

---

# Comparing the User Experience of Search User Interface Designs

**Tomi Heimonen**

Department of Computer  
Sciences, Tampere Unit for  
Computer-Human Interaction  
University of Tampere  
FI-33014 Tampere, Finland  
tomi.heimonen@cs.uta.fi

**Anne Aula**

Google  
1600 Amphitheatre Pkwy  
Mountain view, CA 94043 USA  
anneaula@google.com

**Hilary Hutchinson**

Google  
1600 Amphitheatre Pkwy  
Mountain view, CA 94043 USA  
hhutchinson@google.com

**Laura Granka**

Google  
1600 Amphitheatre Pkwy  
Mountain view, CA 94043 USA  
granka@google.com

**Abstract**

This paper introduces a user experience evaluation method designed to measure implicit and explicit user preference during the use of Web search interfaces. Our aim was to develop a method that scales to the needs of practical product development. We present the method and findings from a pilot study, and discuss the key issues and outcomes.

**Keywords**

Web search user experience evaluation, desirability, forced-choice presentation

**ACM Classification Keywords**

H5.2 User Interfaces: Evaluation/methodology.

**Introduction**

User experience is defined as “aspects of a digital product that users experience directly [...] Learnability, usability, usefulness, and aesthetic appeal are key factors in users' experience of a product.” [7] The evaluation of user experience is typified by the focus on usability and usefulness – it is important to make sure that the right product is being designed for the users' needs and that the product is easy and satisfactory to use, all the while providing benefits in terms of time

saved on task or a decrease in the amount of erroneous interactions while using the product.

Given the complex relationship between effectiveness, efficiency and user satisfaction, it is unlikely that these attributes correlate with one another universally [2]. For example, does greater satisfaction correlate with higher effectiveness? Is a faster and less error-prone user interface always the preferred one? Hornbæk and Law [3] go as far as to propose that users' perceptions do not correlate with objective measures. Clearly, measuring the post-task satisfaction does not provide enough information about the users' preferences as evidenced by their choices and actions during use. On the other hand, user experience design has to also account for business goals and strategy. The findings from user experience evaluations have to be tempered against the findings from performance tracking. In order to provide credible evidence on the applicability of a given design, researchers need to be able to provide enough data to both explain existing usage patterns and predict the effects of design changes.

### **Evaluating Web search user experience**

Web search interfaces differ from traditional productivity applications in terms of their user groups and the purpose of the use. Web search interfaces have no typical users or tasks, although different frameworks of Web search activities have been drawn up based on search log studies. This makes it challenging to set up conditions for user studies and to apply their results to more general search scenarios. Typical user studies present the participants with one or more interfaces and the participants carry out search tasks on different topics. This is combined with interviews, questionnaires and other methods of user feedback collection.

In our experience, traditional research methods do not adequately capture data on the users' actual preferences or credibly indicate which design they would use. When asked which interface one would like to use, participants might be likely to pick the one they used last or use some other criteria not associated with the actual use of the interface, such as memorable branding features. Given the difficulty users typically have with understanding how web search engines work and in formulating their own information needs, it can be difficult for the users to express the rationale for their preferences. Using traditional metrics, we can capture information on how satisfied the users were, but cannot reliably infer which design they would use if given a choice. This position paper outlines the findings of the development and pilot testing of a research method that aims to measure user preference with a two-pronged desirability metric – the implicit and explicit choices by participants during task completion.

### **Forced-choice UX evaluation framework**

In web search user interfaces, such as Google Web Search, it is difficult to introduce significant design changes and subsequently evaluate their effectiveness. The problem stems from the fact that when implementing changes into such a popular tool, one needs to have a significant amount of realistic data to back up any recommendations. Traditional usability studies are too contrived to produce reliable data that would answer questions relating to specific design issues. Live online experiments in which competing designs are released to a limited user population provide a wealth of quantitative data relatively quickly, but do not provide any insight into "why" one interface performs better, and many design changes cannot be tested this way for competitive reasons. With these

issues in mind we laid out two key requirements for a new method of gathering data:

- The results should provide insight on user preferences beyond the traditional methods.
- The method should be easy to set up and administer in parallel in a lab situation to cut down on time to gather enough data.

Our approach tries to augment the satisfaction component of user experience with the notion of desirability, as evidenced by the user's choice of the interface to use for a given task. Our hypothesis is that when presented with a choice of two designs, the user is more likely to choose the one that is more desirable, for whatever reason. Choice in itself does not explain the rationale behind it – the preferred design could appear to be faster, or provide a more aesthetic look and feel. We also need to capture subjective feedback along different dimensions of satisfaction after task completion. Correlations between the choices and feedback can point to strong trends or interesting discrepancies (e.g., users overtly prefer one design while actually using the other) that can lead to new research questions. As a practical approach to studying desirability, we propose to study two different designs at a time in a controlled experiment using a forced-choice paradigm. Forced choice in this context means that the participants are required to select which design of the two they wish to use for a task or a set of tasks, thereby providing their implicit preference. In addition, the participants are probed with open and closed desirability questions to gauge their explicit preferences and feedback on the designs.

The forced-choice paradigm is widely used in psychology, but we could find few studies where it is

applied to studying preference in the HCI context. Santella *et al.* [6] utilized it to compare their gaze-based method to other photo cropping methods, whereby the participants had to choose the most appealing end result of two candidates at a time. The key difference to the previous approaches is that we use choice prior to task completion instead of for rating the end result. In addition to forced choice, we also solicited a variety of subjective ratings. We drew inspiration from the desirability toolkit developed at Microsoft [1, 8]. Attributes relevant to our research question (appealing, busy, complex, attractive, easy to use, and overwhelming) were selected from the desirability toolkit and used along with their opposites as the end points of bi-polar rating scales.

We implemented an application framework to support rapid creation and execution of controlled studies. It facilitates the setting up of the study by using XML configuration files (e.g., search tasks, subjective ratings statements), handles interaction data logging and provides customizable prompts for core activities such as task descriptions and feedback questions. One of the motivations to develop a dedicated software application was the lack of available experimentation tools specific to Web search. We investigated using Touchstone [5] as the experimental platform, but found it unsuitable for our purposes. The experiment design interface is unfortunately not included in the distribution, which means that details of confidential studies would have been stored on an outside host.

In the initial prototype, we implemented support for a basic counterbalanced study design comprised of fixed blocks (tasks performed with a specific design) and forced-choice blocks (participants choose *a priori* which

design to use). Most of the study design logic can be customized within the setup files, making it relatively easy to apply the tool to different search scenarios. One major advantage that this framework provides is that it can be administered in parallel to multiple participants with minimal moderator involvement. The framework creates a unique, counterbalanced set of tasks for each participant. Once started, participants can advance through task and feedback stages by reading directions and following screen prompts.

*Pilot study: Web search results typography*

As a pilot study, we conducted an experiment in which two versions of a Web search results page were compared. The designs differed in terms of their typographical properties, such as font size, color and whitespace around elements. The participants were first exposed to both designs while carrying out two sets of tasks. Then the participants had the choice of selecting which design they would like to use for the next task, in a series of five tasks. Fixed blocks were repeated a second time with the other design to account for learning effects and task-interface interactions. After each block of tasks, the participants were requested to fill in feedback questions related to their preferences (subjective ratings of the design attributes), and after the choice blocks they were also asked which design they would like to use for their own tasks. After the tasks, the participants were presented with a final questionnaire on the perceived differences and preferences related to the designs. The hypothesis was that after having been exposed to both alternatives, the participants would be able to make an informed decision when presented with a choice of designs, a choice that would reflect the user experience afforded by the designs.

### **Methodological findings**

It is not possible to go into too much detail about the results of the particular experiments that we have run. However, we can share some initial findings about the potential strengths and weaknesses of the method. First, any study that attempts to study Web search user experience will have to deal with the issue of branding and prior experience with one of the designs. It has been shown that branding can have a strong effect on users' perception of performance while carrying out web search tasks [4]. As a result, we chose to keep the branding of the two designs tested as similar as possible. The downside to this decision was that we found that when queried after carrying out the tasks, half of our participants were unable to distinguish between the two designs based on their visual characteristics. On the other hand, those participants that could spot the differences did for the most part form quite strong impressions one way or the other.

Second, as the proposed methodology is centered on the forced-choice paradigm, it is important to decide how the design choices are presented. We decided to distinguish the designs with textual labels instead of screenshots when referring to them in questions. This turned out to work relatively well, although some participants had problems recalling which visual style was related to which label. In the future it would be useful to study the effect of using screenshots instead of textual labels. It is possible that if the participants can see the designs before selection, they might make their choice based on something other than their experiences with the interface.

Third, a decision has to be made about whether to have participants choose an interface on per task or per

block of tasks basis. The advantage of per task choice is that we can record more data points per participant. The disadvantage is that the penalty for a “poor” choice is small, making it easy for participants to make arbitrary choices. Per block choice provides less data, but it can increase the validity of the observations. The participants have more at stake with each choice and are perhaps more likely to choose the interface that provides a better experience. Our experiment used per task configuration and the results were mixed. Some of the participants seemed to have no clear preference and made their selections in an alternating way, whereas some participants formed a clear preference, which was also reflected in their subjective feedback.

### Discussion

The driver for the development of this new method was the need to answer design related questions that traditional methods do not adequately address. Based on the pilot study, the forced-choice paradigm seems promising in studying the desirability of interfaces. The combination of usage-based preference and subjective feedback provides us with a better understanding of the effects of design changes on user experience than usability testing or online usage logs alone. Further research is needed to validate the method in larger scale studies. The evaluation method alone cannot answer all questions – it is up to the development team to interpret the results, such as the implications for design when one interface performs worse but is preferred by choice as well as reported preference.

### Acknowledgements

This work was carried out during the first author’s internship with the User Experience Research team at Google, and in part supported by the Graduate School

in User-Centered Information Technology (UCIT). Appreciation is also extended to Aleksandra Sarcevic, Peter Hong, Kerry Rodden, and Jake Brutlag.

### References

- [1] Benedek, J. and Miner, T. Measuring Desirability: New Methods for Evaluating Desirability in a Usability Lab Setting. A paper on the Desirability Toolkit Proceedings of the Usability Professionals Association Conference 2002.
- [2] Frøkjær, E., Hertzum, M. and Hornbæk, K. Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? *Proc. CHI 2000*, ACM Press (2000), New York, NY, 345-352.
- [3] Hornbæk, K. and Law, E.L. Meta-analysis of correlations among usability measures. *Proc. CHI 2007*, ACM Press (2007), New York, NY, 617-626.
- [4] Jansen, B.J., Zhang, M. and Zhang, Y. The effect of brand awareness on the evaluation of search engine results. *Extended Abstracts CHI 2007*, ACM Press (2007), New York, NY, 2471-2476.
- [5] Mackay, W.E., Appert, C., Beaudouin-Lafon, M., Chapuis, O., Du, Y., Fekete, J. and Guiard, Y. Touchstone: exploratory design of experiments. *Proc. CHI 2007*, ACM Press (2007), New York, NY, 1425-1434.
- [6] Santella, A., Agrawala, M., DeCarlo, D., Salesin, D. and Cohen, M. Gaze-based interaction for semi-automatic photo cropping. *Proc. CHI 2006*, ACM Press (2006), New York, NY, 771-780.
- [7] UXmatters Glossary: <http://www.uxmatters.com/glossary>
- [8] Williams, D., Kelly, G. and Anderson, L. MSN 9: new user-centered desirability methods produce compelling visual design. *Extended Abstracts CHI 2004*, ACM Press (2004), New York, NY, 959-974.