Evaluation Comes in Many Guises

Keith Andrews
IICM, Graz University of Technology
Inffeldgasse 16c
A-8010 Graz, Austria
kandrews@iicm.edu

ABSTRACT

As the information visualisation (infovis) community matures, the evaluation of information visualisation techniques is becoming more of a requirement and less of an optional extra. Unfortunately, the term *evaluation* means different things to different people.

Simply encouraging "evaluation" is too general and imprecise. There is a need for clarification as to *what kind of evaluation* is expected at what stage. When reporting their work, authors should clearly distinguish between exploratory, predictive, formative, and summative evaluation.

Keywords

information visualisation, evaluation, user testing, inspection techniques, longitudinal studies, comparative studies, exploratory, predictive, formative, summative

1. INTRODUCTION

As the information visualisation (infovis) community matures, the evaluation of information visualisation techniques is becoming more of a requirement and less of an optional extra [1]. Unfortunately, the term *evaluation* means different things to different people. When asked about whether any evaluation has been done, the range of responses goes from no evaluation (worryingly common), through heuristic evaluation and thinking aloud testing, to longitudinal studies and full formal comparative studies. All of these techniques seem to fall under the umbrella term of "evaluation".

In the context of paper reviewing and project appraisals at least (and probably elsewhere too), there is a need for more precision. Calls for papers should be explicit about *what kind of evaluation* is expected at what stage. The same applies to the published criteria for project appraisals. When reporting their work, authors should describe exactly what kind of evaluation was performed and for what purpose.

For example, was the evaluation performed by evaluation specialists or representative test users? Was the purpose of the evaluation to provide design feedback, to demonstrate effective usage scenarios for a particular class of user, or to objectively compare two or more interfaces?

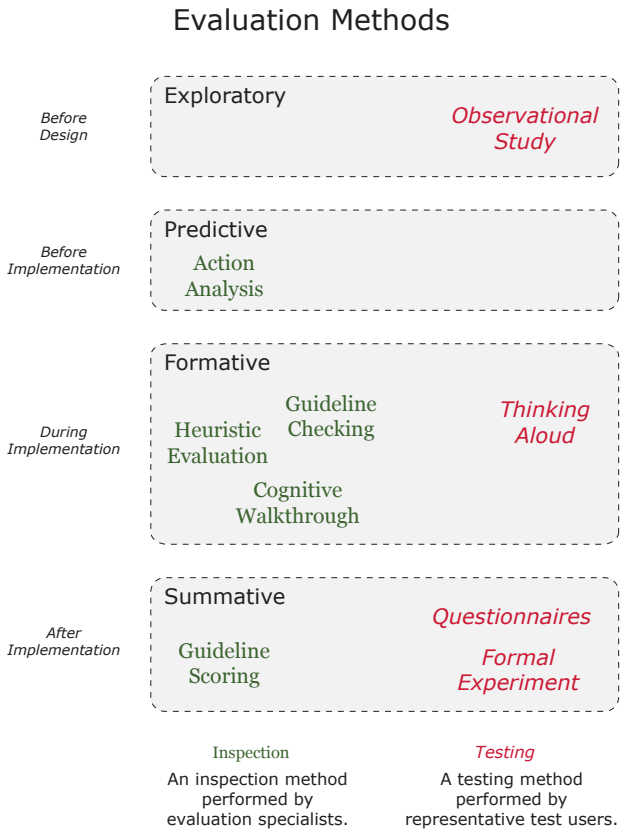

Figure 1: Nine common evaluation methods grouped by purpose and who performs them.

2. TYPES OF EVALUATION

Evaluation methods can be classified into one of two types according to who performs the evaluation:

- *Inspection methods*: specialist evaluators inspect an interface and use their experience and judgement to assess it.
- *Testing methods*: representative test users use one or more interfaces and observations or measurements are made.

Revised Position Paper.
BELIV'08 Workshop, CHI 2008, Florence, Italy.

Method	*Type*	*Purpose*	*Description*
Observational Study	Testing	Exploratory	A longer term study following a small sample of users as they use an interface for their own tasks. Observations and anecdotal evidence are collected and assessed.
Action Analysis	Inspection	Predictive	An evaluator produces an estimate of the time an expert user will take to complete a given task, by breaking the task down into ever smaller steps and then summing up the atomic action times.
Heuristic Evaluation	Inspection	Formative	A small team of evaluators inspects an interface using a small checklist of general principles and produces an aggregate list of potential problems.
Guideline Checking	Inspection	Formative	An evaluator checks an interface against a detailed list of specific guidelines and produces a list of deviations from the guidelines.
Cognitive Walkthrough	Inspection	Formative	A small team walks through a typical task in the mind set of a novice user and produces a success or failure story at each step along the correct path.
Thinking Aloud	Testing	Formative	Representative test users are asked to think out loud while performing a set of typical tasks. The insight gained into why problems arise is used to produce a list of recommendations.
Guideline Scoring	Inspection	Summative	An evaluator scores an interface against a detailed list of specific guidelines and produces a total score representing the degree to which an interface follows the guidelines.
Questionnaires	Testing	Summative	After using one or more interfaces for some typical tasks, test users are asked to rate the interface(s) on a series of scales.
Formal Experiment	Testing	Summative	A larger sample of users performs a set of tasks on one or more interfaces. Objective measurement data is collected and statistically analysed.

Table 1: Nine common evaluation methods, classified according to their type and purpose.

Evaluation methods can also be classified according to their purpose:

- *Exploratory*: Exploratory evaluation provides evidence of how an interface is used and what it is used for.
- *Predictive*: Predictive evaluation produces an estimate of user performance based on an interface design.
- *Formative*: Formative evaluation provides design feedback, often in the form of a list of problems and recommended solutions.
- *Summative*: Summative evaluation provides an overall assessment of a single interface or a comparison of multiple interfaces, often in the form of numerical data which is statistically analysed.

Extending Robert Stake's soup analogy [4, 6]:

> "When the cook tastes other cooks' soups, that's exploratory.
> When the cook predicts the quality of a soup from a recipe, that's predictive.
> When the cook tastes his own soup while making it, that's formative.
> When the guests (or food critics) taste the soup, that's summative."

Figure 1 groups a sample of nine common evaluation methods according to their purpose and type. The methods themselves are described in Table 1. See [2] for further details.

Currently, observational studies are considered exploratory: they are assumed to be done before an interface is built to explore users' goals and needs based on usage of current tools. There is a case for arguing that an observational study of one's own interface after it has been implemented could be considered summative: evidence is gathered from a small number of longer-term users to informally validate the design.

3. EVALUATING INFOVIS TECHNIQUES

Formative techniques such as thinking aloud testing are well-suited to providing development feedback when building an information visualisation system. They should really be done *as a matter of course* during system development, in order to iron out bugs and problems. In my opinion, thinking aloud tests, and indeed all other forms of formative evaluation, should receive mention in an infovis paper, but no more than that. Formative methods lead to better and more usable systems, but neither offer validation of an approach nor provide evidence of the superiority of an approach in a particular context.

The kind of evaluation we want to see being undertaken for infovis systems is that which demonstrates either their utility (fitness-for-purpose) or superiority over other techniques in particular contexts of use. In other words, we should be specifically asking for exploratory and summative evaluations.

Observational studies are used to gather and analyse usage data once a system has been built [3, 5]. Infovis systems are often designed for small numbers of highly skilled users,

where running larger-scale summative studies might be impractical.

However, neither formative testing nor observational studies are suitable for objective comparison of two or more infovis techniques. For comparative studies, formal experiments offer the only objective solution.

4. REFERENCES

[1] K. Andrews. *Evaluating Information Visualisations*. In *Proc. AVI 2006 Workshop BELIV'06*, pages 1–5. ACM Press, Venice, Italy, May 2006. ISBN 1595935622. doi:10.1145/1168149.1168151.

[2] K. Andrews. *HCI Lecture Notes*. 2008. http://courses.iicm.tugraz.at/hci/hci.pdf.

[3] E. Hetzler and A. Turner. *Analysis Experiences Using Information Visualization. IEEE Computer Graphics and Applications*, 24(5):22–26, Sep/Oct 2004. doi:10.1109/MCG.2004.22.

[4] B. Lockee, M. Moore, and J. Burton. *Measuring Success: Evaluation Strategies for Distance Education. EDUCAUSE Quarterly*, 25(1):20–26, 2002. http://www.educause.edu/ir/library/pdf/eqm0213.pdf.

[5] B. Shneiderman and C. Plaisant. *Strategies for Evaluating Information Visualization Tools: Multi-dimensional In-depth Long-term Case Studies*. In *Proc. AVI 2006 Workshop BELIV'06*, pages 1–7. ACM Press, Venice, Italy, May 2006. ISBN 1595935622. doi:10.1145/1168149.1168158.

[6] R. E. Stake. *Evaluating Educational Programmes: The Need and the Response: A Collection of Resource Materials*. OECD, 1976. ISBN 9264115358.