Effects of Tasks at Similar and Different Complexity Levels

Yinglong Zhang
School of Information
University of Texas, Austin
ylZhang@utexas.edu

Jacek Gwizdka
School of Information
University of Texas, Austin
asist2014@gwizdka.com

ABSTRACT
We present preliminary results from a Web search study, where we paid particular attention to the design of complex tasks. We used subjective, behavioral, and cognitive measures to investigate the effects of task complexity and, separately, of search user interface elements. Our findings confirm the expected task properties by design – no differences between tasks at the same level of complexity, but significant differences between tasks at different levels of complexity on most dependent variables. The examined elements of user interface did not have a significant effect on experienced mental workload, however, they affected search behavior, but only on complex tasks.

Categories and Subject Descriptors
H3.3 Information Search and Retrieval: Search process.

Keywords
Interactive information retrieval; evaluation; search tasks.

INTRODUCTION
Understanding search tasks and their subjective perception is important for many reasons. For example, in order to better understand search behavior we need a good understanding of effects of task properties on task performance as well as understanding how users' perception of task properties relates to tasks. In the studies of interactive information retrieval, where search tasks are assigned to users, researchers need to be able to design search tasks of desired properties. Ability to design tasks requires a good understanding of their properties. Not surprisingly, a significant amount of work was devoted to the investigation of search task properties, and, in particular, to researching measures and perception of task complexity. In this poster, we report on preliminary data analysis and consider if task complexity by design is reflected in behavioral, cognitive and subjective measures. In contrast to previous work, we not only compare tasks between the levels of task complexity, but also within the same level of complexity.

RELATED WORK
Task complexity have been considered by numerous researchers (e.g., Byström & Jarvelin, 1995; Gwizdka, 2008; Li et al., 2011; Singer, Norbisrath, & Lewandowski, 2012) and was included in the faceted task classification (Li & Belkin, 2008). In this short paper we can mention only a very limited set of related papers. We feel that it is justified, given many excellent reviews in this area (e.g., in the above cited papers).

Singer et al. (Singer et al., 2012) considered tasks as complex if they require aggregation, discovery, synthesis, and showed that Web users are not always able to assess various aspects of task effort and difficulty after task performance. Wu et al. (Wu, Kelly, Edwards, & Arguello, 2012) discuss design of search tasks using cognitive complexity levels informed by Anderson & Krathwohl's taxonomy of learning. The higher levels of cognitive complexity involved analysis, evaluation, creation (in the increasing order of expected and experienced levels of complexity/difficulty). Toms et al. (Toms et al., 2008) created search tasks for INEX 2007 that ranged from fact-finding and to decision-making. Decision making was considered to be more complex type of task.

METHOD

Experimental Design and Procedure
We conducted a controlled, lab-based experiment of information search on Wikipedia. Participants (N=25; age 18 to 37; 19 females), who were native English speakers and had a normal, to corrected-to-normal vision, came for individual sessions. Each experiment session was completed within 1.5 hours, and participants received $30 for their participation. Participants were asked to complete four search tasks (Table 1). The tasks were designed to be at two complexity levels: simple and complex and were performed by using a commercial test search engine created by Search Technologies Corp. with two variations of user interface (UI) created by us. One interface variation displayed a list of Wikipedia categories sorted by frequency of occurrence in the search results along a list of the retrieved search results (Error! Reference source not found.). The other UI variation did not include the categories. The experiment had a within-subject design with each participant conducting two search tasks of different complexity using each of the search interfaces. Combining the two types of UIs and four different task scenarios, we created eight different task+UI combinations. In order to minimize the order effect, the eight task+UI combinations were used to create 32 rotations, with a constraint that UI is switched after two tasks. The 32 rotations were assigned to participants in a random order. In each task, participants read task description, completed pre- and post- task questionnaires, and searched Wikipedia pages using one of the two search interfaces Participants

77th ASIS&T Annual Meeting, Oct. 31-Nov. 4, 2014, Seattle, WA, USA.

were asked to save the pages they considered as useful to the search task and to add notes to these pages. In addition, participants responded to a secondary task while searching. At the end of a session, they answered exit questionnaire. There were no time limits set for search tasks.

Tasks Selection and Design

We started by selecting four tasks from prior studies conducted by other researchers. The aim of the selection was to obtain a set of two simple and two complex tasks. In our selection of task topics, we made sure that enough information for the tasks existed in Wikipedia. Two tasks were selected from (Wu et al., 2012), one at the low-level of cognitive complexity ("remember") and one at the higher-level of cognitive complexity ("analyze"). One complex task (decision-making) was selected from (Toms et al., 2008), and one simple (fact-finding) from (Gwizdka, 2010). One particular difficulty encountered in our own and others work is how to design tasks that are "significantly" complex, and yet doable by ordinary search engine users in the course of one session. Driven by this concern, we further revised the complex tasks. We used the following criteria: a complex tasks should be (1) difficult to generate queries for; (2) require aggregation of information; (3) need synthesis of information from different web pages. The latter two criteria are based on (Singer et al., 2012).

ID	Type	Task scenario
1	Simple 1	You love history and, in particular, you are interested in the Teutonic Order (Teutonic Knights). You have read about their period of power, and now you want to learn more about their decline. You want to find out: What year was the Order defeated in a famous battle? And you also want to find out which army (or armies) defeated the Order?
2	Simple 2	You recently attended an outdoor music festival and heard a band called Wolf Parade. You really enjoyed the band and want to purchase their latest album. What is the name of their latest (full-length) album? And you also want to know when this band resumed their work together?
3	Complex 1	A local water conversation group requests ideas to expand their efforts. Currently, they pick up debris from local waterways and try to raise awareness about water pollution. In an effort to help out, you volunteer for the group but also, you want to expand their efforts. What other forms of land use are impacting waterways? Which forms of land use have are the highest impact to the environment?
4	Complex 2	A debate is underway after an international logging and mining corporation submitted a bid to buy a local nature reserve. The city needs more jobs but many residents are upset because they find selling a nature reserve as short sighted. And many people actively use the nature reserve for recreation and educational field trips. In an effort to be balanced with support for the community and to be fair to economic development, you decide to investigate both sides further. What are the small and large scale impacts of logging and mining? What are some economic considerations for land preservation? What are your recommendations to the city if the corporation's bid is successful?

Table 1. Search Tasks.

Measures

The independent variables included the level of task complexity (simple/complex) and the variant of search interface (with and without categories). In order to assess how task complexity and search user interfaces affect the experienced task difficulty and search behavior, we collected the following dependent measures:

1. Subjective measures of perceived effort were obtained using the NASA TLX (task loading index) instrument after each search task (Hart & Staveland, 1988).
2. Behavioral measures included time on task; the number of unique queries entered; and the counts of visits to search engine results pages, and to Wikipedia pages (unique and all visits) during each search task.
3. Cognitive dynamic changes of participants' mental workload were measured through a secondary task (Gwizdka, 2010). Our secondary task was based on Stroop effect (MacLeod, 1992).

We also collected eye-tracking data in this experiment; however, the results from eye-tracking data analysis are planned to be included in a future and longer publication.

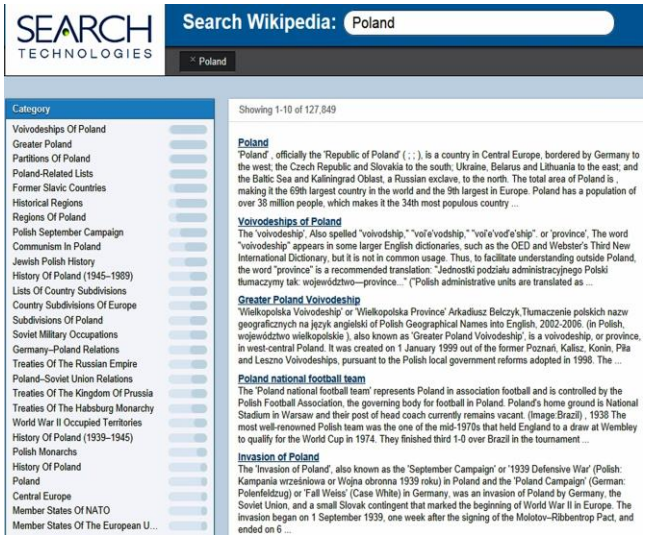

Figure 1. Search engine user interface with search results on the right and categories shown on the left

Hypotheses

The tasks of different complexity levels were expected to be perceived as tasks that demanded, accordingly, different levels of mental effort. Thus, we hypothesized that:

H1: *There will be significant differences between simple and complex tasks, in terms of subjective, behavioral, and cognitive measures.*

Our prior research hinted at helpfulness of presenting tags in a search UI (Gwizdka & Cole, 2013; Gwizdka, 2009) and at a possible effect of semantic categories in search UI on lowering mental effort (Gwizdka, 2010). Hence, we hypothesized that:

H2: *Search user interface with categories will decrease how task complexity and search user interface affect participants' experience of task difficulty, and will also affect behavioral and cognitive measures.*

Considering the careful design of search tasks, we hypothesized that:

H3: *There will be no significant difference in experienced task difficulty and in behavioral, and cognitive measures within simple and within complex tasks.*

RESULTS

Subjective Measures

Different aspects of participants' experienced workload were obtained from three NASA TLX items: mental demands ("How mentally demanding was the task?"), temporal demands ("How hurried or rushed was the pace of the task?"), and overall perception of hard work ("How hard did you have to work to accomplish your level of performance?"). Only the mental demands item showed significant effects of independent variables. We conducted ANOVA controlling for the repeated measures effects; the task complexity had a significant main effect on participant's experienced mental workload ($F(1,24)= 42.21$, $p<.001$), whereas there was no significant main effect of UI variation ($F(1,24)= 2.54$, $p>.01$). Also, no interaction effect was found between task complexity and search user interface ($F(1,24)=42.21$, $p=.84$). This was a somewhat surprising and contrary to H2 that the interface with categories did not decrease participants' experienced workload. On the other hand, agreeing with H1, in complex tasks, participants were more likely to experience a higher workload.

Behavioral measures

Task duration

The results of ANOVA indicated that complex tasks took significantly longer time ($F(1,24)=88.31$, $p<.001$). However, the interfaces with categories hardly reduced the tasks completion time ($p>.07$). There was no interaction effect between task complexity and UI variation ($p> .06$).

Number of queries

Task complexity ($F(1,24)=64.62$, $p<.001$) and UI variation ($F(1,24)=7.93$, $p<.001$) had main effects on the number of queries used in search tasks. An interaction effect between the two factors was significant ($F(1,24)=5.24$, $p<.005$). Based on the results of post-hoc tests (Bonferroni), it was found that in the complex tasks participants entered more queries when using the UI with categories ($p<. 05$) than when the UI without categories ($p<. 001$), whereas in the simple tasks, the UI variation had no significant effects on the usage of queries ($p= 1.0$). Comparing task complexity, we found that, independently of the UI, the complex tasks made participants use more queries than the simple tasks did ($p<. 001$).

Search Engine Result Page (SERP) visits

Similar to the results for queries, the main effects of task complexity and UI variation existed on the number of SERP clicks ($p<.001$). Results of Bonferroni post-hoc tests demonstrated that participants click more frequently in complex tasks when using categories ($p<.05$), whereas in simple tasks similar effect did not exist ($p=.80$).

Wikipedia page visits

Analyzing the number of unique Wikipedia pages that had been visited by participants and the number of pages with revisits, we found that task complexity and UI variation both had significant main effect on them ($p<.001$).

	Task duration [s]	Queries	SERPs	Unique Wiki-pedia pages	Wiki-pedia with revisits
Simple task + UI w. Categories	210 (164.2)	1.64 (1.87)	2.28 (3.14)	1.88 (1.24)	2.32 (2.08)
Simple task + UI without Categories	192.1 (90.4)	1.12 (0.44)	1.24 (0.88)	1.84 (1.27)	2.08 (1.71)
Complex task + UI with Categories	855.7 (382.6)	7.56 (3.93)	10.40 (5.99)	8.00 (5.55)	10.36 (7.79)
Complex task + UI without Categories	725.5 (341.8)	5.40 (3.35)	6.48 (3.77)	5.88 (4.20)	7.40 (6.27)

Table 2. Results of Behavioral Measures (Mean (Std.)).

In addition, an interaction effect existed (Table 2). Bonferroni post-hoc test showed that, in simple tasks and complex tasks, UI variation had no significant effect on the number of Wikipedia page visits ($p>.05$). However, compared with simple tasks, complex tasks significantly increased the number of Wikipedia pages visits ($p<.001$).

Cognitive Measure

We attempted to assess cognitive workload by measuring participants' average response time (RT) to the secondary task. We could not find any significant main effects of tasks complexity and UI variation on the response time. This generally confirms findings from (Gwizdka, 2010), where there were no significant RT effects found at the whole task level, but only during certain task phases. We will perform a similar more fine-grain analysis in the future.

Within and Between Task Type Comparisons

We investigated the similarities and differences within simple and complex tasks. In exploring the data, it was found that the Mauchly's Test for sphericity had been violated. Therefore, we could not use ANOVA and adopted an alternative method, a linear mixed-effects model, in our data analysis (Field, Miles, & Field, 2012). We first created a baseline model, in which an intercept was fixed to 1, there were no predictors, and within-subject effect had been controlled. Next, we created a model, in which a predictor, task ID (Table 1), was added to the baseline model. The maximum likelihood method was used to estimate the model. If the Log likelihood ratio (LLR), for a given criterion variable, is significant between the baseline model and the proposed model, then the introduced predictor has significant effect on that criterion variable.

In this study, we examined the differences in terms of subjective, behavioral, and cognitive aspects. The criterion variables of our linear mixed-effects model were subjective workload, task duration, number of queries, number of SERP and Wikipedia page visits, and the average response time to secondary task. Based on the results of Log likelihood ration test, all criterion variables, except the secondary task RT, were significantly affected by task ID ($p<.001$). In a further analysis, the results of post-hoc

Tukey test indicated that there were no significant differences within the same type tasks (i.e., within the same task complexity level) ($p>.05$), but there were significant differences between task types (i.e., between simple and complex tasks) for the six criterion variables ($p<.001$). Thus subjective and behavioral measures had good discrimination ability between the tasks at different complexity levels. Based on these results, we conclude that the search tasks we designed have similar properties within each task complexity level and different properties between complexity levels.

DISCUSSION

Surprisingly, we found no difference in behavioral measures between search UIs for simple tasks, but found such differences for complex tasks. On complex tasks performed using the interface with categories, participants were more likely to spent more time on tasks, enter more queries, visit more SERPs and more Wikipedia pages. One possible explanation of this phenomenon is that categories not only assist people in narrowing down the search scope, but also provide support for sensemaking. The categories in our experiment came from Wikipedia ontology – taxonomy of terms assigned to articles by their authors or editors. These terms can help participants modify their conceptual maps of search tasks, refine their queries, and make sense of what they read. This finding is in contrast with (Wilson, Hurlock, & Wilson, 2012).

However, everything comes at a price. Although the information contained in categories can be helpful in search, people may use more time and invest more mental effort needed to learn and understand these resources. This may explain it well why no significant difference existed on perceived workload, between using UI with categories and using that without categories. It is likely that the workload of learning additional information provided by categories can offset some benefits brought by them.

Simple tasks require less aggregation, discovery, synthesis, in comparison to complex tasks. In other words, in simple tasks participants were likely to develop appropriate conceptual maps in a quick way and did not need to invest additional time and effort to learn information from categories. Consequently, no significant differences in behavioral measures were found for simple tasks, between two search interface conditions.

CONCLUSIONS AND FUTURE WORK

In this poster, we presented results from preliminary analysis, in which we used subjective, behavioral, and cognitive measures to investigate the effects of task complexity and categories in a search user interface.

We found H2 to be partially supported; there were no significant differences between the UI variations in terms of subjective and cognitive measures, whereas significant differences existed in behavioral measures (except in task time). Considering subjective and behavioral measures, our designed tasks showed good discrimination ability between the tasks at different complexity levels, but such advantage could not be found in the cognitive measure based on the secondary task (partial support for H1, and support for H3). Categories in the search user interface, affected participants' behaviors, but the effect was rarely significant on their perception of task difficulty.

Limitations include analysis performed only at the level of the whole task, thus we could not yet consider dynamics of in changes of mental demands during a search task execution. We will continue our data analysis, and plan to publish a full data analysis that includes eye-tracking data.

ACKNOWLEDGMENTS

Funded, in part, by the IMLS Career Development Grant #RE-04-11-0062-11A awarded to Jacek Gwizdka.

REFERENCES

Byström, K., & Jarvelin, K. (1995). Task complexity affects information seeking and use. *IP&M*, *31*(2), 191–213.

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. SAGE Publications Ltd. Thousand Oaks, CA.

Gwizdka, J. (2008). Revisiting search task difficulty: Behavioral and individual difference measures. *Proc. of ASIST 45*(1), 1–12. doi:10.1002/meet.2008.1450450249

Gwizdka, J. (2009). What a Difference a Tag Cloud Makes: Effects of Tasks and Cognitive Abilities on Search Results Interface Use. *Information Research*, *14*(4). Retrieved from http://informationr.net/ir/14-4/paper414.html

Gwizdka, J. (2010). Distribution of cognitive load in Web search. *JASIST*, *61*(11), 2167–2187. doi:10.1002/asi.21385

Gwizdka, J., & Cole, M. (2013). Does interactive search results overview help?: an eye tracking study. *Proc. of CHI'2013 Ext Abstracts* (pp. 1869–1874). New York, NY: ACM.

Hart, S. G., & Staveland, L. E. (1988). *Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research*. Retrieved from http://archive.org/details/nasa_techdoc_20000004342

Li, Y., & Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. *IP&M*, *44*(6), 1822–1837.

Li, Y., Chen, Y., Liu, J., Cheng, Y., Wang, X., Chen, P., & Wang, Q. (2011). Measuring task complexity in information search from user's perspective. *Proc. of ASIST*, *48*(1), 1–8.

MacLeod, C. M. (1992). The Stroop task: The "gold standard" of attentional measures. *Journal of Experimental Psychology: General*, *121*(1), 12–14. doi:10.1037/0096-3445.121.1.12

Singer, G., Norbisrath, U., & Lewandowski, D. (2012). Ordinary search engine users assessing difficulty, effort, and outcome for simple and complex search tasks. *Proceedings of IIiX'2012* (pp. 110–119). New York, NY: ACM.

Toms, E., OBrien, H., Mackenzie, T., Jordan, C., Freund, L., Toze, S., … A, T. (2008). Task Effects on Interactive Search: The Query Factor. In *Focused Access to XML Documents* (Vol. 4862, pp. 359–372). Springer Verlag.

Wilson, M., Hurlock, J., & Wilson, M. (2012). Keyword clouds: having very little effect on sensemaking in web search engines. *Proc. of CHI'2012 Extended Abstracts* (pp. 2069–2074). New York, NY, USA: ACM.

Wu, W.-C., Kelly, D., Edwards, A., & Arguello, J. (2012). Grannies, Tanning Beds, Tattoos and NASCAR: Evaluation of Search Tasks with Varying Levels of Cognitive Complexity. *Proc. of IIiX'2012*. (pp. 254–257). New York, NY, USA: ACM.