-weigh the potential for response bias (Richman et al., 1999).

Researchers have noted that self-administered questionnaires increase subjects’ willingness to disclose information about sensitive topics (Tourangeau et al., 2000). Initial studies compared differences between traditional modes: pen-and-paper, telephone interview and face-to-face interview. Once computer technology became established as a technique for collecting data, research also compared electronic modes with more traditional modes. Some have found less social desirability responding with electronic questionnaires (Martin & Nagao, 1989; Weisband & Kiesler, 1996), while others have found more social desirability responding with electronic questionnaires (Lautenschlager & Flaherty, 1990) and still others have found no differences between modes (Booth-Kewley, Edwards, & Rosenfeld, 1992). Kiesler and Sproull (2001) found that subjects’ responses to closed questions on an electronic questionnaire were less socially desirable than responses on a pen-and-paper version of the same questionnaire. These researchers also found that responses to open questions were relatively longer and more disclosing in the electronic mode. A meta-analysis of a large number of studies about social desirability responding and questionnaires demonstrated that there were no differences between electronic and pen-and-paper modes, but that there were differences between electronic and interview modes (Richman et al., 1999). Clearly, the relationship between questionnaire mode and response behaviors is quite complex.

One explanation for why there is less social desirability responding in electronic questionnaires is related to anonymity (Richman et al., 1999). Some researchers claim that electronic questionnaires provide greater levels of anonymity to subjects than pen-and-paper or interview questionnaires. Once a subject submits his responses to an electronic questionnaire there is no physical instantiation of these responses – essentially answers “disappear into the computer” (Richman et al., 1999, p. 756). However, people’s perception of computers have likely changed since many of the original studies were conducted and in more recent studies researchers have suggested that instead of heightened anonymity, questionnaires administered electronically might invoke the “big brother” feeling. This, in turn, might increase the occurrence of social desirability responding (Rosenfeld, Booth-Kewley, Edwards, & Thomas, 1996).

Many previous studies of social desirability responding and mode effects have been undertaken in the context of public opinion polling and healthcare, where subjects are often responding to questions about their lifestyle choices; stakes for responding “desirably” might necessarily be higher than in situations where subjects are asked to report their attitudes about a computer system. However, findings from such studies provide one possible explanation for why subjects tend to inflate their ratings of systems in interactive IR experiments. Subjects may view the success or failure of a system as a reflection of their own abilities rather than as a reflection of the system’s abilities. For instance, subjects might believe that responding negatively to questions such

as how easy a system was to learn to use, how easy a system was to use, or how satisfied they were with their performances, reflects on them rather than the system. When people are part of the process as they are in interactive IR, they may feel that they are at least in part responsible for a system's performance. Thus, people may view negative ratings of systems as negative ratings of themselves, and avoid using such ratings.

2.2. Acquiescence

Another factor that is relevant to understanding subjects' tendencies to inflate their ratings of systems is the notion of acquiescence. "Acquiescence, or agreeing-response bias, refers to a presumed tendency for respondents to agree with attitude statements presented to them" (Schuman & Presser, 2004, p. 203). It is important to note that acquiescence is a *presumed* tendency; many studies have remarked how difficult it is to clearly demonstrate that such a bias exists, although it seems like a reasonable explanation for why subjects tend to inflate their ratings of systems. Schuman and Presser (2004/1996) indicate that psychologists tend to view acquiescence as a personality trait, while sociologists and survey investigators tend to characterize acquiescence as a form of deference or as the result of a respondent's ignorance about a particular topic. Although it is often suggested that reversing the direction of items can correct acquiescence, changing the wording of items can change their meaning and in many cases this does not address a larger agreement issue: subjects are agreeing not so much with the statement, but with the experiment. This phenomenon is known as a demand effect.

2.3. Demand effects

Demand effects occur when subjects have an expectation of how they should behave in a particular research setting. These effects are described as *demand* effects because subjects may perceive that there is a demand for them to behave in a particular way. The source of the demand can be the researcher who behaves differently when subjects are evaluating the researcher's system of choice, placing a demand on subjects to react a particular way to one system and not another. The source of the demand can also be the subject's interpretation of the experimental situation, in which case the subject's behavior is contingent on him interpreting what desired effects are in any given experimental situation. In the context of interactive IR experiments, it seems reasonable for subjects to interpret that desired effects are likely to translate into positive system ratings, which might suggest why subjects tend to inflate their ratings of systems. In addition, subjects may not want to offend the researcher by rating a system poorly.

2.4. Cognitive and physical effort

An alternative perspective on differences observed due to questionnaire mode concerns the varying amounts of cognitive and physical effort required of subjects to complete questionnaires (Schwarz, Strack, Hippler, & Bishop, 1991; Tourangeau, 1984). Researchers generally agree that answering a question requires subjects to perform several tasks: interpret the question and understand its meaning; generate an opinion or reflect on past behaviors, which typically consist of retrieving information from memory; and communicate and possibly edit responses. Questionnaire mode most often impacts efforts related to interpreting the question, understanding its meaning and communicating and editing responses.

Cognitive efforts related to interpreting and understanding questions are similar in pen-and-paper and electronic modes since the question (and often response set) are visible to subjects. However, cognitive effort is high in the interview mode since this information is usually read to the subject and he must keep it in memory while interpreting and forming a response. Different levels of effort are also required for communicating responses. To communicate responses via pen-and-paper questionnaire a subject writes, via electronic questionnaire types, and via interview questionnaire speaks. While the effort required to communicate responses in each of these modes is ultimately related to the subject's abilities to write, type or speak (or read for that matter), in general, one can assume that writing requires the most effort and speaking the least.

Editing one's response requires both cognitive and physical effort, since it involves aspects of response generation and communication. The amounts of cognitive effort in the pen-and-paper and electronic modes are

somewhat similar. Pen-and-paper mode requires slightly more physical effort than the electronic since a person has to add/remove material with a pen or pencil. The cognitive demands of editing one's response in the interview mode are high, while the physical are still relatively low. A person must remember his previous response, identify what parts he wants to edit, and then manage to communicate both the location of the edit as well as the actual content of the edit.

In the context of responding to questionnaires in interactive IR experiments, subjects may have a more difficult time answering questions via interview mode because of cognitive demands. This effect is likely to be particularly marked for open questions, which already place high cognitive demands on the subject when formulating responses. Subjects may provide lengthier responses to open questions in the interview mode because this mode requires the least physical effort to communicate responses (Groves, 1978). However, these responses may not necessarily be of a higher quality, given the higher cognitive demands of this mode.

2.5. *Open and closed questions*

In constructing questionnaires, another major decision that must be made concerns question format. Common question formats include open and closed, and each has its own set of merits and demerits. Payne (2004/1951) identifies the merits of open questions and states that, "the free-answer is uninfluenced, it elicits a wide variety of responses, it makes a good introduction to a subject, it provides background for interpreting answers to other questions. It can be used to solicit suggestions, to obtain elaborations, [and] to elicit reasons" (p. 143). In addition, open questions allow subjects to report their own feelings, beliefs and impressions without being encumbered by a response set. Drawbacks to open questions are that they take longer to administer and responses are often difficult to interpret and analyze. People typically use different words to describe the same thing and some subjects are better at clarifying and explaining their responses than others. Since different subjects are likely to contribute different amounts of feedback, there is a danger that the opinions of a small number of subjects will dominate and bias results. In relation to questionnaire mode, subjects in a pen-and-paper or electronic questionnaire might skip open questions, only provide partial responses or provide responses which do not make sense. When open questions are administered in the interview mode, interviewers can ask for elaborations, explanations, and clarifications, and can probe subjects in a number of other ways. However, when administered during an interview, the impact of the interviewer on the interaction is more salient (Groves, 1978).

Krosnick (1991) and Brewer et al. (2004) provide some evidence that suggests that subjects often engage in a type of satisficing (Simon, 1957) in order to reduce the cognitive burden placed on them when responding to open questions. Brewer et al. (2004) describe respondents' satisficing response behavior: "Because formulating a full and complete answer may be too effortful, they may offer a compromise response that they think is 'good enough'" (p. 876). Krosnick (1991) distinguishes between two types of satisficing, weak and strong. In the case of weak satisficing, subjects' responses only reflect a small portion of what they know because to search their entire memories for exhaustive responses would be too effortful. Often subjects truncate this search process as soon as enough information has come to mind to form a reasonable judgment (Krosnick, 1991).

In the case of strong satisficing, subjects' responses reflect little or nothing about what they actually know and believe. One tactic identified by Brewer et al. (2004) is to answer the first in a series of related questions truthfully and then repeat that response (or variations of) in response to remaining questions. In the context of interactive IR, subjects who are fatigued at the end of a long experiment or those who try quickly to complete a questionnaire might engage in such tactics. For instance, a subject might identify one feature of the system or one usage instance and stay focused on this feature or instance when responding to questions, rather than considering his entire search experience and all features of the system. It is unclear how mode might be related to satisficing behavior in face-to-face interviews. For instance, the interview mode might help mitigate satisficing because subjects may feel more accountable for producing unique responses. On the other hand, more satisficing might be observed in the interview mode because the pace of the interview may make subjects feel more pressure to respond quickly. With this added pressure, subjects may not take the time to produce measured responses.

Closed questions allow researchers to gather a larger amount of data in less time using standard questions and scales for eliciting responses. Data elicited from closed questions are usually easier to analyze

and interpret. Although it is possible for subjects to skip one or more closed questions, in general, the amount and form of data elicited from subjects via closed questions is much more homogenous than that elicited via open questions. Closed questions also have several limitations. As noted previously, questions can be subject to response biases such as acquiescence. Moreover, scales are not based on a true number line and scale values are subject to individual interpretation; one subject's 6 may be another subject's 5. Response sets provided for closed questions do not always capture the extent of a person's opinions and researchers can introduce many biases when they construct scales and scale labels. With respect to interactive IR, data elicited via closed questions are often skewed toward positive ratings with little variance. Researchers often find themselves comparing system ratings that differ by a single point or by tenths of points. Thus, it is imperative that researchers began to identify better ways of designing and administering closed questions.

2.6. Purpose of study

The literature reviewed above presents several factors that have been proposed to impact subjects' response behaviors to questionnaires. In this study we investigate some of these factors in the context of an interactive IR experiment. Specifically, we investigate the relationship between questionnaire mode and subjects' responses to a usability questionnaire comprised of closed and open questions administered during an interactive IR experiment. The questionnaire modes we investigate are pen-and-paper, electronic and interview. At least one study has been conducted in the area of library science comparing mode effects in the context of a survey administered to academic reference librarians (Hayslett & Wildemuth, 2004). This study compared pen-and-paper and electronic versions of a survey and focused primarily on response rates and sampling bias. Aside from the studies cited earlier (Rosenfeld et al., 1996; Weisband & Kiesler, 1996), we were unable to find any studies of questionnaire mode effects in the HCI literature. To our knowledge, no study of questionnaire mode has been conducted in the context of interactive IR.

This study investigates the impact of social desirability responding and demand effects on subjects' responses to closed questions, which were designed to elicit numeric ratings of two systems. Our hypothesis is that subjects will rate a system more positively in the interview mode than in the pen-and-paper and electronic modes. Social desirability theory predicts that subjects will want to present themselves as favorably as possible. Research examining social desirability responding indicates that the greatest potential for this to occur will be in the interview mode.

We also wanted to investigate if the proximity of the researcher to the subject would create greater demand effects, which would also lead to more positive system ratings. The interview mode produces the greatest demand effect in this regard since the researcher is sitting next to the subject, engaged in a pseudo-conversation and making eye contact with the subject. We thus believed that this might also contribute to more positive ratings of the system from subjects in the interview mode vs. the pen-and-paper and electronic modes. It should be noted that we consider the source of the demand in this situation the subject's interpretations of the situation and not the researcher's behaviors.

Unfortunately, it is not possible to separate social desirability responding and demand effects in this instance, since they both offer plausible explanations for why differences may occur and previous research predicts that these differences will be in the same direction. Although our findings will not allow us to state definitely which of these variables had the greatest impact on the results, our findings can potentially allow us to say something about the relationship between questionnaire mode and subjects' responses.

This study also investigates the relationship between questionnaire mode and subjects' responses to open questions. Our hypothesis is that subjects will provide the shortest and least informative responses in the pen-and-paper mode because it requires more physical effort for subjects to complete. We hypothesize that subjects will provide the lengthiest and most informative responses in the interview mode, since this mode requires the least physical effort for subjects, there is a greater motivation for subjects to respond and since interviews are generally believed to elicit better responses to open questions. Although the higher cognitive demands in this mode may act to reduce the informativeness of responses, we believe the decreased physical demands will exert a stronger influence on behavior.

3. Method

The purpose of this study was to investigate the effects of questionnaire mode on subjects' evaluations of retrieval systems. This investigation was conducted within the context of an interface design study whose purpose was to investigate the effectiveness of search term highlighting on retrieval performance. The investigation of mode effects was integral to the study, and the experiment was purposefully designed to explore both mode effects and interface effects. We chose to use an experimental IR system, rather than a baseline system (such as Google) because we felt that this would give subjects more things to discuss during the questionnaires. If we presented a system with which subjects were already familiar, it is unlikely that they would have much to say.

Specific details of the interface investigation are not reported in this paper, but they are mentioned since they are necessary to understand the study design. The study was a 2×3 between-subjects factorial design; that is, each subject used one interface and experienced one questionnaire mode. Search term highlighting had two levels (highlighting or no highlighting) and questionnaire mode had three (pen-and-paper, electronic and interview). We created a balanced design where questionnaire mode was distributed equally across highlighting levels.

Two researchers conducted this study and several steps were taken to minimize researcher effects. First, we systematically rotated researchers across conditions so that each researcher administered an equal number of cases with each mode–interface combination. Second, a detailed experimental protocol was developed and used during the study which prescribed how the experiment would be administered (including scripts for interacting with subjects). Finally, during analysis, we included researcher as a variable to determine if there were significant differences in subjects' responses based on researcher.

3.1. Participants

An email was sent to the entire undergraduate population at the University of North Carolina to solicit subjects for this study. We accepted the first 48 people who responded to the email and accepted additional students as alternates. Subjects were assigned randomly to condition and compensated US\$20.00 for participation. In total, 52 subjects completed the experiment, but only 51 are used in analysis (17 per mode). One subject was considered an outlier and excluded from the study because this subject's age was over seven standard deviations above the mean. The mean and standard deviation for age including the outlier was 21.21 and 4.19; excluding the outlier the figures were 20.67 and 1.47. Since the subject differed so substantially from the rest of the subjects on this variable, we felt her exclusion would provide us with a more cohesive sample.

Sixty-seven percent of the subjects were female and 33% were male. Ninety percent of the subjects were undergraduates, while 4% were graduates and 6% were 'other' (e.g., continuing education students). Thirty percent of the subjects were social science majors, 24% were humanities majors, 22% were science majors, 22% were part of the professional schools (e.g., business) and 2% did not respond to this question. The means and standard deviations for subjects' search experiences and frequencies of Web searching were as follows: 3.04 (.49) and 3.73 (.53). For search experience, "1" indicated very inexperienced and "4" indicated very experienced. For frequency of Web searching, "1" indicated less than monthly and "4" indicated daily. As expected, this group is fairly experienced with respect to searching for information and searches the Web with a great deal of frequency.

3.2. Interface and system

In this study, subjects used the conteXtual relevance feedback system (XRF) to find and save documents for four search topics (Harper & Kelly, 2006). The XRF interface allows users to save documents in piles and perform relevance feedback using documents contained within these piles. Fig. 1 shows the XRF interface for a search on the topic "tropical storms, loss of life and damage," where the user is sorting documents according to the name of each identified storm. A query is entered by selecting the search pile (1), search results are displayed in (2), and the full texts of documents are displayed in (3).

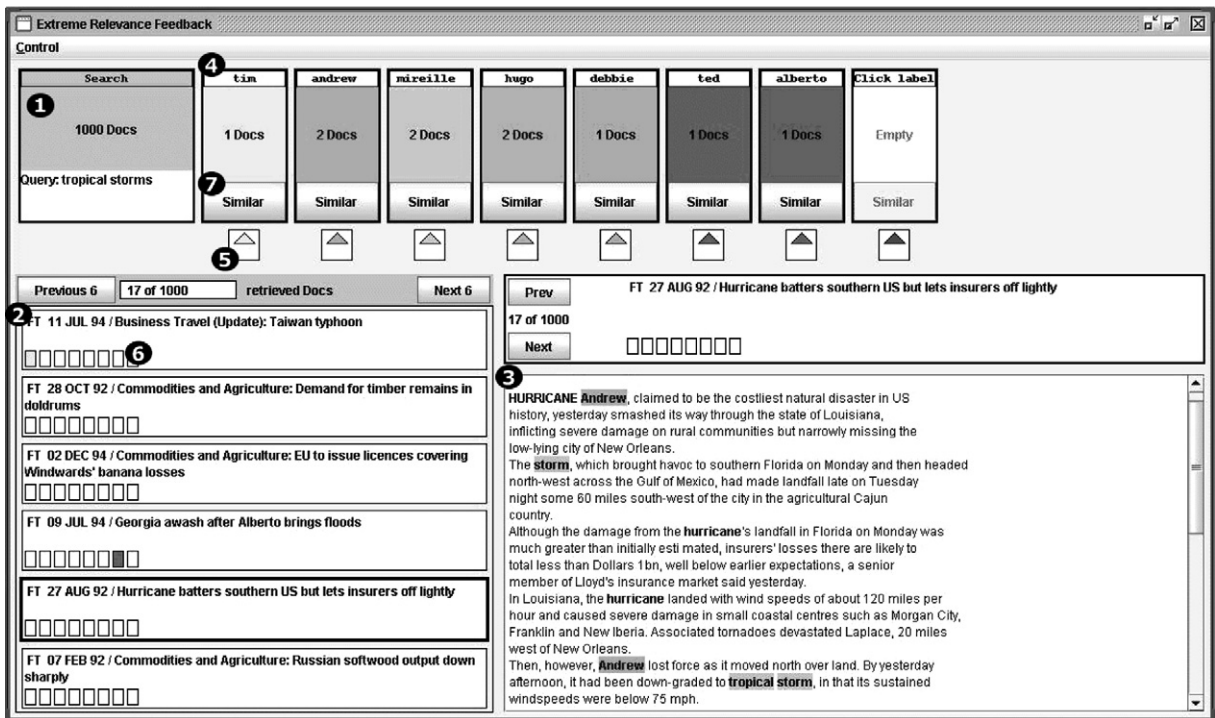


Fig. 1. Screen shot of the XRF interface with highlighting.

XRF enables a user to save documents in one or more piles (4) and create labels for these piles. Piles have preset colors to help with identification. A document can be moved into and out of a pile by using the arrow under each pile (5). Small pile icons below the title in each result show piles in which a document is saved (6). The contents of a pile can be reviewed by clicking on a pile. Relevance feedback is invoked by clicking the *Similar* button on a pile (7), where the relevance feedback (RF) process assumes that all documents in a pile are relevant. Support for highlighting varied in this study, with one system providing no highlighting and another providing some. In each opened document, the ‘highlighting’ system highlighted subjects’ current query terms (grey highlight), previous query terms (bold face), and terms used by subjects as pile labels (pile color highlight). Interested readers are referred to Harper and Kelly (2006) for more information about the XRF system; we note that the interface described therein did not support highlighting.

3.3. Collections and tasks

The TREC-8 Interactive Track collection, consisting of a corpus of 210,158 articles from the Financial Times of London 1991–1994, a set of aspectual search topics, and a set of relevance judgments was used in

Number: 408; **Title:** tropical storms

Description: What tropical storms (hurricanes and typhoons) have caused property damage and/or loss of life?

Instances: In the time allotted, please find as many DIFFERENT storms of the sort described above as you can. Please save at least one document for EACH such DIFFERENT storm. If one document discusses several such storms, then you need not save other documents that repeat those, since your goal is to identify as many DIFFERENT storms of the sort described above as possible.

Fig. 2. Example TREC-8 Interactive Track topic.

Number: 408; **Title:** tropical storms

Imagine that you are enrolled in an environmental sciences course and you are interested in learning more about *tropical storms (hurricanes and typhoons)*. It seems that tropical storms are becoming more destructive and you decide to investigate past storms. Your instructor asks you to prepare a short, 5-page paper investigating past tropical storms. Your instructor asks you to use historical newspaper articles as your sources of information and to collect comprehensive information on different tropical storms and impacts of each storm. Specifically, your goal is to *identify different tropical storms that have caused property damage and/or loss of life*, and to find as much information as possible about each storm.

Fig. 3. New version of TREC-8 Interactive Track topic.

this study (Hersh & Over, 1999). An example aspectual recall tasks entitled ‘tropical storms’ is displayed in Fig. 2.

We wanted subjects to complete tasks that required them to find comprehensive and exhaustive information, so we modified the six TREC aspectual topics. The new topics asked subjects to find documents covering as many aspects as possible (comprehensiveness) and also as many documents as possible relating to each aspect (exhaustiveness). We further added text describing an information seeking scenario by combining the description and instance fields. The modified version of the tropical storms topic is displayed in Fig. 3. The original text from the TREC-8 Interactive Track topic is italicized. We used five of six available topics in this study: one topic was used in video demonstrations of each system and four were used by subjects. A Latin-square was used to rotate topics.

3.4. Procedures

The study took approximately 1.5 h to complete. When subjects arrived to the laboratory, they completed a consent form and a background questionnaire. Following this, subjects were presented with a video demonstration of the system that they would use with an example search scenario. Next, subjects were presented with their first search scenario. Subjects had up to 15 min to search for documents related to each scenario. After subjects completed one scenario, they were asked to complete a Post-Search Questionnaire, which asked them to make estimates of a number of things, such as the amount of time they spent searching for documents related to their first query, and the number of documents they saved. Data from these questionnaires are not reported in this paper; however, the mode of delivery for these questionnaires matched the mode used in the Exit Questionnaires. In total, subjects completed four search topics and four Post-Search Questionnaires. At the end of the study, subjects completed the Exit Questionnaire. The Exit Questionnaire is the focus of this study.

3.5. Exit questionnaire

The content of the Exit Questionnaires was identical across modes. The first 21 questions were closed questions designed to assess the usability of the systems. These were not questions per se, but statements where subjects indicated the strength of their agreement. All 21 of these questions were assessed with 7-point scales, where 1 = strongly disagree and 7 = strongly agree. Circles, or radio buttons, corresponded to each numeric value on the scale and subjects marked the value that represented their beliefs.

A set of eight general usability questions were taken from the USE Questionnaire (Lund, 2001), while the remainder of the questions ($n = 13$) were developed to evaluate features specific to the XRF systems. The content of all questions and the order in which they were asked are displayed in Table 3. These questions were designed to evaluate the following aspects of usability, with the number(s) corresponding to the question(s) assessing each aspect in parenthesis: ease of learning (2), ease of use (3–8), usefulness (9–10), effectiveness

(11–19) and satisfaction (20–21). We included one additional question to assess the similarity of the search methods used by subjects in this study to those they normally use when searching the Web.

The following four open questions were included as questions 22–25 of the Exit Questionnaire: (a) What were the most positive things about using this system and why? (b) What were the most negative things about using this system and why? (c) How would you improve this system and why? (d) Is there anything else that you would like to tell us about this system and your experiences using it? These general questions are typical of those asked during interactive information retrieval studies.

In the pen-and-paper mode, subjects were provided with a print version of the questionnaire, which was produced using a word processing program. In the electronic mode, subjects were provided with an electronic version of the questionnaire, which was produced using XHTML and CSS. Subjects viewed the electronic questionnaire with the Firefox Web browser.

We used separate programs to produce the pen-and-paper and electronic questionnaires because we did not want the pen-and-paper questionnaire to look like a printed Web document. The reason for this was that we did not want subjects to even consider the questionnaire mode – if subjects were given a questionnaire that had clearly been printed from the Web, then they might wonder why they were not completing the questionnaire on the Web. For each questionnaire, we used identical layout, font and design, including identically sized visible areas for responses. Additional space for responding to open questions in the pen-and-paper mode was found on the back of the questionnaire (and in the margins), while the text areas in the electronic mode were set to wrap and to have no limit on the number of possible rows (i.e., all or part of the response was visible to subjects at all time and subjects could type as much as they liked).

For the interview mode, a protocol was developed which prescribed how interviews were to be conducted, including instructions for asking follow-up questions. It should be noted that scripting interviews in this manner necessarily means that the interviewer loses some flexibility, which of course, is one of the benefits of conducting interviews. However, we wanted to minimize interviewer effects as much as possible, especially since one interviewer was more experienced than the other. In general, follow-up questions were only asked if the subject did not answer the entire question or if the researcher did not understand the subject's response. Interviewers were instructed to maintain a neutral facial expression and to nod if appropriate. For closed questions, interviewers presented subjects with a scale for responding (to reduce cognitive effort), read each statement to subjects, and asked subjects to simultaneously point to and say aloud the number on the scale that represented their opinions. The interviewer recorded subjects' responses on a printed version of the questionnaire. It is obviously impossible to make all interviews identical, even in cases when there is a single interviewer. However, large-scale survey efforts employing multiple interviewers (cf., US Census) have demonstrated that through training, differences can be minimized. Interviews were also recorded and transcribed later for analysis.

3.6. *Analysis of responses*

The analysis of quantitative data from the closed questions was somewhat straight-forward – numeric scores were entered into SPSS and analyzed. However, the analysis of data from the open questions was a bit more difficult. We initiated a content analysis of responses to the open questions to divide subjects' responses into units that could be analyzed, a process known as “unitizing.” Neuendorf (2002) describes the unit in content analysis as, “an identifiable message or message component, (a) which serves as the basis for identifying the population and drawing a sample, (b) on which variables are measured, or (c) which serves as the basis for reporting analyses. Units can be words, characters, themes, time periods, interactions, or any other result of ‘breaking up a ‘communication’ into bits’ (Carney, 1971, p. 52).” (p. 71). “Unitizing” is the process of separating a stream of actions or words into discrete units. Unitizing spoken words, particularly those created in a conversational setting such as an interview, are notoriously difficult because people usually do not speak in complete sentences and can often have many fits and false starts. Neuendorf (2002) cites evidence (Newton, Engquist, & Bois, 1977) that demonstrates that while people are generally good at experiencing a stream of words as coherent units auditorily, attempts to instruct coders to unitize such speech acts formally are often met with failure, especially when articulating unitization rules.

To unitize subjects' responses to open questions, two of the researchers coded responses independently with the goal of subdividing responses into the smallest possible unit. This usually meant subdividing based on

Table 1
Example of how responses from two subjects were unitized

Question	Subject 13	Subject 34
Positive	<i>“Being able to put documents into piles is very helpful. It helps organize the info. you find [in] your search. I like the color-coding and the highlighting of words – very helpful.”</i> [4]	<i>“The color coding for the various saved terms because it made it easier to locate items that were similar to what you were looking for.”</i> [1]
Negative	<i>“Didn’t always return results that were relevant.”</i> [1]	<i>“It often turned up items that were not relevant to my search query.”</i> [1]
Improve	<i>“Make is possible to search for tourism AND Italy, for example, so it would return only documents with both words, not articles about Italy or just tourism.”</i> [1]	<i>“When I clicked on similar I would only allow items that contain that text to come up in the results box because it is misleading to have all those results come up if you only the first three have to deal with your topic.”</i> [1]
Anything else	<i>“No.”</i> [0]	<i>“it Overall it was fairly easy to use. The color coding/similar save article features were the best part about it. I also liked how you could quickly view which articles you saved.”</i> [3]

Units are in italics. Bold indicates text shared by two or more units. Number of units appears in brackets.

mentions of features (e.g., piles, highlighting, *Similar* feature) and reasons. For instance, in response to the first question regarding the most positive things about the system and why, subjects would often identify one or more features and state why they felt they were positive. In many other cases, subjects only identified a feature without providing a reason. We thus refer to our units as features/reasons, which reflect that units can consist of a feature, if no reason is provided, or a feature–reason pair, if both a feature and reason are provided. The initial coding process captured raw occurrences, so if the same feature, or feature–reason pair was mentioned more than once, it was counted as many times as it appeared. Table 1 presents example responses from Subjects 13 and 34 along with how these responses were divided into units. Subject 13 was in the pen-and-paper condition and Subject 34 was in the electronic condition. These responses are representative of all subjects’ responses.

Before the unitizing process began, coders met to discuss the process and to code several responses together. After coding units independently, coders meet to compare results and discuss differences. In general, there was very low agreement about what constituted a unit – in most cases agreement was around 60% and with interview responses, this number was even lower. Responses produced in the pen-and-paper and electronic modes tended to be in the form of sentences or phrases, which made unitizing responses a bit easier (as demonstrated by the examples in Table 1). Responses produced in the interview mode were usually quite long, with a lot of incomplete utterances, false starts and backtracking. Because of the low agreement, coders engaged in a consensus method to resolve disagreements and produce a final listing of units for each response.

We developed several measures to characterize subjects’ responses. Our hypothesis regarding these responses indicated expected differences in the lengths and informativeness of responses. We operationalized length as the raw number of words in each response. This measure included stop words because in most cases

Table 2
Measures of response informativeness

Dimension	Measure
Length	The number of words per response
Raw units: question	The number of units contained in subjects’ responses to a single question. Duplicate units are included in this measure
Unique units: question	The number of unique units contained in subjects’ responses to a single question. If the same units are mentioned twice in the response, then they were only counted once
Raw units: response	The number of the units contained in subjects’ responses to all four questions. Duplicate units are included in this measure
Unique units: response	The number of unique units contained in subjects’ responses to all four questions. If the same unit is mentioned in response to two different questions, then it is only counted once
Efficiency	The total number of unique units divided by the total number of raw units (unique units: response/raw units: response)

stop words were necessary to understand the meaning of subjects' responses. We conceptualized informativeness using several different dimensions and operationalized these dimensions accordingly. All of these measures are based on the basic units (features/reasons). Table 2 presents all such dimensions and how we measured them.

4. Results

Before reporting results regarding the two main hypotheses of this study, it is necessary to rule out the possibility that any potential differences were caused by experimenters. *T*-tests were conducted using experimenter as the independent variable and responses to closed question, and measures of length and informativeness of responses to the open-ended questions as dependent variables. Results from these tests were non-significant allowing us to eliminate the possibility that potential differences were due to experimenter effects.

4.1. Closed questions

The first hypotheses stated that subjects in the interview mode would rate systems more favorably than subjects in the pen-and-paper and electronic modes. Table 3 shows each closed question, means and standard deviations for scores in each mode, *F*-scores associated with the one-way ANOVA and results from the post-hoc tests (*Diff.*). The degrees of freedom for all tests is 2, 48. Statistically significant *F*-scores are marked with asterisks. Results reported in the *Diff.* column indicate between which modes statistically significant differences were detected with Scheffe's post-hoc tests; for example, $E > I$ means that the mean score for the electronic mode was significantly higher than the mean score for the interview mode. For 10 of the 21 questions, statistically significant differences in scores were detected between modes, and in all of these cases scores assigned by subjects in the electronic mode were significantly higher than scores assigned by subjects in the interview mode.

Table 3
Means, standard deviations, and *F*-scores for closed questions from the Exit Questionnaire according to mode

Question	Pen-and-paper (<i>P</i>)	Electronic (<i>E</i>)	Interview (<i>I</i>)	<i>F</i>	Diff.
1. The search methods I used in this study were similar to those I use when I normally search the Web	4.69 (1.66)	5.39 (1.38)	4.76 (1.56)	1.10	–
2. The system was easy to learn to use	6.38 (.50)	6.78 (.43)	6.00 (.79)	7.54**	$E > I$
3. The system can be used effectively without instruction	4.38 (1.63)	4.89 (1.61)	4.71 (1.10)	.530	–
4. I didn't notice any inconsistencies when I used the system	5.19 (1.23)	6.22 (1.11)	5.47 (1.66)	2.64	–
5. It was easy to pose queries to the system	5.94 (1.12)	6.39 (.85)	6.29 (.85)	1.06	–
6. It was easy to navigate the search results	6.00 (1.21)	6.50 (.86)	5.41 (1.62)	3.25*	$E > I$
7. The color-coding of the piles made sense to me	6.63 (.81)	6.89 (.32)	6.53 (.80)	1.34	–
8. Overall, the system was easy to use	5.94 (.85)	6.61 (.50)	5.88 (.99)	4.47*	$E > I$
9. The system gave me control over my searching activities	5.69 (1.14)	6.06 (.87)	5.35 (.99)	2.15	–
10. The system made things I wanted to accomplish easy to do	5.56 (.81)	5.94 (.99)	5.00 (1.32)	3.44*	$E > I$
11. The system helped me find relevant documents quicker than search systems I normally use	5.06 (1.48)	5.50 (1.34)	4.24 (1.20)	3.98*	$E > I$
12. In general, it was easy to find relevant documents with the system	5.56 (1.15)	6.28 (.75)	5.06 (.96)	7.10**	$E > I$
13. It was easy to understand why documents were retrieved in response to my query	5.19 (1.05)	5.56 (1.24)	5.41 (1.23)	.415	–
14. It was easy to determine if a document was relevant to a task	5.00 (.82)	5.72 (.96)	5.76 (1.25)	2.88	–
15. The system helped me identify different aspects of my task	5.06 (1.12)	5.44 (1.29)	5.53 (.94)	.791	–
16. The system helped me explore different aspects of my task	5.44 (1.15)	6.06 (1.11)	5.06 (1.30)	3.14*	$E > I$
17. The system helped me find documents that were relevant to different aspects of the task	5.50 (1.10)	5.89 (1.53)	5.47 (.94)	.636	–
18. The various functions of the system were well integrated	5.75 (.93)	6.39 (.61)	5.53 (.87)	5.33**	$E > I$
19. Overall, the system was effective in helping me complete search tasks	5.75 (.86)	6.44 (.62)	5.47 (.94)	6.69**	$E > I$
20. If this system were available for use, I would use it frequently	5.38 (1.59)	6.11 (1.02)	5.06 (1.20)	3.13*	$E > I$
21. Overall, I am satisfied with my performance	6.06 (.93)	6.00 (.77)	5.65 (.99)	1.05	–

Diff. column represents significant differences between modes as detected by post-hoc tests. (* $p < .05$, ** $p < .01$).

Table 4
Patterns of scores, frequency of occurrence and number of significant differences detected

Pattern	Frequency	Significant differences
$I < P < E$	13	10
$P < I < E$	5	0
$P < E < I$	2	0
$I < E < P$	1	0

Although no support was found for our first hypothesis, these results provide some insight into mode effects in interactive IR experiments. Readers are reminded that the basic assumption is that lower scores represent more critical, and hence more valid, scores since subjects tend to inflate their ratings of systems. Overall, the general trend was that scores assigned by subjects in the interview mode were lower than scores assigned by subjects in the pen-and-paper and electronic modes. Table 4 shows each pattern that occurred in the data, the frequency with which these patterns occurred, and the number of statistically significant results that each pattern represented. Patterns represent relationships between scores, so that $I < P < E$ means that the average score for the interview mode was the lowest, for the pen-and-paper mode second lowest and for the electronic mode highest.

This data seem to suggest a trend for subjects in the interview mode to be more critical in their ratings of systems than subjects in the electronic mode. For 18 out of 20 questions, subjects' responses in the electronic mode were higher than subjects' responses in the interview and pen-and-paper modes, and subjects' mean ratings in the electronic mode were never the lowest. Moreover, all 10 of the statistically significant differences displayed in Table 3 were in the same order. In most cases, differences between the interview and pen-and-paper modes were negligible, which suggests that subjects in the pen-and-paper mode were also more critical than subjects in the electronic mode. However, post-hoc tests did not demonstrate significant differences between the pen-and-paper and electronic modes.

Another way to examine the data is to explore the extent to which subjects' ratings vary. This analysis provides an indirect way to explore acquiescence, where greater acquiescence will be indicated by less variance across ratings assigned by a single subject. In other words, little variance will be observed in cases where a subject responded to all questions with the same or similar scores (more acquiescence) and greater variance will be observed in cases where a subject used a variety of ratings (less acquiescence). Since all ratings were on the same scale and in the same direction, we created variables representing the average and standard deviation for each subject across all 21 closed questions. The means and standard deviations for the average and variance variables are presented in Table 5. Overall, subjects in the electronic mode had a significantly higher average rating than subjects in the pen-and-paper and interview modes, $F(2, 48) = 4.42, p < .01$, which is consistent with the previous results, and a lower variance than subjects in the pen-and-paper and interview modes, although not significantly so. It is interesting to note that again, scores elicited via the pen-and-paper and interview questionnaires are similar to one another, but slightly different from those elicited via the electronic. Overall, it appears that all subjects, regardless of mode, are only using a small range of numbers to describe their attitudes about the system.

4.2. Open questions

The second hypothesis stated that subjects' responses to open questions in the interview mode would be longer and more informative than subjects' responses in the pen-and-paper and electronic modes. Fig. 4

Table 5
Means and standard deviations for average rating and variance for each subject across all 21 closed questions according to mode

Mode	Average rating	Variance
Pen-and-Paper	5.33 (.71)	.979 (.280)
Electronic	6.05 (.60)	.903 (.284)
Interview	5.42 (.72)	.972 (.282)

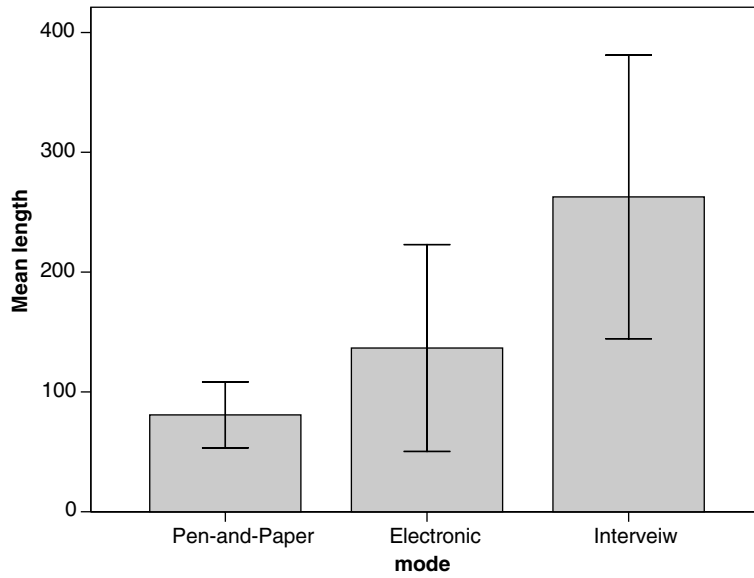


Fig. 4. Average lengths of subjects' responses to open questions according to questionnaire mode (bars represent \pm one standard deviation).

displays the mean lengths of subjects' responses to all four questions according to mode, along with bars representing the standard deviations (\pm one standard deviation). The average lengths and standard deviations of

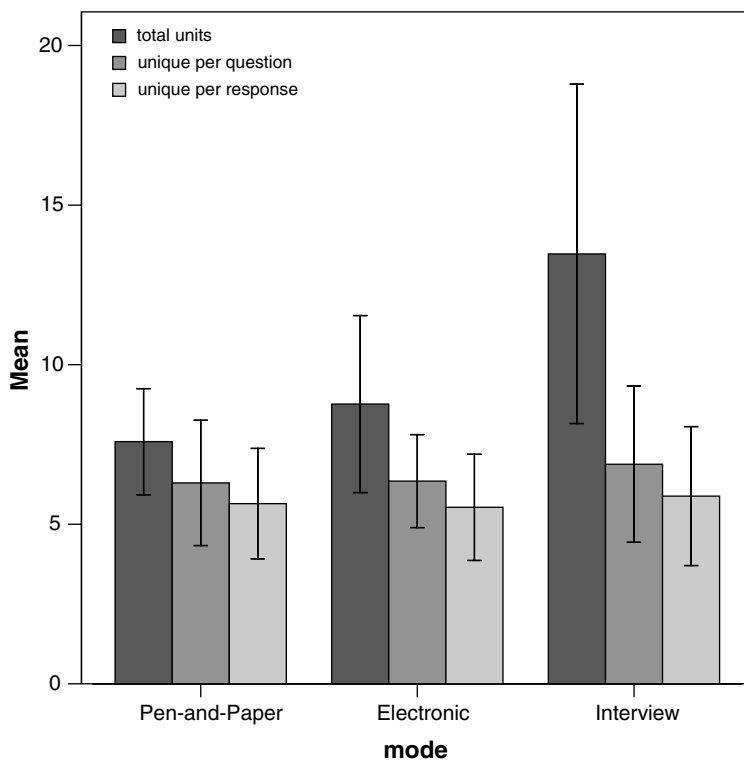


Fig. 5. Average number of units, unique units per question and unique units per response according to questionnaire mode (bars represent \pm one standard deviation).

subjects' responses in the pen-and-paper, electronic and interview modes were: 80.82 (27.47), 136.65 (86.35) and 262.76 (118.36), respectively. These differences were highly significant, $F(2,48) = 19.94$, $p < .000$ ($R^2 = .45$) and post-hoc tests detected statistically significant differences for all pairs. These results provide support for our hypothesis regarding length of subjects' responses to open-ended questions according to questionnaire mode.

Looking at the lengths of subjects' responses provides only one way to compare differences between modes; we were also interested in looking at the informativeness of subjects' responses. There is more physical effort involved in responding to pen-and-paper questionnaires than electronic questionnaires, and more effort with electronic than with interview. The average lengths of subjects' responses reflect this to a certain extent and are therefore, not too surprising. Examining the informativeness of subjects' responses is crucial since subjects in the interview mode could merely be repeating themselves, discussing irrelevant topics, or simply using more words to express the same number of ideas. Fig. 5 shows the average number of units identified in subjects' responses, as well as the number of unique units per question and per response. With respect to total units, we see the same trend in Fig. 4, although not quite as pronounced (pen-and-paper: 7.59 (1.66); electronic: 8.76 (2.74); and interview 13.47 (5.32)). These difference were statistically significant, $F(2,48) = 12.76$, $p < .000$ ($R^2 = .35$) and post-hoc tests detected significant differences between all pairs. Thus, it appears that subjects in the interview mode are expressing more ideas than subjects in the pen-and-paper and electronic.

When we look at the uniqueness measures, however, the figures converge. These results indicate that overall, each mode is eliciting a similar amount of usable feedback. While it appears that subjects in the interview mode are expressing more ideas, many of these ideas are the same; subjects are not necessarily providing any new ideas, but rather seem to be repeating their ideas. The means for unique units per question for pen-and-paper, electronic and interview were 6.29 (1.96), 6.35 (1.41) and 6.88 (2.45), respectively, and the means for unique units per response were 5.65 (1.73), 5.50 (1.62), and 5.88 (2.18), respectively. None of these differences

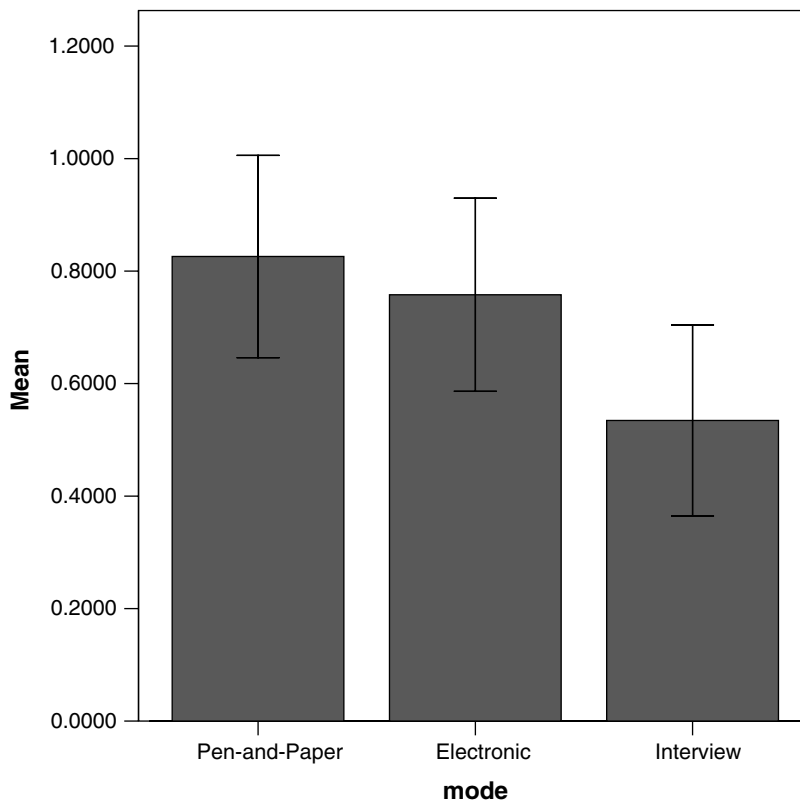


Fig. 6. Efficiency of subjects' responses (bars represent \pm one standard deviation).

were statistically significant. Thus, subjects identified similar amounts of unique units at the question level and at the response level regardless of mode. Note that for each measure, greater standard deviations occurred in the interview mode, which demonstrates that there was greater variability in this mode.

Subjects in the pen-and-paper and electronic modes provide a similar amount of usable feedback as subjects in the interview mode, but do they do so in fewer words? Fig. 5 seems to suggest this, but to investigate this question we look at the efficiency of subjects' responses. As a reminder, efficiency was measured by the number of unique units identified by a subject divided by the total number of units identified by that subject. Fig. 6 shows the efficiency of subjects' responses. There were significant differences in the efficiency measure across condition, $F(2,48) = 13.08$, $p < .000$ (pen-and-paper: .83 (.18), electronic: .76 (.17) and interview: .53 (.17)). Post-hoc tests showed that differences between all pairs were statistically significant. Combined with the results in the preceding paragraph, these results suggest that while subjects in the pen-and-paper mode identify the same amount of unique units as subjects in the electronic and interviews modes, they do so in fewer words, with less repetition. It appears that the extra effort required to communicate one's responses in the pen-and-paper mode may have some impact on the conciseness of subjects' responses.

After examining subjects' responses to open questions, we noticed two other behaviors, which might be construed as indicators of satisficing. We noticed a number of back-references in subjects' responses. Back-referencing describes the behavior where a subject responds to one question by referencing his response to another question. The most egregious example of this was subjects writing "see previous response," which was observed three times in each of the pen-and-paper and electronic modes, but only once in the interview mode. The χ^2 test did not show that this distribution was statistically significant, but it appears that back-referencing is more likely to happen in self-administered questionnaire modes. The interview mode may make back-referencing more difficult to do because of the interaction with the interviewer. Given that subjects in the interview mode used significantly more units to communicate a similar amount of unique feedback, it may be the case that back-referencing in the interview mode happens through repetition, rather than explicit reference to previous responses. We observed one extreme satisficing tactic in the electronic condition: a subject copy-and-pasted his response to one question into the response area for another question. Although we only observed a single instance of this behavior, it seems reasonable that electronic questionnaires might encourage more extreme forms of satisficing.

5. Discussion

In this study we investigated the relationship between questionnaire mode and subjects' responses to a usability questionnaire comprised of closed and open questions administered during an interactive IR experiment. Social desirability theory has been used extensively to explain a number of differences that have been observed between questionnaire modes, and it was used in part to motivate the hypotheses of this study. Three questionnaire modes (pen-and-paper, electronic and interview) were explored with 51 subjects who used one of two information retrieval systems. We hypothesized that subjects' quantitative ratings of systems would be more positive in the interview mode than in the pen-and-paper and electronic modes. We further hypothesized that subjects would provide longer and more informative responses to open questions in the interview mode than in the pen-and-paper and electronic modes.

Results demonstrated that subjects' quantitative ratings of systems in the electronic mode were significantly more positive than subjects' ratings in the pen-and-paper and interview modes. The general trend was that scores assigned by subjects in the interview mode were lower than scores assigned by subjects in the pen-and-paper and electronic modes, and that scores assigned by subjects in the interview and pen-and-paper modes were more alike than those assigned by subjects in the electronic mode. Overall, subjects in the interview and pen-and-paper modes were more critical in their ratings of systems than subjects in the electronic mode. These results suggest that researchers should use the interview mode to elicit responses to closed questions.

Although no support was found for the first hypothesis, our results provide some important insight into mode effects in interactive IR experiments. Social desirability theory would predict that subjects' ratings in the interview mode would be more positive than subjects' ratings in the pen-and-paper and electronic modes. However, it does not appear that this theory offers a good explanation of what happened in this study. It may

be the case that subjects in the interview mode believed that their responses were more likely to be valued and used since they were presenting them directly to a researcher and, as a result, were more critical in providing feedback. Accordingly, it may be the case that subjects in the electronic mode were least critical because they believed they were submitting feedback to a 'black hole' and it was unclear to them when a person would review their comments. In some ways, this is related to the idea of anonymity described earlier, only in an opposite way than was proposed by previous researchers (Richman et al., 1999).

Another possible explanation for these results is related to the mode of delivery for the entire experiment and the concept of flow (Csikszentmihalyi, 1997). All subjects used a computer to complete the primary portion of the experiment (using an experimental IR system). While subjects in the pen-and-paper and interview modes switched interaction styles after using the system to complete the Exit Questionnaire, subjects in the electronic mode did not switch styles, and instead followed a link to an electronic version of the questionnaire and continued to use the mouse, keyboard and monitor to communicate responses. It is proposed that the flow that was maintained in the electronic mode caused subjects to be less critical and thoughtful. In this situation, flow prevented users from recognizing a task switch and adjusting their behaviors accordingly. Subjects' flows in the pen-and-paper and interview modes were interrupted (or perhaps disrupted is more appropriate) which functioned to signal the end of one part of the experiment and perhaps gave them a few moments to reflect on their experiences before they began the next. This, in turn, may have caused them to be more critical and thoughtful. Overall, these results suggest that in interactive IR experiments, some interruption should happen to signal the end of one part of the experiment before subjects move on to another part. Eliciting usability ratings through interviews appears to be one way to accomplish this. This can also be accomplished by providing subjects with a break or asking subjects to complete an unrelated task away from the computer.

Another possible explanation for why subjects' ratings were more critical in the interview mode is related to the pace of the questionnaire. Previous studies have demonstrated that the pace of an interview can often impact response quality, especially in the context of telephone interviews (Groves, 1978). It is conjectured that in this study, the interview mode slowed down the pace of the questionnaire which may have resulted in more critical ratings. Subjects were unable to rapidly skim questions and response sets; instead they were required to listen carefully and move at a pace set by the interviewer.

Although there were significant differences in usability ratings according to mode, overall, subjects' ratings were still quite high and most subjects used only a small range of numbers on the higher end of the scale to characterize their opinions. Subjects still seem to exhibit acquiescence, whether in relation to their tendency to agree with attitude statements or in relation to their tendency to agree with the experiment (demand effects). Although we did not vary the direction of our statements to test for item acquiescence, acquiescence still appears to be a problem that needs to be addressed.

With respect to subjects' responses to open questions, results of the study demonstrated that subjects in the interview mode provided significantly longer responses to the four open questions. Subjects' responses in this condition also contained significantly more units, although there were no differences in the number of unique units subjects identified across mode. While subjects in the interview mode identified more units than subjects in the pen-and-paper and electronic modes, many of these units were repetitions of previously identified units. Results showed that subjects in the pen-and-paper mode were significantly more efficient in communicating their responses than subjects in the electronic or interview modes, and that subjects in the electronic mode were significantly more efficient than subjects in the interview mode.

Overall, subjects in the pen-and-paper mode produced the most concise responses, which is likely related to the increased physical effort involved with producing responses. Although the lower physical effort associated with the interview mode may have encouraged subjects to produce more data, it was not necessarily better data. Instead, it appears that the interview mode may have encouraged subjects to engage in more satisficing behaviors. While the decreased pace provided by the interview mode was beneficial for closed questions, it may have caused subjects to respond more quickly to open questions to avoid uncomfortably long silences. Because these responses were generated more quickly, they likely contained more repetition. There was also no visual instantiation of subjects' responses, and this too, may have caused more repetition. The pen-and-paper and electronic modes were also not free of satisficing behaviors. There were six instances of back-referencing in the pen-and-paper and electronic modes, while only one instance was observed in the interview mode. Each mode apparently elicits a different type of satisficing behavior.

It is important to note one very important benefit to having subjects respond to open questions via pen-and-paper or electronic questionnaire: subjects' responses in the pen-and-paper and electronic modes were much more well-formed than subjects' responses in the interview mode. This made analyzing the data much easier. Subjects' responses in the pen-and-paper and electronic modes were typically well-formed sentences. Subjects' responses in the interview mode were typically disjointed utterances, with a number of fits and starts, and lots of back-tracking and digression. Even though there was more physical effort required of subjects to communicate their responses in the pen-and-paper and electronic modes, this extra effort seems to have resulted in much more well-formed, concise and understandable responses. Editing can be cumbersome in the pen-and-paper mode and this may have also impacted the conciseness of subjects' responses in this mode.

The relative well-formedness of subjects' responses had many implications for the overall costs associated with conducting this research. Not only were there greater research costs associated with analyzing data produced in the interview mode, there were also greater costs associated with transcribing interviews. The process of unitization was extremely arduous in this condition, which likely affected the reliability of the coding. Surprisingly, costs associated with entering data collected via pen-and-paper mode were minimal because data was entered by the researcher via Web form immediately following subjects' experimental sessions or during consecutive experimental sessions. Thus, with respect to eliciting subjects' responses to open questions, results of this study suggest that pen-and-paper or electronic questionnaires are optimal; if the latter technique is used it is important to include a break between subjects' uses of the system and their completion of questionnaires.

While this work represents a reasonable contribution to the literature, its results are by no means definitive and there are some limitations to consider. As with all other studies, the particular users, tasks and systems may have influenced the results. Our users were paid undergraduate students from one university in the US who completed four tasks with two experimental systems. Although we included a variable for interviewer and were careful with our study design, there still may have been some differences caused by interviewer. Finally, the process of unitizing responses was very difficult; our attempts to independently code units resulted in low reliability, which required us to follow this approach with the consensus method.

6. Conclusions

Questionnaires are an important part of interactive IR studies and are used extensively to collect a wide range of measures from users, most notably system usability ratings. However, subjects' general tendencies to inflate their ratings of systems call into question the validity of data collected via questionnaires and make it virtually impossible to make reliable system comparisons. This, in turn, places severe limitations on what researchers are able to learn about interactive IR systems and processes. Therefore, identifying better ways of collecting usability data from subjects should be one of the chief concerns for interactive IR researchers.

In this study, we investigated some factors related to subjects' response behaviors to questionnaires. There are two main findings of our study. First, closed questions administered in the interview mode elicited lower ratings on average than pen-and-paper or electronic modes. Second, open questions in the pen-and-paper or electronic modes resulted in more concise and coherent responses, with no appreciable reduction in information content, even though responses in interview mode tended to be significantly longer. Furthermore, verbal responses obtained in interview mode were far harder to process and analyze. These results suggest that, at least for some types of interactive IR experiments, the post-system questionnaire takes the form of an interview for closed questions, followed by pen-and-paper or electronic mode for open questions. The interview serves the useful purpose of clearly demarcating the performance of the task(s) from the evaluation of the system. It seems likely that this would also hold for post-task questionnaires, but this would need to be confirmed by further study.

Our results provide some insight into the relationship between questionnaire mode and subjects' response behaviors. Although it may seem cliché to say so, more work should clearly be done in this area. It is vital to the field of interactive IR that adequate attention is paid to understanding more about how research design can potentially impact results, to developing and improving our data collection techniques, and to training and educating new researchers on the intricacies of behavioral research. It is sometimes easy to forget that our study results are only as good as our study methods.

Acknowledgements

We would like to thank Professors Barbara Wildemuth and Stephanie Haas from UNC for their helpful discussions about this paper.

References

- Booth-Kewley, S., Edwards, J. E., & Rosenfeld, P. (1992). Impression management, social desirability and computer administration of attitude questionnaires: Does the computer make a difference? *Journal of Applied Psychology*, 77(4), 562–566.
- Carney, T. F. (1971). Content analysis: A review essay. *Historical Materials Newsletter*, 4(2), 52–61.
- Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 56(1), 71–90.
- Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), paper no. 152.
- Brewer, N. T., Hallman, W. K., Fiedler, N., & Kipen, H. M. (2004). Why do people report better health by phone than by mail? *Medical Care*, 42(9), 875–883.
- Bulmer, M. (2004). *Questionnaires, V.1*. Thousand Oaks, CA: Sage Publications.
- Csikszentmihalyi, M. (1997). *Finding flow: The psychology of engagement with everyday life*. New York: Basic Books.
- Czerwinski, M., Horvitz, E., & Cutrell, E. (2001). Subjective duration assessment: An implicit probe for software usability. In *Proceedings of IHM-HCI 2001 conference, Lille, France* (pp. 167–170).
- Dumais, S. T., & Belkin, N. J. (2005). The TREC Interactive Tracks: Putting the user into search. In E. M. Voorhees & D. K. Harman (Eds.), *TREC: Experiment and evaluation in information retrieval* (pp. 123–153). Cambridge, MA: MIT Press.
- Groves, R. M. (1978). On the mode of administering a questionnaire and responses to open-ended items. *Social Science Research*, 7, 257–271.
- Groves, R. M. (1989). *Survey errors and survey costs*. New York, NY: John Wiley & Sons.
- Harper, D. J., & Kelly, D. (2006). Contextual relevance feedback. In *Proceedings of the first symposium on information interaction in context, Copenhagen, Denmark*.
- Hayslett, M. M., & Wildemuth, B. M. (2004). Pixels or pencils? The relative effectiveness of Web-based versus paper surveys. *Library & Information Science Research*, 26, 73–93.
- Hersh, W., & Over, P. (1999). TREC-8 Interactive Track report. In D. Harman, & E. M. Voorhees (Eds.), *The eighth text retrieval conference (TREC-8)* (pp. 57–64).
- Kiesler, S., & Sproull, L. S. (2001). Response effects in the electronic survey. *Public Opinion Quarterly*, 50, 402–413.
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213–236.
- Lautenschlager, G. J., & Flaherty, V. L. (1990). Computer administration of questions: More desirable or more social desirability? *Journal of Applied Psychology*, 75(3), 310–314.
- Lund, A. M. (2001). Measuring usability with the USE questionnaire. *Usability and User Experience*, 8(2). <http://www.stcsig.org/usability/newsletter/0110_measuring_with_use.html>.
- Martin, C. L., & Nagao, D. H. (1989). Some effects of computerized interviewing on job applicant responses. *Journal of Applied Psychology*, 74, 72–80.
- Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage Publications.
- Newton, D., Engquist, G., & Bois, J. (1977). The objective basis of behavior units. *Personality and Social Psychology*, 35, 847–862.
- Nielsen, J., & Levy, J. (1994). Measuring usability – Preference vs. performance. *Communications of the ACM*, 37(4), 66–75.
- Payne, S. L. (2004/1951). Who left it open? A description of the free-answer question and its demerits. In M. Bulmer (Ed.), *Questionnaires, V.1* (pp. 131–147), Thousand Oaks, CA: Sage Publications. (Reprinted from Payne S. L. (1951). *The art of asking questions* (pp. 170–184). Princeton, NJ: Princeton University Press.)
- Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology*, 84(5), 754–775.
- Rosenfeld, P., Booth-Kewley, S., Edwards, J. E., & Thomas, M. D. (1996). Responses on computer surveys: Impression management, social desirability, and the big brother syndrome. *Computers in Human Behavior*, 12(2), 263–274.
- Schuman, H., & Presser, S. (2004/1996). The acquiescence quagmire. In M. Bulmer (Ed.), *Questionnaires, V.2* (pp. 319–347), Thousand Oaks, CA: Sage Publications. (Reprinted from Schuman, H., & Presser, S. (Eds.) (1996). *Questions and answers in attitude surveys: Experiments on question form* (pp. 203–230). London: Sage Publications.)
- Schwarz, N., Strack, F., Hippler, H.-J., & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology*, 5, 193–212.
- Simon, H. A. (1957). *Models of man*. New York: Wiley.
- Tague-Sutcliffe, J. M. (1992). The pragmatics of information retrieval experimentation, revised. *Information Processing and Management*, 28(4), 467–490.
- Thomas, P., & Hawking, D. (2006). Evaluation by comparing result sets in context. In *Proceedings of the conference on information and knowledge management (CIKM '06)*.

- Toms, E. G., Freund, L., & Li, C. (2004). WiIRE: The Web interactive information retrieval experimentation system prototype. *Information Processing & Management*, 40, 655–675.
- Tourangeau, R. (1984). Cognitive sciences and survey methods. In T. B. Jabine et al. (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73–100). Washington, DC: National Academy Press.
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. New York, NY: Cambridge University Press.
- Weisband, S., & Kiesler, S. (1996). Self disclosure on computer forms: Meta-analysis and implications. In *Proceedings of ACM special interest group on computer human interaction (CHI '96), Vancouver, BC* (pp. 3–10).