JUS
JOURNAL OF USABILITY STUDIES

Vol. 2, Issue 3, May 2007, pp. 102-111

Expanding Usability Testing to Evaluate Complex Systems[1]

Janice (Ginny) Redish
Redish & Associates, Inc.

6820 Winterberry Lane

Bethesda, Maryland, USA, 20817

T: 1-301-229-3039

ginny@redish.net

When you think of usability testing, do you think about working with someone for an hour or two, watching and listening as they do a series of short, discrete scenarios where there is a clear ending or a correct answer for each scenario? For most of us that's a typical usability test. It's the type of usability test assumed by the Common Industry Format (ISO/IEC 25062, see http://zing.ncsl.nist.gov/iusr/). It's the type being discussed for the CIF-Formative (Theofanos and Quesenbery, 2005). It's the type that the various CUE studies have focused on (see http://www.dialogdesign.dk/cue.html).

But not all systems lend themselves to short, discrete scenarios. Not all scenarios have clear endings or known, correct answers. How do we evaluate the usability of systems that are too complex for our typical usability testing protocols?

What do I mean by a complex system?

I am focusing here on complex information analysis: the work that domain experts do when solving open-ended, unstructured, complex problems involving extensive and recursive decision-making (Mirel, 2003, 2004; Albers 2003).

[1] This essay is excerpted and adapted from a paper written in collaboration with Jean Scholtz of the Pacific Northwest National Laboratory (also a UPA member) and presented by Ginny Redish at a symposium on HCI and Information Design to Communicate Complex Information, February 2007 (Redish and Scholtz, 2007).

Complex information analysis takes place in many domains, including the following:

- Store managers evaluating inventory against future needs (Mirel, 2004).
- Corporate and government project managers allocating resources among many projects (Mirel, 2004).
- Nurses dispensing medicine (Mirel, 2004) and many other health care professionals in many situations (operating room systems, intensive care systems, neonatal care systems, patient records, and so on)
- Intelligence analysts bringing together many sources
- Emergency responders prioritizing logistics
- Train drivers (Olsson and Jansson, 2005; Olsson, Johansson, Gullikssen, and Sandblad, 2005)
- Customer service representatives who may have to interpret the customer's presentation of a problem and use multiple sources of information to help the customer
- Many others

How do complex systems differ from those we typically encounter for usability testing?

As Mirel says (2003, 233), "complex tasks and problem solving are different in kind not just degree from well-structured tasks." These complex systems differ from the world of well-structured tasks in at least these ways:

- Information overload is endemic. People must sift through more information than they can deal with. They must figure out how to allocate attention efficiently among an overabundance of information and information sources.
- Data analysis and recursive decision-making are cognitively very burdensome; people have little cognitive workload available for dealing with unusable interfaces.
- Information is often incomplete. It may be unreliable. In some domains, people may have to sort deceptive from meaningful information.
- In some domains, for example, intelligence, there may be no way to know at the time of analysis if the result one gets is right or wrong.
- In many domains, for example in medicine, transportation, intelligence, and the military, time may be critical. Good decisions made too late are bad decisions. And wrong decisions may have catastrophic effects. Getting it right can indeed be a matter of life or death.
- These people are typically domain experts. However, they may not be computer or systems experts. The demands of their work may make it difficult for them to put much time or effort into the learning curve of new programs or new presentation methods.
- Often, analysts and decision-makers are different people. An important question may be how data presentations and complex systems allow the people

searching, gathering, and analyzing information to convey facts and interpretations to decision-makers.

- Visualizations are often a critical presentation method for complex information systems. There is a need, therefore, to study the usability of specific ways of visually representing specific types of data for specific types of users. And there is also the need to study how people develop and use visualizations in the larger context of the work they do. (Scholtz, 2006, calls this larger context the visual analytic environment.) We must evaluate components (individual visualization methods and screens) and the entire system (the environment in which the visualizations are used). I'll talk more later in this essay about the need for this two-level evaluation (components and entire systems).

What else must we consider beyond ease of use?

Ease of use – what we typically focus on in usability testing – is critical but not sufficient for any product. Usefulness (utility) is as important as ease-of-use. If the product does not match the work that real people do in their real environments, it may be an easy-to-use solution to the wrong set of requirements. (This is the main point of Mirel, 2003, 2004.)

Moreover, to understand how people use systems successfully in these complex domains, we may need to include other types of evaluations beyond usability and utility (Scholtz, 2006). We may need to understand and evaluate how well the suite of tools, the presentation methods, and the entire environment support the following:

- collaboration among users (and between users and others, such as decision makers, who may not themselves use the system)
- creativity and innovation (Fischer, 2005)
- interaction (of the user with the same system over time or with a variety of systems that should – but may not – interconnect)
- iteration (the same user returning to the system, wanting to retrieve previous analyses or records, and so on)
- reduction of human error (For a review of various methods proposed to study this issue, see Shorrock and Kirwan, 2002. For an example of applying predictive human error analysis [PHEA], see Parush, et al., 2004.)
- situation awareness (Endsley, 2000; Endsley, Bolté, and Jones, 2003)

Why have we not done more about new usability testing techniques for these complex systems?

Within the usability community, the focus in considering these complex, open-ended systems has, quite correctly, been in pre-design studies – in understanding the domain experts' work. We want designers and developers to get it right beforehand, not to try to fix it through evaluation later.

Pre-design usability studies are absolutely necessary. However, they are not enough. All design and development projects require evaluation as they move from concept to prototype to functioning system. We still need formative evaluation (usability testing)

techniques for complex information systems as they are being designed and developed.

What does this mean for us as usability specialists?

Here are just some of the points we must consider as we expand our usability testing techniques for complex information systems for domain experts:

Collaborating with the domain experts

Most of us are usability or design or communication experts. We are not experts in the domains I am talking about in this essay. And, becoming expert in these domains is not a trivial undertaking. This makes it very difficult to apply user-free formative evaluation techniques in which the usability specialist serves as surrogate user, such as cognitive walkthroughs (Polson, Lewis, Rieman, and Wharton, 1992) and persona-based / task-based / heuristic-based evaluations (Chisnell, Redish, and Lee, 2006).

In all evaluations, working with the client to understand the potential users and their scenarios is very important. In the situations we are discussing here, it is critical. The domain experts must be partners in the evaluation, just as they must be partners throughout the planning, design, and development of the systems.

Collaborating with other specialists

If we, as usability specialists, concentrate on issues of effectiveness, efficiency, and satisfaction, we may have to work with specialists who focus on utility, collaboration, creativity and innovation, interaction, iteration, and situation awareness to get the overall evaluation that is necessary to ensure a successful system. Evaluations like these should not be done in assembly line fashion being passed from one specialist to another but in team work, as a collaborative endeavor.

Furthermore, in many of these domains, security and privacy are major issues. In domains like intelligence and medicine, issues arise about dealing with real data about real events and real people or trying to set up entire data sets with surrogate data.

Getting the right users

We know how critical it is that our usability test participants represent the people who will use the system. We usually worry about getting a false positive result – that the product does fine in usability testing, but when it gets to the users, they have lots of problems with it. A false negative result – that the product shows lots of problems in usability testing, but the real users would not have those problems – is also worrisome. False negatives may be more likely in the systems we are considering here – if we do not have the domain experts as usability test participants.

Relevant issues, therefore, include the following:

- How do we get the time of the domain experts to participate in usability testing?
- How much incentive (money or other) is necessary to engage their participation?
- If surrogates must be used, what are the essential characteristics of the targeted users that we must match?

Getting the right scenarios

In all usability tests, we want realistic tasks. For these complex domains, how will we as non-experts even be able to define good tasks unless we work with the

domain experts? And the tasks must be complex enough to represent the realities of the world in which the systems will actually be used. How do we get the right level of complexity? How do we set up usability testing with the time and environment that is realistic? As Caroline Jarrett says, for these complex domains, teams must "go into the field and trap cases 'in the wild' to use as tasks for usability evaluations" (email to author, 1/16/07).

Understanding how difficult it may be to set goals and tasks

For a usability test of any of these complex systems, we are likely to be able to specify the users and the context of use. The goals, however, will usually be at a higher level than typical usability testing goals, and they may be much harder to specify.

Furthermore, these initial, high-level goals may be vague, such as, "What in this patient's records will help me understand how to interpret this patient's current complaint and relate that to the patient's overall heath?" or "What are we overstocked on and would putting that on sale be good for our bottom line?" or "Is there a trend in this data that I should make my boss aware of?"

These goals (and especially the subgoals to achieve the larger goal) are likely to change as our domain expert moves through the data. Also, our domain expert may be trying out "what if" scenarios using the data to explore aspects of, or possible solutions to, the larger goal.

In the context of the possibly vague and almost certainly shifting goals of a complex information analysis, it may be very difficult to define *a priori* what constitutes effectiveness or efficiency in a given scenario. In almost all information gathering and analysis tasks, people satisfice. They stop at a point where they are satisfied enough with what they have achieved.

Accepting experts' judgment of completion and effectiveness

If we do not ourselves know the entire data set that is being used for the usability test, we may well not know the answer to a given task. We must rely on the usability test participants' judgments that they have arrived at a reasonable solution. In still other domains, again such as intelligence or medicine, the rightness of an answer can only become known over time. What do we set as a measure of success for our usability test?

Doing usability testing of both components and entire systems

Systems for complex problem solving often include many pieces (components, tools). Usability testing at the component level may be possible and very useful for some situations. And, our typical usability test protocols will often work for testing specific components.

However, we must also test the suite of components together at some point because the only true measure is the user's success at solving the problem, however large that problem may be. And, typical usability testing is too short, too "small task"-based, and not context-rich enough to handle the long, complex, and differing scenarios that typify the work situations that these complex information systems must satisfy.

What might we do?

No single methodology or measure is going to work for usability testing of all these systems for domain experts doing open-ended problem solving. As Mirel reminds us (2003, 250): "To be analytically useful, interactive data visualizations have to be designed to allow users to employ and see the results of the analytical methods relevant to the lines of reasoning in their particular area of specialty for a given type of problem. These lines of reasoning are not generic. They are social and contextual." And Mirel's point is valid for all systems, not just those that use interactive data visualizations.

(Scholtz, Morse, and Potts Steves [2006] begin to develop a list of dimensions and factors along which these complex information systems vary. Redish and Scholtz [2007] expand that list.)

For usability testing of complex information systems, some of the techniques we might consider include:

- conducting usability studies outside of the laboratory, for example at conferences where designers, developers, and domain experts meet
- using multiple evaluators to observe different team members in collaborative work
- building simulations (with consideration of how well the simulation captures enough richness and complexity of the real work)
- developing situation awareness assessments (for domains where that is appropriate)
- using think aloud (especially cued retrospective where concurrent think aloud would pose too much additional cognitive workload) Redish and Scholtz (2007) include an extensive discussion of the research on various types of think aloud and implications for usability tests of these complex systems.
- implementing unattended data capture for portions of a long-term evaluation, used along with observations and interviews

What has been tried?

Here are four very brief case studies. They are all from research projects, not from usability testing of commercial systems, but they can give us ideas for expanding our usability testing techniques.

Testing with simulated situations within a typical usability testing time frame

Patterson (1999) reports on a study to find out whether expert intelligence analysts, working on a topic that was not their primary specialty, would find, select, and use the best sources available to answer a given question when they had a large data set of documents and a short time frame. In this case, the researchers knew which were the best documents and what the analysts should report. What they learned was how different analysts searched, what data they kept, how much time they were willing to spend, and how much they relied on their own knowledge compared to using the data in the documents.

Taking advantage of conferences and contests

Contests have been used in several domains as a way of focusing attention and evaluation on components for moving a field forward. For example, the Message Understanding Conference (MUC) was started in 1987 to do qualitative evaluation of the state of the art in

message understanding, and the InfoVis Contest was begun in 2003 to create an Information Visualization Repository of resources to improve the evaluation of information visualization techniques and systems (Thomas and Cook [Eds.], 2005, 152).

Based on the success of contests in these other domains, the Visual Analytics Science and Technology (VAST) conference, held in the Fall of 2006, included a contest as a way of bringing developers together with domain experts. Commercial organizations and university teams that are researching and developing visual analytic tools showed how well (or poorly) their systems worked for a problem set by the contest organizers. The teams whose software did best in a first round where the developers acted as users then got to see how their systems worked for actual domain experts. A domain expert was assigned to each team to use the system to work on a second problem set by the organizers. So this was a sort of usability test at a conference with the added flavor of different development teams competing to be the most effective, efficient, and satisfying system to complete the assigned task. (Grinstein, et al., 2006; www.cs.umd.edu/hcil/VASTcontest06).

A week-long formative evaluation in a real environment

Scholtz, Morse, and Potts Steves (2006) report on a study that included week-long evaluations. Volunteer intelligence analysts participated for two weeks. In the first week, they were trained on the new system and then spent several days on a practice task. After they finished the practice task, they were tested to ensure that they could use the basics of the new system.

They were then given a week to research a particular question and generate a report on it – typical of the work that they do. Analysts were debriefed daily, in person on the first day and through an online form during the rest of the week. Further data was collected through the Glass Box (Cowley, Nowell, and Scholtz, 2005). The Glass Box is software that captures keystrokes, links followed, and search terms used. It can gather a continuous recording of the computer displays along with audio from participants. All entries are time stamped. Users can make annotations at any time. The users here were specifically asked to explain what they were doing any time they were away from the system for 15 minutes or more – in part to capture times when they were gathering data from people or paper or analyzing information offline. (Scholtz [in Redish and Scholtz, 2007] notes that data from the Glass Box can be difficult to analyze as it is in complex relational databases that require scripts to sort out.)

An evaluation like this is more of an instrumented pilot release than a typical usability test, but it did allow the evaluation team to see the entire analytic process, see what features of the software were used and when, see what documents the analysts read, what information they took from which documents, etc.

Formative evaluation in a partnership with domain experts

In the Scandinavian tradition of participatory design (Schuler and Namioka, [Eds.],1993; Greenbaum and Kyng, 1991), domain experts are part of the design and development team throughout the process. Olsson, Johansson, Gullikssen, and Sandblad (2005) describe several participatory projects with domain experts. One of those projects is TRAIN (Traffic Safety and Information Environment for Train Drivers).

In the TRAIN project, the Uppsala University researchers spent three years doing an ethnographic study of train drivers at work. They then involved six train drivers, with a spread of experience, gender, and company they worked for, in iterative participatory exploration and design of a new interface for the engine cab of Swedish trains. System engineers and two HCI researchers worked with the train drivers. (The design phase of the project is discussed in Olsson and Jansson, 2005.)

At the end of the 2005 paper, Olsson and Jansson said that their next step was evaluating the prototype in a simulator, comparing it to the currently existing system and measuring performance and situation awareness. In early 2007, Jansson informed us (email to the author, February 1, 2007) that their first evaluations showed that it is difficult to get a complex enough scenario in a simulated environment to capture the expertise of the domain experts, and, also, that it is difficult to operationalize situation awareness in the train domain. They are working on developing a simulation that is complex enough and that will allow train drivers to act close enough to the way they do on their jobs.

Conclusion

Many of us have expanded our repertoire of usability testing techniques beyond the controlled laboratory setting. Many of us would agree with Joe Dumas (Dumas, 2003) that usability testing today is not a single technique but a spectrum of related techniques. However, the much broader considerations needed to ensure success of complex information systems for domain experts doing open-ended, recursive analysis may require measures and methods beyond even the wider spectrum of usability testing techniques that we have so far developed. These are critical systems. We should be much more involved with them than we have been.

Acknowledgements

I thank Caroline Jarrett, Avi Parush, Jean Scholtz, and Whitney Quesenbery for comments on earlier drafts of this essay.

References

Albers, M. J. (2003) Complex problem solving and content analysis, in M. J. Albers and B. Mazur (Eds.), *Content and Complexity: Information Design in Technical Communication*, Mahwah, NJ: Lawrence Erlbaum Associates, 263-284.

Chisnell, D., Redish, J. C., and Lee, A. (2006) New heuristics for understanding older adults as web users, *Technical Communication*, 53 (1), February, 39-59.

Cowley, P., Nowell, L., and Scholtz, J. (2005) Glass Box: An instrumented infrastructure for supporting human interaction with information. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences* (HICSS'05) - Track 9 - Volume 09.

Dumas, J. S. (2003) User-based evaluations. In J. Jacko and A. Sears (Eds.), *The Human-Computer Interaction Handbook*. (pp. 1093-1117) Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Endsley, M. R. (2000) Theoretical underpinnings of situation awareness: A critical review, in M. R. Endsley and D. J. Garland (Eds.), *Situation Awareness Analysis and Measurement*. Mahwah, NJ: Lawrence Erlbaum Associates.

Endsley, M. R., Bolté, B., and Jones, D. G. (2003) *Designing for Situation Awareness – An Approach to User-Centered Design*. Boca Raton, FL: CRC / Taylor & Francis.

Fischer, G. (2005) Creativity and distributed intelligence, Report of the NSF Workshop on Creativity Support Tools (pp. 71-73), downloaded from http://www.cs.umd.edu/hcil/CST.

Greenbaum, J. and Kyng. M. (1991) *Design at Work: Cooperative Design of Computer Systems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Grinstein, G., O'Connell, T., Laskowski, S., Plaisant, C., Scholtz, J., and Whiting, M. (2006) VAST 2006 Contest – A Tale of Alderwood, *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pp. 215-216.

Mirel, B. (2004) *Interaction Design for Complex Problem Solving – Developing Useful and Usable Software*. San Francisco: Morgan Kaufmann.

Mirel, B. (2003) Dynamic usability: Designing usefulness into systems for complex tasks, in M. J. Albers and B. Mazur (Eds.), *Content and Complexity: Information Design in Technical Communication*, Mahwah, NJ: Lawrence Erlbaum Associates, pp. 233-262.

Olsson, E. and Jansson, A. (2005) Participatory design with train drivers – a process analysis, *Interacting with Computers*, 17, 147-166.

Olsson, E., Johansson, N., Gullikssen, J., and Sandblad, B. (2005) *A Participatory Process Supporting Design of Future Work*, paper from the Department of Information Technology, Human Computer Interaction, Uppsala University, Sweden, retrieved from http://www.it.uu.se/research/publications/reports/2005-018/2005-018-nc.pdf

Parush, A., Erev-Yehene, V., Straoucher, Z., Rugachev, M., Kedmi, E., and Markovitch, R. (2004) *The safety implications of command and control tasks: A comparative study*, Research Center for Work Safety and Human Engineering, Israel Institute of Technology, HEIS-04 (available in Hebrew from the first author at Avi_Parush@carleton.ca).

Patterson, E. S. (1999) A simulation study of computer-supported inferential analysis under data overload, *Proceedings of the Human Factors and Ergonomic Society*, 43rd Annual Meeting, pp. 363-367.

Polson, P. G., Lewis, C., Rieman, J., and Wharton, C. (1992) Cognitive walkthroughs: A method for theory-based evaluation of user interfaces, *International Journal of Man-Machine Studies*, 36, 741-773.

Redish, J. C. and Scholtz, J. (2007) *Evaluating complex information systems for domain experts*, paper presented at a symposium on HCI and Information Design to Communicate Complex Information, Memphis, TN: University of Memphis, February. (Paper available by request to the first author at ginny@redish.net.)

Scholtz, J. (2006) Beyond usability: Evaluation aspects of visual analytic environments, *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pp. 145-150.

Scholtz, J., Morse, E., and Potts Steves, M. (2006) Evaluation metrics and methodologies for user-centered evaluation of intelligent systems, *Interacting with Computers*, 18, 1186-1214.

Schuler, D. and Namioka, A., (Eds.) (1993) *Participatory Design: Principles and Practices*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Shorrock, S. T. and Kirwan, B. (2002) Development and application of a human error identification tool for air traffic control, *Applied Ergonomics*, 33, 319-336.

Theofanos, M. and Quesenbery, W., 2005, Towards the design of effective formative test reports, *Journal of Usability Studies* 1 (1), November, 28-46.

Thomas, J. J. and Cook, K. A. (Eds.) (2005) *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. Piscataway, NJ: IEEE Press.

Janice (Ginny) Redish is President of Redish & Associates, Inc. in Bethesda, Maryland, USA. Ginny has been helping colleagues and clients communicate clearly and create usable products for more than 30 years. Ginny was the founder and first director of the Document Design Center at the American Institutes for Research. Since 1992, she has been an independent consultant focusing on usability of software, documents, and web sites.

Ginny is co-author of two of the major books on usability:

- *A Practical Guide to Usability Testing* (with Joseph Dumas, Intellect Books, 1999)
- *User and Task Analysis for Interface Design* (with JoAnn Hackos, Wiley, 1998).

Her new book on writing for the web: *Letting Go of the Words – Writing Web Content that Works* (Morgan Kaufmann, 2007), will be available at UPA 2007.

Ginny is internationally recognized as a dynamic speaker and workshop leader. She keynoted the UPA Conference in 2004 and has trained hundreds of writers, usability specialists, and subject matter experts in user-centered design, specific usability techniques, and clear writing.

Ginny is a graduate of Bryn Mawr College and holds a Ph.D. in Linguistics from Harvard University.