A pragmatic framework for selecting empirical or inspection methods to evaluate usability

Paul Englefield
Usability Competency Centre, IBM United Kingdom Limited,
PO Box 31, Birmingham Road, Warwick, Warwickshire, CV32 5JL
paul_englefield@uk.ibm.com

Abstract
Within the literature of human-computer interaction there is a vigorous debate on the relative merits of two classes of evaluation methods; those that carry out an empirical study of users' task performance and those that employ experts to inspect a design. The central themes in this debate are effectiveness and cost-efficiency. While these concerns are important in commercial usability work, an analysis of project goals and constraints may be more useful in selecting and justifying methods. This paper proposes a model for selecting methods based on the above criteria. It also suggests that, as inspection methods are more grounded in the concerns of practice than those of science, their selection should also be influenced by factors arising from practice.

Keywords
Methods, empirical, inspection, practice, comparison

Empirical methods
Empirical methods rely on observing and measuring representative users attempting typical tasks. For example, in a *user test*, a researcher might: a) recruit a sample of participants from a demographic profile supplied by the client, b) invite each participant to use the proposed design to complete a set of tasks designed to exercise critical aspects and c) observe the user's behavior while collecting performance data such as task success or subjective satisfaction.

Empirical studies are grounded in the methods of science. Researchers are typically concerned with issues such as reliability and validity. Comparisons can be made with exploratory and experimental studies in the behavioral sciences.

Expert methods
Expert methods rely on skilled practitioners assessing a design proposal within a structured process. For example, in a *heuristic evaluation* [Nielsen, 1994], a panel of experts inspects a design with respect to a set of heuristics in order to identify, categorize, and prioritize design errors.

By contrast to empirical methods, expert methods are grounded in the methodology of practice rather than science. Researchers are typically concerned with rigor, coverage and judgment. Parallels can be identified with techniques such as inspections in software engineering, assessment in education, editing in publishing, and auditing in accountancy.

Academic studies of relative effectiveness
Although there is evidence suggesting that heuristic evaluation alone is comparatively effective and economical [Jeffries et al, 1991], other research: a) questions the validity of this research approach [Gray and Salzman 1998], b) considers the risk of false positives [Cockton and Woolrych, 2001,] and c) identifies evidence to support empirical methods or a combination of methods [Karat et al, 1992] as complementary tools.

The central theme in this research is cost-effectiveness, measured by ratio between the number of findings discovered and the time taken. Although this research tradition recognises cost as a consideration, the focus is the effectiveness of the method as an instrument for acquiring data to support research.

A commercial perspective

While cost-effectiveness is clearly an important concern in industry, other factors may dominate the selection of methods. These factors can be categorized as project goals and project constraints.

Clients articulate a range of motivations for commissioning studies. They may be broadly considered as diagnostic or related to change agency. To draw an analogy, a doctor might prescribe a diagnostic blood test to check for the presence and type of infections where a dietician might conduct a review with a patient in order to persuade them of the value of adopting a healthier lifestyle. Likewise one usability study might diagnose the presence and type of errors where another might be concerned with gathering evidence of damaging design problems to support the investment or organisational change needed to adopt user-centred design practices.

Commercial experience suggests that empirical studies are more effective than expert studies in developing evidence to support practice change. Where expert opinion can be discounted, user performance statistics, attitude measurements, direct quotations and video extracts are difficult to repudiate.

Where the purpose of the study is diagnostic, distinctions between methods are more subtle. In general, the output from a heuristic evaluation can be better structured for identifying necessary changes to both designs and processes. Where evaluators code each finding with respect to the impacted user task, the panel in error and the heuristic that has been violated, simple quantitative analysis can reveal "hot spots" that require action. For example, if a heuristic evaluation of a retail website records many severe findings related to violations of consistency and standards while attempting to purchase a product, then a practitioner could easily identify the need for defining and rigorously applying a style guide to prevent lost sales. The additional identification of the pages associated with these errors would support a rigorous assessment of the business impact and point designers to the areas of the site that required design attention. Because equivalent data from a user study is typically structured around user behaviour rather than principles and design elements, interpretation and development of recommendations is less intuitive and transparent.

Another critical distinction contrasts *sensitivity* and *breadth*. In our experience, empirical studies tend to be more sensitive to subtle socio-technical design errors that may be hard to detect by inspection. For example, in a recent comparative study of alternative information architectures, participants reported that a design proposal well supported by theory was perceived as patronizing. Conversely, the coverage of empirical studies may be limited by the diversity of the sample employed. For example, an evaluation based on a sample corresponding to a customer demographic would be unlikely to reveal even basic accessibility issues easily detected by inspection. By contrast, a heuristic evaluation where the expert panel includes an accessibility specialist reliably identifies a broad range high-level accessibility issues.

In practice, every project is bounded by a set of constraints. A competent practitioner is expected to identify and respond to these constraints by selecting and adapting methods. For example, access to representative users is frequently limited by concerns about costs, availability, and confidentiality. Where evaluators have multiple concurrent commitments, it may also be difficult to assemble a team of practitioners in a single time and place for an extended period. A client may also place time constraints on a study, requiring a short turnaround from brief to results. Clients may also have absolute cost constraints and be prepared to sacrifice some degree of coverage and rigour for economy.

In our experience, expert methods are easier to manage in a constrained environment. Firstly, expert evaluations tend to require somewhat less effort to design, execute and analyse. There are excellent opportunities for reuse of study design components and software tools can be usefully deployed to reduce the costs of the administrative and data management aspects of the engagement. Secondly,

because access to users is not required, the potentially significant costs and delays associated with recruitment are eliminated. Thirdly, heuristic evaluation can be executed asynchronously, gathering input from evaluators working in different locations at different times.

Methods integration

Commercial practitioners commonly need to perform evaluations within a broader methodological framework. Such frameworks may require specific outputs from evaluation activities in order to ensure continuity across the design process. In this context, the need to support an embracing methodology is also likely to influence the choice of evaluation method.

A specific case concerns the need for a framework not only to guide design but also to demonstrate a significant return on investment in that framework [Vredenburg et al. 2002]. For example, the IBM User Engineering method [IBM Ease of Use 2003] uses quantitative measures not only to inform and validate design with respect to business and user goals but also to confirm the economic value of user-centered methods. While early formative testing within User Engineering can be handled by either inspection methods or empirical methods, only empirical methods are appropriate for the summative testing required to meet these goals.

A model for selecting methods

The following figure proposes a model for method selection derived from the above analysis of commercial practice.

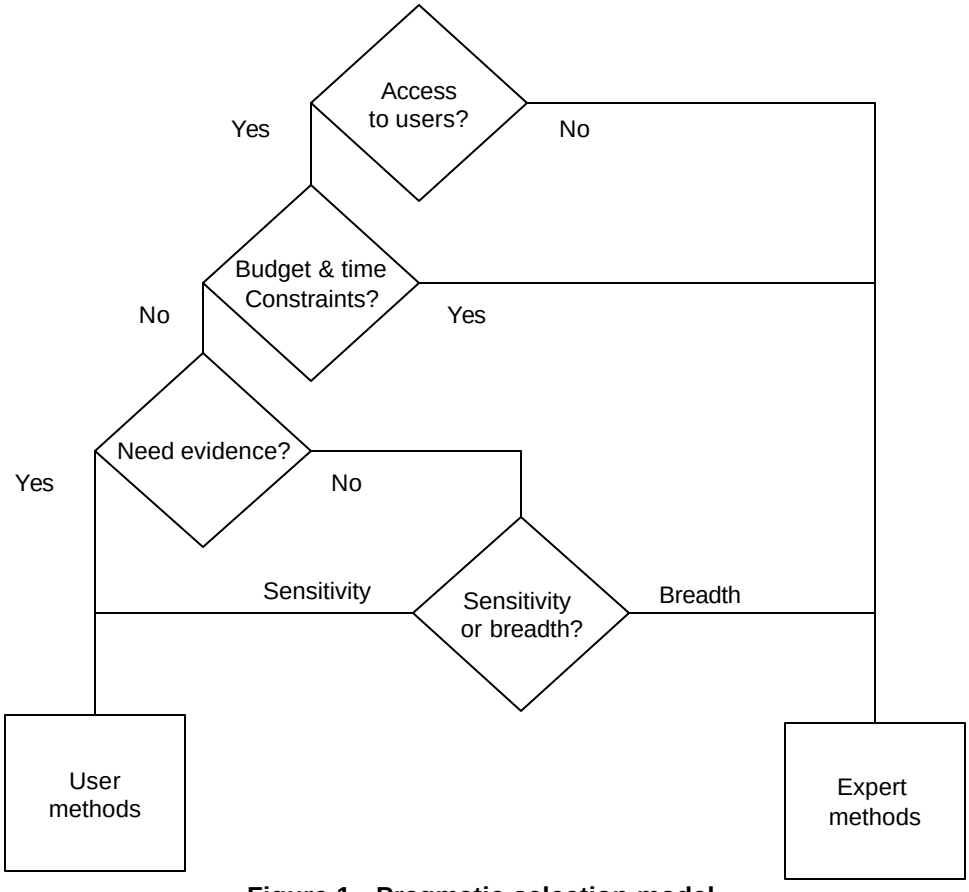

Figure 1: Pragmatic selection model

While this model addresses the issue of cost-effectiveness discussed in the academic literature, it does so within a broader framework drawn from the concerns of practice. This approach recognizes the role of the pragmatic practitioner in negotiating research designs that respect the values of scientific method while recognizing the additional complexities of working with clients motivated by concerns other than those of science.

The model has been used in training for inexperienced evaluators and found to be useful in articulating the issues found in practice and defining a clear and practical decision path for discriminating between these groups of methods.

Case studies

The following case studies illustrate applications of the above model in commercial practice.

Client A sells retail products directly to consumers by means of a web site. They commissioned a study of the site in order to confirm internal analyses of usability issues and to assess the appropriateness of adopting user-centered design methods for a future iteration of the site. Users were easily obtained by supplying a market research agency with a well-defined demographic profile of the target audience for products available on the site. The brief called for a rigorous study with no requirement for a fast turn-around. Given the high commitment of some internal business specialists to the existing design, empirical evidence was considered useful to ensure that any design issues were recognized and addressed. Following the model, a user study was recommended and carried out. The evidence acquired by the study was found to be valuable in communicating the severity of the findings to business and technical stakeholders and in gaining commitment to adoption of user-centered methods.

Client B develops software to support business analysts in constructing and analyzing models of organizational processes. They commissioned a study to benchmark the usability of the current release in order to inform a planned major redesign project. In this engagement, the additional cost of recruitment, training and compensation of suitably experienced analysts would have exceeded the budget. Results were also required within a short period of the brief to avoid delays to the project plan. Following the model, a heuristic evaluation was proposed and carried out. The study identified the need for a style guide and a change management process to ensure a more consistent, refined user experience. Given a mature relationship with the client and strong management support, an implementation plan was rapidly agreed without the need for persuasive empirical evidence.

Conclusions

While issues of cost-efficiency are important to practitioners, other factors may be more critical in selecting methods in a commercial engagement. These factors can be seen to be related to the concerns of practice rather than to those of science. The criterion-based model described above provides a framework for systematically identifying relevant project goals and constraints and recommending appropriate methods.

Where the demands of science take precedence, empirical methods provide compelling evidence and rigorous methodology. Where the concerns of practice dominate, inspection methods may be preferred. If access to users is limited, budgets are constrained, timeliness is essential and breadth of focus considered critical, inspection methods are strongly indicated.

References

Cockton, G., Woolrych, A., 2001
Understanding inspection methods: lessons from an assessment of heuristic evaluation. In *People and computers XV - Interaction without frontiers*

Gray, W.D., and Salzman, M.C., 1998 Damaged merchandise? A review of experiments that compare usability and evaluation methods. In *Human-Computer Interaction*, 13

IBM Ease of Use, 2003
User Engineering
www.ibm.com/easy

Jeffries, R., Miller, J, Wharton, C., Uyeda, K., 1991
User interface evaluation in the real world: a comparison of four techniques, Published in *Human factors in computing systems CHI'91 conference proceedings*

Karat, C-M., Campbell, R., Fiegel, T., 1992. Comparison of empirical testing and walkthrough methods in user interface evaluation. In *Human factors in computing systems CHI'92 conference proceedings*

Nielsen, J. 1994
Usability Engineering, Academic Press

Vredenburg, J. et al. 2002
Vredenburg, J., Mao, J., Smith, P., Carey, T.
A user of User-centered design practice.
ACM SIGCHI 2002