

Reality Testing Mobile and Entertainment Applications

Gitte Lindgaard¹, Bruce Tsuji², Shamima Khan³

¹²³Human Oriented Technology Lab, Carleton University, Ottawa, Ontario, Canada K1S 5B6

¹gitte_lindgaard@carleton.ca; ²btsuji@connect.carleton.ca; ³schoolandwork@gmail.com

With a focus on ecological validity and behavioural measures, our interest is in reality testing of applications based on new or emerging technologies. In particular, mobile and entertainment applications are simply not amenable to classic laboratory testing. Furthermore, standard usability measures of efficiency and effectiveness are not always relevant to these new situations and environments. The challenge is to balance the desire to isolate the effect of individual variables against the business demands of providing meaningful (ecologically valid) and timely feedback to development groups. Another significant issue in usability testing is identifying behavioural measures that capture the right phenomena and yield the information required. Finally, many of the domains of greatest interest are exactly those that have the most severe restrictions on industrial confidentiality. Thus, our opportunities to learn from each other are substantially reduced.

Keywords: ecological validity; usability testing; behavioural measures; confidentiality; industrial secrets; reality testing; new applications; new products

Projects involving mobile and/or novel technologies tend to severely challenge our ability to design ecologically valid [2] usability tests. In this paper, we discuss two recent examples of such projects that we have worked on. One concerned a walkie-talkie-like service running on cell phones, and the other was destined for home entertainment. Both were tested at an early stage of design and both were new to the world, so our tools were non-existent or unreliable prototypes. In addition, due to the novelty of both products, participants were completely unfamiliar with the functionality and features these provided.

The difficulties of obtaining valid user opinion/acceptance data on novel and mobile products are as real as obtaining reliable performance data. Very little is available on this subject in the public arena [4][12]. One major challenge is to create the 'right' usage context(s). For example, the problems of obtaining data from tests of mobile usage are rarely solved by using a mobile usability laboratory and conspicuous video/audio equipment. Setting up varying 'realistic' conditions such as differences in lighting, in noise levels, or in the degree to which space may be crowded, is almost impossible in the laboratory. Yet, when the technology under scrutiny is not already available in the market place, it may also be impossible to field test reasonably advanced prototypes. Often it is even hard to know or predict the kinds of behavioural performance data one should collect, and short of performing longitudinal studies, it is unlikely that the data would reflect actual future usage patterns because people's usage habits change as they evolve.

In one set of experiments seeking to determine people's tolerance for delays in a push-to-talk technology over a cellular network, we imagined how people would respond in behaviourally measurable ways when their level of tolerance was exceeded. As is customary in psychology, we thought we could identify the point at which 'enough is enough' where people lose their patience with a system characterized by highly variable delays. The product was a new-to-the-world cellular service that enabled users to enjoy walkie-talkie-like capabilities from their cell phones [e.g. 9]. Users could press a cell phone button to initiate a 2-way voice call with a pre-determined destination. The push-to-talk over cellular (PoC) service is analogous to a voice version of Instant Messaging. Unfortunately, the nature of the PoC implementation meant that a response time delay of as much as 20 seconds might ensue after the initial button press. We asked two important questions: first, would users find a response time delay of that magnitude acceptable, and second, would they tolerate huge variability in the system response time? Emergency response workers were among the intended audience for the service. Clearly, much of their work is performed under very stressful time- and safety-critical circumstances.

We predicted that 'impatience' would manifest itself in rapid, repeated button presses. Thus, the tolerance threshold could be measured by logging button presses and observing the function of these over a range of different delays and by using different scenarios in which the participants were co-actors. In order to log the button presses and identify the threshold-function, and because it would be impossible to test the product in real-life

emergencies, we were forced to carry out the tests in the laboratory [5][6]. In order to observe differences in the patterns of button presses, we tried artificially to create situations that would be perceived as 'critical' or 'uncritical'. These attempts completely failed. Even priming our university participants by showing a very dramatic, realistic video clip before commencing the task to get them excited, did not evoke the predicted responses. The 'conversation' in which the participant was subsequently to engage using the technology was crafted to suit the particular situation depicted in the video. Yet, participants never exhibited the predicted behaviour [7] although such priming has been successfully used by social psychologists for decades. It is unclear if this lack of success is due to our choice of behavioural measure – we have no evidence suggesting that impatience is actually expressed via increased button presses – or if it is simply due to the sheer artificiality of the laboratory setting.

In another set of experiments we tested a novel type of digital TV service offered over telecommunications networks as opposed to cable or satellite [e.g. 3]. While the new implementation promised hundreds of channels it also meant that users could expect delays of as much as 10 seconds when changing channels by pressing a channel up or down key [8]. Thus, we were again faced with a new service unknown to users, and again the task was to determine the point at which users will lose their patience. As before, our challenge was to identify one or more behavioural measures that would capture the difference between a relaxed and patient user and an annoyed, impatient user.

Since our first set of experiments showed that people either did not reach that level of annoyance at which their loss of patience becomes obvious, or they were too polite to show it, we knew that even artificial situations depicting very serious emergencies would not work, and that we would not be able to observe the behaviour we had identified as being indicative of such a change. Instead of attempting to predict changes in users' behaviour and collect behavioural data, we decided to vary the delay times systematically between and within blocks of trials so that delays were constant in some trials and variable in others. After each trial, which comprised 15 channel switches, we administered a set of standard Mean Opinion Scores (MOS). On one half of the occasions, participants entered the number of a channel highlighted on the screen, and on the other half, they stepped to the highlighted channel by using the up or down key. They then watched the relevant program for a few seconds before the experimenter, who controlled the program, moved to the next stimulus, again a certain channel highlighted on the screen. As the delays increased, user satisfaction dropped, but there was no difference between variable and constant delays. Thus, we learned very little about the point at which people lose their patience with variable and/or very lengthy delays in the system response time.

We believe our spectacular failure may largely be due to the lack of ecological validity, or face validity, a factor that has long been recognized as a major challenge in experimental laboratory research involving thinking human beings [10]. The fact that participants agree to take part in an experiment, go to a laboratory to perform certain more or less meaningless actions, are observed throughout a session, and are paid at the end, is known to impact motivation [1]. Participants typically try to guess the experimental purpose and behave in ways they believe will support the experimenter's hypothesis as well as trying to impress the experimenter by performing as fast and accurately as they possibly can, which may say nothing about the same unobserved behaviour performed in a natural setting. These add up to what Orne [10] called "demand characteristics". Availability of sufficient monetary and human resources can occasionally alleviate some of the artificiality of laboratory research, especially in research involving time-, safety-, or mission-critical applications, for example studies of passenger evacuation in large aircraft [e.g. 11]. Performing ecologically valid and reliable laboratory research that can effectively generalize to field settings with limited resources and short time frames remains a significant issue for human factors research.

Critical issues we would like to discuss

- When new applications/new products have no precedents and there are no ways to test using similar products or services, how do we decide what to measure?
- How do we meet the challenge of mimicking a small mobile device with a novel service that is not yet available?
- What kinds of data should ideally be collected in environments such as those described here?
- How can we be sure the data we decide to collect are meaningful and able to address the topic under consideration?
- How might we provide indicators of future commercial success in addition to gathering valid usability data?
- How can we make up for a definite absence of ecological validity when we are forced to test products/services in a lab?

About the authors

Lindgaard and Tsuji have between them approximately 50 years of relevant experience in wired and wireless telecommunications, business intelligence, data mining, medical/health applications, CRM, data security, e-commerce, consumer electronics, streaming media, and other domains. They have substantial university and industrial experience and have worked as HCI practitioners, instructors, consultants, managers, as well as clients of HCI services. Dr. Lindgaard is a full professor in the Department of Psychology, Carleton University, the Director of the Human Oriented Technology Laboratory at Carleton University and she holds the NSERC/Cognos Chair in User-Centred Design. Bruce Tsuji is a PhD candidate at Carleton University. He has been a user interface designer, usability tester, product manager, and has held executive positions in marketing and sales. Bruce is a co-inventor on seven Canadian and US patents and his awards include a best-in-show from the Consumer Electronics Show as well as a PC Week Analysts' Choice Award.

References

- [1] Adair, J.G. (1984). The Hawthorne effect: A reconsideration of the methodological artefact. *Journal of Applied Psychology*, 69(2), 334-345.
- [2] Brunswik, E. (1955). Representative design and probabilistic theory, *Psychological Review*, 62, 193-217.
- [3] <http://www.instat.com/press.asp?ID=972>
- [4] Lee, I., Kim, J. & Kim, J. (2005). Use contexts for the mobile internet: A longitudinal study monitoring actual use of mobile internet services, *International Journal of Human-Computer Interaction*, 18(3), 269-292.
- [5] Lindgaard, G., Khan, S, & Tsuji, B. (2004a). Push-to-Talk over Cellular (PoC): *A Review of Relevant Literature*, Unpublished HOTLab Technical Report HCIGLSKBT04POC03050527.
- [6] Lindgaard, G., Khan, S, & Tsuji, B. (2004b). *HOTLab observations of the Telus Mike/Direct Connect (PoC) Service*, Unpublished HOTLab Technical Report HCIGLSKBT01POC01050527.
- [7] Lindgaard, G., Khan, S, & Tsuji, B. (2004c). *Initial call setup delays in Push-to-talk over Cellular (PoC)*, Unpublished HOTLab Technical Report HCIGLSKBT02POC02050527.
- [8] Lindgaard, G., Khan, S, & Tsuji, B. (2004d). *Channel switching delays in digital TV systems*, Unpublished HOTLab Technical Report HCIGLSKBT03DTV01050527
- [9] <http://www.motorola.com/content/0,,2038-4398,00.html>
- [10] Orne, M.T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications, *American Psychologist*, 17(11), 776-783.
- [11] Wilson, R., Thomas, L. & Muir, H. (2003). *VLTA emergency requirements research evacuation study*, Work Package 3, Task 3.1 Report – Test plan for evacuation trials, European Commission, VERRES Project.
- [12] Zhang, D. & Adipat, B. (2005). Challenges, methodologies, and issues in the usability testing of mobile applications, *International Journal of Human-Computer Interaction*, 18(3), 293-308.