Beyond Usability: Designing for Evaluating Experience

Joseph 'Jofish' Kaye
Cornell University Information Science
301 College Ave, Ithaca NY 14850 USA
jofish at cornell.edu

INTRODUCTION

My recent work has involved building systems for communicating intimacy between couples in long distance relationships. This has a number of difficulties: we don't know how to measure intimacy, we don't know how to communicate intimacy, and we don't know how to measure if or how much intimacy we're transmitting. All of these make it difficult for to evaluate these systems. Figure 1

Figure 1. Measuring intimacy: Ariel loves Eric 19 hearts.

shows a cartoon version of the way this evaluation *doesn't* work: if only it was as easy as just counting the hearts.

Submission for DIS 2006 Workshop: Exploring the Interrelationships between the Design and Evaluation of Interactive Systems

More than anything else, this work emphasizes the fact that *usability* alone is an insufficient topic for evaluation: it is necessary to evaluate for a richer and deeper understanding of experience.

EVALUATION

The standards of evaluation used in traditional HCI have descended from its legacy in designing efficient user interfaces, such as the human factors work that arose from building cockpits for military pilots. Evaluation methods are favored that give reproducable, rigorous answers to questions such as the optimal spacing or ideal sizes of buttons. (See [8] for a survey of applications of the GOMS model, for example.)

Recently, some HCI researchers have become concerned with a fuller expression of the user experience, one that values the phenomenological, felt experience of the user, rather than the psychophysical, analytical reaction. [2,13] How, then, to gather adequately rich descriptions of these experiences as so to inform evaluation sufficiently as to be able to understand the users' experience?

In this short paper, I hope to sketch out some of the work I and others have been doing to attempt to answer these questions. In particular, I want to disucss two approaches to describing others that influenced our strategies in designing evaluation, and detail two attempts to evaluate a particular piece of software.

Cultural Probes

Gaver et. al.'s article on cultural probes [4] has been one of the most influential articles in recent HCI history. We were inspired by both the rich results the designers got back from the probes, but also by the degree of involvement the phobes encouraged from the participants.

Gaver et. al. provided potential stakeholders in technological changes to an area with packets containing pre-addressed postcards with leading questions, maps to be filled out, and cameras with a list of requested images. Participants filled out the postcards and maps and took the photos over the course of a few weeks, and returned the results to the design team. It's important to note that Gaver et al. were not trying to solve questions of evaluation; they were trying to provoke inspirational responses from potential users, for designers to use in building appropriate technologies. The variety of materials and open-ended nature of the tasks subjects were asked to do with the

materials gave rich, situated answers for the deisgn team to work with.

Gaver et. al. were using probes to inspire design, and we are using them to inspire evaluation. This phraseology seems a little strange: we're used to thinking about design as something that needs to be inspired, whereas evaluation is something that's comparatively passive, coming at the end of a development cycle. But thanks to, among others, Dunne & Raby's evaluation of the Placebo project [3] Gaver & Dunne's informal invaluation of the Presence project [4], and Höök et. al.'s evaluation of an interactive art piece called the Influencing Machine [7], we've come to see evaluation as something that requires creativity and inspiration as much as design does, rather than an objective, passive and measured response to rational criteria. [9]

Thick Description

Another inspiration is Geertz's notion of *thick description*. [6] In essence, thick description requires a detailed account of the culture and the context around a specific action. He uses the example of winking: with thin description, an eye twitch and a wink are the same. It's only the thick description of the context and culture that lets us understand the role of the wink in sharing a conspiracy, or even parodying another sharing conspiracy. Thick description is about supplying the context along with the content to facilitate understanding of the experience.

Our difficulty with much of traditional evaluation techniques is that they were designed for tighly controlled, thin description. This allows for rational optimiziation, but doesn't leave room for *culture*, which we feel is key in building truly relevant, pervasive computing devices. We do recognize that truly thick description as Geertz describes it requires years of observation, and traditional ethnographers and anthropologists would no doubt recoil in disgust at our description of our subjects responses as thick description, but our point of comparison is the hour-long user study, not three years in a mud hut, and we'll take the best we can get.

IMPLEMENTATION

A topic for evaluation

Our object of study is a piece of software called the Virtual Intimate Object, or VIO, which is designed for couples in long distance relationships to communicate intimacy. [10] Both couples install the VIO software, which appears as a circle in the taskbar of a Windows machine, or in a small window of its own on a Mac. When one member of a couple clicks on their VIO, the other member's VIO turns bright red, and then fades over time, initially rapidly and then slowly, returning to transparency after twelve hours. This process is shown in Figure 2. It's also possible for users to check on the current appearance of their partner's circle by moving their mouse over the VIO without clicking, also shown in Figure 2.

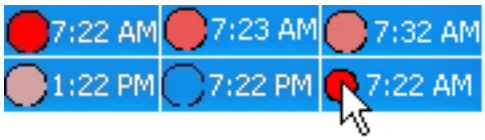

Figure 2: VIO, showing change over time.

The great advantage of this piece of software is that it's extremely simple. It uses minimum bandwidth, and within its very limited boundries, it's extremely configurable: the colours, fade times, and icon choices can be changed as much as the researcher desires. There's no inherent usage scenario built into the software: we use it to tell a story about couples in long distance relationships, but it could be used for any other pair of people or devices needing to communicate in a minimal manner, or even scale to a one-to-many or many-to-many model depending on the scenario. As such, we feel it's an excellent system for evaluation.

User Studies

We did two rounds of user studies on the VIO. In the first, we provided subjects with a printed logbook to fill out over the course of their first week using the software; in the second, we attempted to replicate our first results using an online software survey system on a larger sample size. In each case, we used a similar set of criteria.

In all cases, our questions fell into one of three categories:

- questions about the technology under evaluation
- questions about the relationship that we hoped would be affected by the software
- questions about the survey itself.

We made an serious effort to ensure that these questions were provocative and open ended. Our hope was that these questions would serve to defamiliarize participants with their standard mental constructions of their relationships, and encourage reflection. [1,11] For the first category, questions included

- What's the thing you hate the most about using the VIO?
- Draw a picture of your ideal intimate object.
- What was your favorite movement involving your VIO?

Questions about the relationship were if anything even more open ended, and included

- What colour is your relationship? Why?
- What's the nicest thing your partner has done for you since you've been in a long distance relationship?
- What season is your relationship? Why?

We also asked questions about the survey itself. We felt this was important to try and understand the role of the survey itself in gathering information about the user experience, much as ethnography has come to recognize the ethnographer as actively participating in the experience of gathering her ethnography, not merely as a passive observer, such as in [12]. These questions included

- What question should we have asked you to understand your use of the VIO?
- What question do you remember answering earlier in these surveys?
- What should we do differently with these surveys?

Methodology

Participants initially filled out a reasonably long questionnaire that asked them questions about their relationship: how far away from their partner were they, why were they in a long distance relationship, how often did they talk or use other forms of communication. They were then given a link to download the software.[1]

Once the software was installed, each partner was asked to fill out one page of their logbook a day. In the first study, they were asked to do this for one week, and in the second study, they were asked to do this for two weeks. Each day followed an identical format. First, subjects were asked how many times they had clicked their VIO, and how many times they thought their partner clicked their VIO. Next, subjects were asked to rate the following statements on a scale of one to seven:

- How close do you feel to your partner today?
- How satisfied do you feel by your relationship today?
- How connected do you feel to your partner today?

and were then asked to explain one of their answers to these questions. They were then asked to do the same with the following questions:

- How big an impact did your frequency of VIO use have on your partner's day?
- How positive do you feel about your VIO today?
- How interested do you feel in your VIO today?
- How comfortable do you feel about your VIO today?

Finally, subjects were asked three to five of the open ended questions listed above, which we feel were the key point of the probes. (In fact, in the first survey, we found no statistically significant results from these questions. The decision to keep them in the second version was because we feel they provide a counterpoint to the open-ended questions, making the questions seem more inspiring than they would otherwise be – although that's a hypothesis we haven't addressed directly.)

We made an effort to distribute the three kinds of questions evenly across the testing period, although we consciously retained questions that would be improved by significant VIO use (such as "What was your best experience with the VIO?") to the end.

After the testing period ended, we conclude with a follow-up questionnaire that was extremely similar to the pre-test questionnaire, with the addition of some more questions about their experiences of the VIO.

LEARNING FROM OUR MISTAKES

Our initial survey was small: we had five couples using the VIO. On the second, we placed requests for help on various communities of couples in long distance relationships on LiveJournal, a blog hosting site, after requesting permission from the community organizers. Our response was very good: we ended up with approximately eighty people completing the survey pre-test.

We had received criticism from CHI reviewers and others about the brevity of our one-week study, and decided to try to address this by asking our subjects to do a two-week long study. We found that two weeks of daily questions was just asking too much of subjects recruited without a deeper tie to the research: less than ten percent of our subjects replied to the final questionnaire two weeks later.

We also felt there was a fundamental problem with the online surveys, which couldn't be overcome with design. Receiving a cultural probe *feels like being given a gift*. There's something exciting about receiving a cultural probe, and you want to do your best to make the people who put it together happy. You want to go and take the pictures, fill out the postcards, draw the maps, because it's a way of saying thank you back to the researchers who provided all those toys. There's a novelty to even the mundane, be it Japanese-themed notebooks, or a personalized camera with your name on it, or Finnish tubes of glitter glue (with left-hand threads!) Without the social pressure in response to the gift, it's hard for even the most enthusiastic recipient to consistently provide answers.

Finally, we're still trying to understand the role of interaction between the quantitative and qualitative in these studies. We had two kinds of qualitative data: the answers to the 7-point scale questions (which gave no statistically significant data in the first study), and the server logs. The design of our software meant that each interaction with the software was logged by our server: we could tell whenever users clicked their VIOs, and we can even see when the VIO is running, even if it's not being used. In the first study, we were able to compare click rates with, for example, answers to specific questions. (Notably, our couple with the lowest number of clicks per day was also only couple who replied 'winter' and 'fall' to *What season*

[1] Interested readers may wish to install the software themselves: it is available from http://io.infosci.cornell.edu/

is your relationship?) I'm currently in the process of looking the correlations between the server logs and the qualitative and quantitative survey data in the second study.

ACKNOWLEDGMENTS

I thank my subjects for their generous donation of their time, effort and energy. I have developed these ideas in the course of a variety of discussions and talks, including the Workshop on Innovative Approaches to Evaluating Affective Interfaces at CHI 2005, the Less is More Conference at Microsoft Research UK, the HUMAINE workshops at SICS in January 2006, a presentation at the Viktoria Institute, Gothenberg, and a presentation at alt.chi at CHI 2006. I'd also like to thank Phoebe Sengers for her consistent support, and the members of the Culturally Embedded Computing Group.

The first study was done with Mariah K. Levitt, Jeffrey Nevins, Jessica Golden and Vanessa Schmidt, with advice from Kirsten Boehner and Jeff Hancock.

REFERENCES

1. Bell, G., Blythe, M., and Sengers, P. 2005. Making by making strange: Defamiliarization and the design of domestic technologies. *ACM Trans. Comput.-Hum. Interact.* 12, 2 (Jun. 2005), 149-173.
2. Dourish, Paul. (2001) *Where the Action Is: The Foundations of Embodied Interaction.* Cambridge, MA: MIT Press.
3. Dunne, A. & Raby, F. (2001). *Design Noir: The Secret Life of Electronic Objects.* Basel, Switzerland: August /Birkhaeuser.
4. Gaver, W. & Dunne, A. (1999). Projected Realities: Conceptual Design for Cultural Effect. *Proc. CHI 1999.*
5. Gaver, B., Dunne, T., and Pacenti, E. 1999. Design: Cultural probes. *interactions* 6, 1 (Jan. 1999), 21-29.
6. Geertz, C. (1973) *The Interpretation of Cultures*. New York: Basic Books. Chapter I: Thick Description: Toward an Interpretative Theory of Culture.
7. Höök, K., Sengers, P., Andersson, G. Sense and Sensibility: Evaluation and Interactive Art. *Proc. CHI '03.*
8. John, B. E. and Kieras, D. E. 1994 *The GOMS Family of Analysis Techniques: Tools for Design and Evaluation.* Technical Report. UMI Order Number: CS-94-181., Carnegie Mellon University.
9. Kaye, J. 'J.' *I just clicked to say I love you:* Rich evaluations of minimal communication. *Ext. Abs.. CHI'06.*
10. Kaye, J. 'J'., Levitt, M. K., Nevins, J., Golden, J., and Schmidt, V. 2005. Communicating intimacy one bit at a time. In *Ext. Abs. CHI '05*
11. Sengers, P., Boehner, K., David, S., and Kaye, J.'J'. 2005. Reflective design. In *Proc. Critical Computing '05*. 49-58.
12. Tsing, A.L. (1993) *In the Realm of the Diamond Queen: Marginality in an Out-of-the-Way Place*. Princeton, NJ: Princeton University Press.
13. McCarthy, J. & Wright, PC, (2004). *Technology as experience*. Cambridge, MA: MIT Press.