Towards Lightweight Digital Library User Evaluations

George Buchanan
Centre for HCI Design
City University
London, United Kingdom
+44 20 7040 8469
George.buchanan.1@city.ac.uk

ABSTRACT
Human Computer Interaction (HCI) research has prioritized the evaluation of systems by values consistent with psychological practice, aiming to achieve high reliability for relatively subtle effects. For many digital library (DL) evaluations, the aim is to identify major effects for practical working purposes: e.g. to iterate a design process. Timeliness and cost are often key criteria for selecting a study method in such circumstances, and again psychological standards are in excess of what is required. However, lightweight methods must retain methodological soundness and aim to achieve results with known and planned shortcomings.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: User Issues

General Terms
Design, Experimentation, Human Factors.

Keywords
Digital libraries, iterative design, lightweight evaluation

1. INTRODUCTION
Traditional human-computer interaction methods have increasingly borrowed from the domain of academic psychology. Whilst this is a welcome boost to the reliability and soundness of core HCI research, it is the contention of this paper that for many engineering purposes, this pursuit of psychological rigour is misplaced.

To conduct an experiment that reflects the concerns of classic research in either physics or psychology requires a great degree of skill. Not only have hidden, controlled and uncontrolled variables to be accounted for, but also the experimenter must understand the intricacies of a number of methods of experimental design (e.g. questionnaire, interview), computer programming (e.g. to manufacture detailed logs), statistical and qualitative analysis. To achieve a high level of ability in all these is a major intellectual undertaking [3].

When designing an operational digital library system, to take one example, a librarian may certainly lack complete mastery of such skills, but also must conduct an experiment with limited resources of time and money. Whilst ideally grounded experimentation should guide a high-quality design process, realistically grants often are in the region of thousands or tens of thousands of dollars. When a good study would cost in the region of five to ten thousand, this represents a burden that would cripple the core project. Therefore, some modification must be practically made.

Similarly, for a PhD student in digital libraries, particularly from a software engineering background, many evaluations are pilot studies to scope future work and to validate the basic soundness of core parts of the work. A full understanding of user cognition is not yet required, and much of the context of use may be known from the prior constraints and scoping of their program of research. Again, the student will probably neither possess nor seek to possess the full range of HCI or CHI evaluation methods.

In this position paper, I argue for some simple scoping and boundary setting for practical user evaluations that simplify the variables that could affect an experiment and that should produce moderately reliable (approx. 90% statistical reliability) for major effects in a digital library. These parameters are calibrated from ten years' experience of conducting HCI experiments on DL systems, and reflect the experience of an HCI academic who has worked in the context of DL engineering practice and research.

2. SCOPING A STUDY
Many practical DL studies aim to prove that one system or design is "best". In academic terms, an unqualified "best" is readily criticized as naïve. For the practical purposes of conducting a meaningful HCI experiment, such a broad-ranging goal is unworkable. "Best" or "better" must be turned into precise and particular measures that can be observed in a user study.

Similarly, many libraries face proving to funders that their system is "usable" or "effective" for an anonymous set of users. Again, this implies a wide and poorly controlled scope that contains far too much variance to provide good quality results in a small study.

If a user evaluation is to have any underlying reliability, these issues of scope and variance need to be addressed ruthlessly. For example, a collection of old photographs from a city's history may well be of interest to a host of potential users: enthusiasts for old photographic methods, family historians, local historians, those who like pictures of old trains, or a schoolteacher preparing for a class. Each of these types of users will address their information goals in a different way, influenced by their experiences doing other tasks in other libraries, and the skills that they developed through this accidental "training". To conduct a meaningful study with a small set of users, then there is more likelihood of achieving a reliable result when focusing on only one type of user.

In the field of HCI, the ideas of "personas" and "scenarios" has proven popular in recent years. Fictional composite of known persons (e.g. an elementary school teacher with limited IT skills) are created – the personas – and described in the context of their

likely use of a system (e.g. of a DL of their local town in preparing a class on the Depression of the 1930s). Whilst fashion can be misleading, librarians often have a good understanding of some of their patrons, and can readily identify some relevant interests and abilities of some potential users of any library.

Whilst funding bodies may seek evidence of a big impact on all users, this is simply not achievable without very large numbers of participants. As already stated, individual and group variance means that each group would need many representatives participating in a study, resulting in a large total. Similarly, covering all potential uses for even one client base raises many variations, and again becomes arduous.

So, the natural conclusion of this reasoning is that identifying one set of patrons (e.g. elementary schoolteachers) and then testing their particular needs against a system is more likely to result in meaningful data with a smaller number of participants. In experimental terms, we usually contrast this as being a choice between "internal validity" – the results are certainly correct for the group being tested – against "external validity" – that the results can be projected as being trustworthy for other groups.

If a particular group is identified, then recruiting participants can be simplified by a targeted campaign on places where the target group is often found (e.g. the local history society for family historians). Once some volunteers are found, they can be used to encourage their peers to participate. Likewise, the skills and goals of their potential use are more straightforwardly identified and planned for in creating a user study.

3. SCOPING TASKS

Just as who is doing what – at an abstract level – can be used to limit the work of a study, similarly task selection is critical to producing a good quality evaluation. Users let to browse around a collection will have such a variety of experience that it is not at all clear what is exceptional and what is systematic. HCI of course uses defined tasks to reduce such variables, and lightweight techniques should seek to do the same.

Small-scale user studies cannot identify subtle effects of low magnitude, so tasks should focus only on those parts of the design that are known to be critical to the user and/or believed to have defects (or differences between two designs).

4. SCOPING ANALYSIS

In addition to reducing the scope of the study in general, and the specific tasks, a lot can be done to limit the analysis work at the end of the study. Presuming that a tight focus has been determined for the study design, the spectrum of possible outcomes will in most cases have been reduced substantially. This has significant outcomes for collecting information and analyzing it.

For example, in the case of an observer studying the user's activity during a study session, a simple ticklist can be prepared for the observer to note instances of a particular action (e.g. a user making a mistake in selecting from a list), and similar strategies can be used during an interview (e.g. if an interviewee expresses an inability to use advanced search options). This approach can make encoding qualitative data much more rapid during a study, and analysis quicker afterwards. This rather rigid method in fact reproduces the concept of "coding" in more rigorous qualitative methods, but identifies some key codes in advance.

Even simple quantitative analysis can be fast-tracked: for many comparisons, the same spreadsheet can be reused as a template for more than one study. Modern versions of Excel, for example, can easily produce simple t-test and Chi-squared confidence data, and a template will minimize the possibility of error and the work involved.

5. GENERAL OBSERVATIONS

HCI has increasingly sought to produce results with levels of statistical reliability in the range of 95 to 99%. Heuristics have emerged of needing more than twenty or ideally in the order of fifty participants. However, these heuristics are again based on relatively small effects. Comparing a blue whale with an ant – for the sake of argument – certainly does not require ten of each, or even a miniature poodle with a Great Dane. When confidence levels of over 80% are acceptable, numbers of ten to a dozen participants are often more than sufficient to achieve useful results – and this certainly applies in many practical situations.

Similarly, using all participants on each of two interfaces – if two designs are compared – can be done well using a careful Latin-squared design, where order and learning biases can be countered. Given the effort of recruiting volunteers, I see little value in not using a "between subjects" design where more information is gained from each person.

6. DISCUSSION

Clearly this method of producing short, sharp studies has its limits. The aim is to potentially obtain some modestly reliable outcomes. For academic pilot studies and practical circumstances where the big picture is often well understood, arriving at high confidence is a misleading goal. A high price may be paid for a study of modest strategic value.

A series of lightweight evaluations may ultimately, in any case, yield a compound set of evidence that provides a final conclusion as sound as any one large-scale study. In digital libraries, often our problems are clear and have very visible impacts that can be measured by simple apparatus. We need to deeply consider whether in all events gold-standard HCI studies are required – and if not, in which circumstances.

In my own research, lightweight studies have proved an effective first step in a larger research process (e.g.[1]) and small groups of a dozen participants have given very clear performance outcomes (e.g. [2]). Relying on heuristics (e.g. so-many participants required) is leading us into often self-defeating traps when so many problems are large-scale and readily detected by simpler means.

7. REFERENCES

[1] Buchanan, G., Pearson, J. 2008, "Improving Placeholders in Digital Documents", Procs. European Conference on Digital Libraries, Spinger, 1-12.

[2] Buchanan, G., Owen, T., 2008, "Improving skim reading for document triage". Procs. Symposium on Interactive Information in Context (IIiX), BCS, 83-88.

[3] Cairns, P.A., Cox, A., "Research Methods for Human-Computer Interaction", Cambridge University Press, 2008 ISBN 9780521690317