User-Centered Evaluation Methodology for Interactive Visualizations

Theresa A. O'Connell and Yee-Yin Choong
National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD, 20899-1070 USA
{theresa.oconnell, yee-yin.choong}@nist.gov

ABSTRACT
This position paper describes a methodology to evaluate innovative, complex interactive visualizations (IV). Our research goal is to develop innovative evaluation methods and sensitive metrics to understand how users interact with IVs. Focusing on experimental design, we highlight ways this methodology and its associated metrics are specialized for IV evaluation.

Categories and Subject Descriptors
H.5.1 [**Multimedia Information Systems**] *evaluation/ methodology*, H.5.2 [**User Interfaces**] *evaluation/methodology*

General Terms
Measurement, Evaluation, Human Factors.

Keywords
User-Centered Evaluation; Interactive Visualizations.

1. INTRODUCTION
The Visualization and Usability Group (VUG) at the National Institute of Standards and Technology (NIST) is conducting research on evaluating interactive visualizations (IV) developed for information analysis. These IVs are highly complex, multi-source, multi-perspective, collaborative interfaces that are aimed at empowering novice analysts to perform information analysis. The information analysts need tools that will facilitate information discovery and exploitation, facilitate trend and relationship discoveries, and support analytical thinking.

There are inherent difficulties in applying traditional formative methodologies to evaluating any new technology. Our approach is to identify the unique aspects of interaction with innovative IVs and to develop evaluation methodologies that address these unique aspects of IVs. One driver behind our research is a lack of metrics that address the unique human-computer interaction (HCI) aspects of IVs. Based on our understanding of our targeted users, their needs and tasks, their workplaces and their work styles, we have started to develop a set of metrics for the purpose of measuring analysts' interaction with IVs. The goal is sensitive metrics that are customized to analysts' needs and experiences, yet still place the burden for usability on the IVs. We are applying these metrics and methods in studies of analysts using IVs to perform typical workplace tasks. To date, our research has focused on prototypes. We anticipate that, with the maturity of the products whose progress we monitor, we will have the opportunity to apply our methods and metrics longitudinally, furthering our understanding of their sensitivity.

2. User-Centered Evaluation Methodology
In this section we discuss the IV usability evaluation methods we use, highlighting ways they differ from traditional usability evaluation methods. Our work most often focuses on the component level, but the metrics we use help us to draw conclusions that encompass the system level. Our methodology starts with understanding analysts' work styles; the demands of their work environments; analysts' typical workplace needs and goal, especially with respect to IVs and IV utility. This knowledge helps us understand how analysts will interact with the IVs.

We perform formative, user-centered evaluations on IVs for information analysis. We have employed our metrics in a variety of rigorous user-centered methods in evaluating IVs. The focus of our current efforts is often the early stage of an IV's development life cycle. Therefore, in many of our studies, we work with IV prototypes. Evaluating prototypes with real users provides valuable user feedback to the development teams. The results identify design aspects that can be improved, and can help the development team set priorities for design changes.

Each evaluation involves five major phases: experimental design, dry run, evaluation, data analysis and reporting. In this position paper, we focus on the experimental design as it is the blueprint for our methodology. The experimental design for each formative evaluation is a rigorous and iterative process with constant reviewing and refining. It starts with defining the evaluation's high level goals, i.e. what we need to learn. The development and NIST teams collaborate closely on defining these goals. Because the tools are complex, so are the goals. Typically, they consist of studying targeted facets of the tool such as its ability to empower analysts to find and assemble evidence, discover and demonstrate patterns, take notes, do sensemaking, form and consider hypotheses, develop insights, and produce a report. Often we must measure the tool's ability to promote collaboration. Each facet that we study tracks to an aspect of our understanding of analysts. For example, we study continuity because we know that analysts' work can be long term, often requires them to monitor different

topics at the same time and that interruptions are common. We state the goals in questions, e.g., Does the IV help analysts remember where they were and empower them to resume work after an overnight break?

We define the user profile specifying the characteristics of representative users for the evaluation. It is essential that people who perform analysis as part of their everyday work participate as users. The user profile specifies experience using analytic tools and performing information analysis.

A key factor to evaluation success is to have realistic tasks and realistic large data sets that require visual analysis. Tasks and data must simulate analysts' real-world experiences. We have learned that, if they are not realistic, analysts consider them inadequate and may transfer this impression to their impression of the tool. We often work with non-subject analysts to develop realistic tasks that state a background and a problem. In the experimental design, we include evaluation and practice tasks.

We then define the methods and metrics. We create a chart to track data collection instruments and metrics to our questions (Table 1). Logging software and sometimes eye-tracking collect objective quantitative data. Metrics must be sensitive to the targeted users and their interactions with targeted IV functionality. The experimental design specifies each measure, metric and data collection instrument as well as the logging output format.

Table 1. Metrics collection.

Data Collection Goal Is to Answer This Question	Data Collection Instrument	Data Collection Metrics
Can analysts take notes?	Logging	• Number of notes created by each analyst • Number of copy and pastes • Average length of hand-typed notes
	Eye tracking	• Total number of fixations (unique fixations, re-fixations) in the notes panel during note taking • Percentage of time devoted to viewing notes panel and scanning its entries
	Exit Survey	• Satisfaction with note taking experience (1-7)
	Observation	• Analysts' comfort level during note taking • Number of critical incidents during note taking

We collect and quantify analysts' opinions in a series of surveys using Likert scales. Analysts complete surveys before training, after training and after interacting with the IV. These surveys also have open-ended questions so that analysts can explain their scalar ratings. We compose interview questions to collect anecdotal subjective data about interactions with targeted functionality. Another good source of data is observations of users interacting with the IVs, by trained professionals in human-computer interaction. We usually use screen capture software to record analysts' interactions and comments. Because most of our studies investigate collaboration among analysts, we do not often use think-aloud protocols as this would intrude on the collaboration.

We record the number and type of critical incidents. Because adoption of new technologies can be frustrating, we expand our definition of critical incidents beyond technical show-stoppers to include any time that the analyst abandons the tool, e.g., stopping work out of frustration or because of strong negative reactions.

The experimental design specifies which of the teams, NIST or developers, is responsible for each part of the study. It includes complete surveys and interview questions. It specifies the evaluation environment, which for IVs can require multiple displays and system upgrades such as powerful graphics cards.

The complexity of IVs makes learnability an important aspect of the evaluation. The experimental design specifies training and training materials. If it is expected that training will be formal with instructors, the developers provide instructors. If the intention is a self-paced tutorial, analysts take the tutorial. At the end of training, we administer a competency test that requires analysts to demonstrate that they can exercise each of the IV's targeted functionalities. We log the amount of time analysts spend learning to use the IV, broken down by training time and practice time. We collect analysts' statements about training and learnability. We ask them how long it would take to acquire the level of expertise necessary to perform tasks in the workplace.

The experimental design includes schedules for the entire evaluation process and the evaluation day. Because the studies often involve collaboration and continuity, the schedules can be very complex, specifying points where analysts must exchange work products or move to a new workstation. Evaluations usually span at least two days so that we can verify if the IV empowers analysts to resume work after an overnight break.

In dry runs, we exercise each targeted function and any required supporting functionality. Then an analyst user undergoes training and performs all tasks. We log interactions and examine the logs to verify that we are collecting needed data. An analyst completes the online survey to verify that the questions will collect the intended data.

A good experimental design can facilitate report writing because it describes every step of the experiment. Because the IVs we evaluate are innovative, the report stresses utility, e.g. it discusses the degree to which the targeted functionality will be useful in the workplace. It discusses readiness for insertion and whether the IV improves the user experience.

3. CONCLUSIONS

NIST has historically employed a variety of rigorous user-centered methods, measures and metrics to achieve the goals of formative evaluations. We feel that this workshop will provide us opportunities to learn different viewpoints from other researchers, as well as to share our own experience.

4. ACKNOWLEDGMENTS

This work was done as part of the Disruptive Technology Office A-SpaceX Program.