An Explorative Analysis of User Evaluation Studies in Information Visualisation

Geoffrey Ellis
Computing Department
University of Lancaster
Lancaster, LA1 4YW, UK
+44 (0)1524 510340
g.ellis@comp.lancs.ac.uk

Alan Dix
Computing Department
University of Lancaster
Lancaster, LA1 4YW, UK
+44 (0)1524 510319
alan@hcibook.com

http://www.hcibook.com/alan/papers/beliv06-evaluation/

ABSTRACT

This paper presents an analysis of user studies from a review of papers describing new visualisation applications and uses these to highlight various issues related to the evaluation of visualisations. We first consider some of the reasons why the process of evaluating visualisations is so difficult. We then dissect the problem by discussing the importance of recognising the nature of experimental design, datasets and participants as well as the statistical analysis of results. We propose explorative evaluation as a method of discovering new things about visualisation techniques, which may give us a better understanding of the mechanisms of visualisations. Finally we give some practical guidance on how to do evaluation correctly.

Keywords

Explorative evaluation, Information visualisation, Evaluation, Case study

1. INTRODUCTION

How often do we come across a paper describing a new visualisation technique and the future work section at the end states “we intend to undertake a thorough user evaluation” or words to that effect? This is certainly what one of the authors found whilst undertaking a survey of papers in his collection mostly concerned with reducing display clutter in some way. One aim of the survey was to learn from other user evaluation studies to find out about types of participants, experimental details and datasets. He discovered that out of 65 papers describing new visualisation application or techniques, 11 did indeed state that a user evaluation was part of the future work. However a more surprising finding was the fact only 12 out of the 65 papers described any evaluation at all.

So the first question is why do less than 20% of the authors in this literature sample report user evaluations and over 60% do not even think it is worth mentioning? The second question arising from our review of the user studies presented in these 12 papers addresses the effectiveness of the evaluation. Two of the experiments appear to be flawed; 5 seem to be problematic but at least they gave some useful results; one was an informal study with a single user but was interesting and the result of two were possibly a foregone conclusion. We are therefore left with just two user studies that we considered to be successful!

There are a number of interesting papers in the literature that highlight some of the problems of evaluating visualisation techniques [7,14,15]. We have experienced some of these difficulties ourselves (e.g. finding suitable datasets and participants), but it is hard to believe that there are only 2 user studies from the original set of 65 papers that seem to be particularly useful.

This said, we should note that in the 5 papers that the authors have published together on aspects of visualisation, only 1 includes any user evaluation and that would be classed as unacceptable by the criteria of this paper. So, lest our critique seem over harsh, we include ourselves in it! Note too that where we refer to specific papers below this is not to say that they are particularly bad (often the opposite), but that they illustrate more general issues.

The papers that did not report any user evaluation are not considered here. However, we should point out that most of these do attempt to justify the significance of the application by means of examples, arguing that their particular technique has advantages over existing methods, although direct comparison is not common. Some papers do report empirical studies based on simulations but this is primarily to demonstrate efficiency.

Section 2 looks at some of the aforementioned user studies, highlighting some apparent problems and less frequent successes. Section 3 makes suggestions on why it may be particularly difficult to evaluate visualisations, considering their complexity, dataset, measurement and analysis. Section 4 looks at some of the broader issues of evaluating visualisation and proposes that regarding evaluation as explorative is often a more appropriate standpoint. Finally, in section 5 we offer some more practical advice on how to avoid some of the pitfalls and perform useful and successful evaluations.

In summary we find that (a) 'evaluation sucks' [8] (b) it sucks because it is hard and (c) but if you think about it differently it may not be so bad after all and so (d) you can actual do it right.

BELIV 2006 Venice, Italy.

2. CASE STUDIES

This section highlights some possible problems and successes from the user evaluation studies described in the collection of information visualisation papers of one of the authors. Note that this is not intended to be a comprehensive review of the literature as it focussed on a relatively small collection of papers with some relevance to clutter reduction.

2.1 Studies with foregone conclusions?

Some user evaluation studies include experiments, which generate results that are possibly a foregone conclusion. For instance, in a study [16], the distortion of edge lines in a dense graph layout is being examined to help users understand which nodes are connected to which lines. One method gently bends the lines, so they separate, but the curve draws the eye towards the end nodes. The other method, also using a lens metaphor, gives a circular distortion, so the end of the distorted line is pointing anywhere but towards the end node! Even without a study, the best method is rather obvious. As is often the case, many participants were involved and a fair amount of data gathering and analysis was performed to 'show' that one method was better than the other. However, the definitive answer to the study lies in the users' comments of the circular distortion method – "edges bend a weird way", "awkward and not useful" and "I don't like this"! Users of the other method instead made comments such as "works great" and "identifies routes very well".

Another example [1] of what might be deemed as unnecessary was an investigation where users were asked to click on a particular number (actually a rectangular box containing the number), when the numbers in the range of 1 to 99 were displayed in a window on the screen. The experiment arranged the numbers in three different ways: un-ordered, partially ordered, and ordered left to right, top to bottom. It is not surprising that the latter task was completed in the least time, while the first took a lot longer. To be fair, the experiment was comparing the users' results against a calculated metric, but was the experiment noted above really necessary?

2.2 Wrong sort of experiment?

Other types of user studies that came up in our literature sample appear to be inappropriate in that, instead of evaluating the visualisation per se, they tend to test something else. Some authors have commented that performing a full user evaluation (what ever that is!) is beyond the scope of the project and hence a small trial and/or testing a minor part of the system is all that can be done at that stage. However, it might be the feeling that some reviewers (and conferences) like to see the inclusion of an evaluation, whether or not this is useful, and that drives some of us to include user studies!

For example, one experiment [6] was carried out to discover if users could find information on a map quicker by utilizing popup labels or by zooming in to read the labels. The popup labels were found to be faster despite ignoring the time to zoom in the map. However, comments from the participants revealed that the zooming was often disorientating, so one could argue that it was in fact a focus+context problem that was being tested. If the users of the 'zoom' interface had been given a magnifying lens (electronic version!) that enabled them to read the small text in a localized area, would the results then have been the other way round? In addition, activating the popup labels required far greater precision in manipulating the mouse pointer than the zoom interface, so was this more a dexterity test? We are not saying that the labelling technique is poor; on the contrary, it appears to be a very good solution for presenting specific information in a crowded space, but the problem lies with an experimental design that is not giving much insight into the benefits of the new technique.

In another experiment [17] users were asked to find patterns using two different types of parallel coordinate plots. Patterns were loosely defined as either clusters or outliers. In the standard version, participants found 8 or 9 out of the 25 patterns that were deemed by an expert to be significant, whereas with the enhanced version users discovered about 16 patterns – a sizeable increase. Should we not question which of these patterns are the important ones? It is certainly useful to discover more relationships in the data, but if we are missing the vital one, then is quantity that important?

Finally, in a study [11] involving a new interface showing web search results, participants spent at least 16 hours on 3 different interface configurations. One relatively small change to the standard interface reduced the query search time by 25%, but this was probably expected based on their previous work. In the third configuration, 7 parameters were changed and sorting out the effects of each of these would be practically impossible. In terms of the timing data, there was little difference between the standard interface and the third configuration, and based on users' preference, its likeability was about the same. Yet, looking at the individual responses, it is evident that some users really liked the third configuration whilst others hated it – a clear example of the dangers of averaging results! Users' comments also revealed that increasing the size of the text was probably the most significant benefit of the third version, something that an analysis of the data would never have uncovered. So, were these experiments really necessary? It could be argued that, observing a few users in an informal evaluation and recording their comments could well have provided as much understanding of the new technique.

2.3 Fishing for results?

In one of the studies we looked at [12], the visual interface application was designed to help users browse and understand large document collections. The first experiment compared the effectiveness of the visual interface with a text-based search interface and found it to be not so effective. The authors expected this as the application was designed primarily for browsing. A second experiment was conducted to assess whether the visual interface gave the users a better understanding of the structure of the document collection. Users were asked to draw a tree structure to represent the topics in the collection. These were analysed and as one would expect, the visual interface user group produced diagrams that were more similar to each other than the text searchers. This was used as evidence to show that the visual interface was better in producing a more coherent view of a large document collection. As the authors pointed out, it is difficult to assess how much knowledge a user has gained from a browsing activity. However, presenting one group of users with what amounts to a picture to copy, and using the fact that they mostly ended up with similar pictures to infer that the application is effective in this context is like fishing for results!

2.4 What makes a good study?

As mentioned in the introduction, 2 out of 12 studies in the literature set appear to be successful – in that they effectively demonstrate the potential benefits of some application through user evaluation.

For instance, in an evaluation of an interface showing web search results [4], the authors ran subsequent tests by changing an appropriate parameter each time in order to tease out the reasons behind the test results. This method of deciding what to investigate next, based on previous results, seems to be a good approach and probably has a greater chance of understanding the interaction mechanisms than 'let's do lots of test runs with any number of independent variables and hope to sort it out in the end'. In order to adopt this iterative approach, one clearly needs participants who can return for a series of experiments, thus increasing the overheads, but then fewer users may be required. The other significant feature of this study is that the authors attempt to generalize their results, something that is sadly lacking in some of the papers.

Another noteworthy example [13] is the evaluation of a new visualisation technique that did not time anything – not a stop-watch or timing device in sight! It involved getting a large group of 'real' users from a wide range of jobs but they all had a good knowledge of the domain and the data set. After the prototype was demonstrated to a small group of potential users, they were given the opportunity to use the interface with some typical data and were then asked to comment on this and come up with potential advantages and problems. A wealth of useful data was collected and as others have reported [7], domain experts are often worth their weight in gold.

3. PROBLEMS OF EVALUATION

Why is it apparently so hard to evaluate visualisations effectively? In fact there are many reasons, some shared with general user interface evaluation, and some more particular to visualisation.

3.1 Complexity

The visualisation process consists of many complex interactions and is thus difficult to be treated as a whole. However, we can attempt to understand the mechanisms that drive it.

3.1.1 Interpretation and credit assignment

In common with other user interfaces, visualisations typically embody many assumptions and theoretical views. Carroll and Rosson's *claim's analysis* [2] seeks to expose the many factors affecting the usability of interfaces, e.g consistency in the layout of navigation buttons or the colour of highlighted data items. These often interlocking 'claims' are implicit in software, but this is not a way of thinking that is common amongst those producing novel visualisation techniques. Even, if one has managed to articulate the multiple claims embodied in a visualisation, simple end-to-end timing measures or user satisfaction score gives little indication of which of these have been important in the success (or otherwise!) of a technique.

3.1.2 Mechanism

Even a simple interaction with a visualisation will include multiple stages and steps at both a coarse level (e.g. getting to know a data set and then finding items in the data set) and at a fine level (e.g. visually scanning for some feature, then moving the mouse and selecting a node, then evaluating pop-out detail).

Again end-to-end measures are not the most helpful in working out which steps are making a difference. For example, an experiment might find no difference between the performances of two visualisations, but the task involved included both 'getting to know' a data set and finding specific features. It may be that one technique is in fact better at the former and the other at the latter, if one realises this it might suggest ways of creating a hybrid technique, but without this both just appear the same.

Even worse, the aspects of the interaction that cause a difference may be completely irrelevant to the essential qualities of a visualisation technique. In one of the papers we studied, after pages of timing and accuracy data, some users were quoted as saying that one visualisation was preferred to the other because the font was bigger and it was easier to read. While this was in some part related to the nature of the visualisation (a form of fish-eye), it is likely that some small 'fix' may well have been able to avoid this problem with the second visualisation – were the differences seen due to a simple detail of the implementation?

3.2 Diversity

Another reason that makes evaluation of visualisations harder lies in the diversity of tasks, data sets and participants.

3.2.1 Variety of data sets

Different visualisations deal with different kinds of data. While there has been some attempt to create a standard, (e.g. the FADIVA network a few years ago), we still do not have well-developed and easily available standard datasets in the way that the information retrieval community do (e.g. TREC). This means that visualisation evaluation (and even simple demo-ing), is limited by the availability of data, or compromised by inappropriate, or artificially generated data.

3.2.2 Indeterminacy of tasks

Different tasks are better supported by different visualisations. In a recent evaluation, standard outliner-style TreeViews were compared with PieTrees [10], which have a constant value–area mapping similar to TreeMaps. Not surprisingly, tasks such as 'find the biggest' were fastest using the PieTree whereas finding a specific named node, where the area mapping did not act as a heuristic, was fastest with the TreeView.

It would be very easy to have chosen a task that had made one or other look better and think this was definitive – indeed, it is natural to choose tasks (and datasets) that suit a novel technique i.e. ones that it is good at. Furthermore, as a researcher, there is a temptation to deliberately choose the tasks that make one's method look good – referees are often unforgiving of a truthful paper that says a technique has strengths and weaknesses.

Not only do tasks differ, but the real tasks we want are usually open ended. If the user knows beforehand what is important to see in the visualisation, then there are typically better ways of looking for it: aggregates, searches etc. Visualisations are often at their best for more exploratory tasks, but these are precisely the tasks that are hardest to replicate in an experiment.

3.2.3 Individuality of people

Students may be useful but … The majority of the studies used students, often computing students as their subjects. Clearly, students are convenient. They are nearby and can be persuaded to give up a few hours either because we convince them that they are doing something worthwhile or there is some monetary incentive.

In some cases this is fine, for example, interaction or perception experiments (e.g. check colour or size of objects that are to be selected or manipulated in some way) require little knowledge of the visualisation domain, hence students would be suitable. However, there is a large amount of literature dealing with cognitive issues, which may well guide the designer.

But where a more realistic task (i.e. where the task matches the application domain) is used in the evaluation, participants need a clear understanding of the problem that the visualisation tool is attempting to solve and also, one might argue, an understanding of the data itself. In such cases, the chances of assessing the usefulness of the tool using students will be slim, as we found during the informal testing of our Sampling Lens [5]. Users liked the lens-based tool as it revealed patterns within a parallel coordinate plot in areas where, in the absence of the lens, there were too many overlapping lines. However, when we asked the users what the patterns meant, they did not really understand it; they just thought it was cool!

Other researchers [7,15] have suggested that better information can be obtained by using either a small number of domain experts involved in more qualitative studies, expert visual designers or HCI expert reviewers. Of course, it is more difficult to get access to such a group of people.

The acceptability of a particular kind of experimental subject is dependent on the exact details of the experiment or system. However, we ought to consider these: (i) it is always important to explicitly consider the potential effect of the type of participant on the interpretation of the results; and (ii) understanding the mechanism is again essential so that by considering the details of the interaction one can determine aspects that are capable of evaluation by non experts.

Recognising individuality is also important when analysing results. Different cognitive styles may lead to a particular technique working well for one group and not for another – the overall averages may hide this and simply appear inconclusive. Finding who a technique is really useful for may be more important than making it work pretty well for everyone.

3.3 Measurement

Issues of accuracy, precision and significance of statistical data are fundamental when discussing the relevance of experimental results.

3.3.1 125.2 seconds to do what?

Studies often present end-to-end times to do a certain task, and sometimes the average of a set of tasks. What does this tell us about the interaction and more importantly, the understanding of the user? In some cases, time may be of the essence, for example in an in-car information display, but more often 'time' seems to be an easily measurable proxy for 'ease of use' … and not necessarily the most accurate!

Numbers are powerful when we understand what they mean, but they can also be misleading. In particular, it is often easy to let precision fool one into an impression of accuracy. In many of the experimental results we see the time to do a task or set of tasks given to the nearest tenth of a second. e.g. 125.2. While this may be a true representation of the measured value, if the level of variation is +/– 17 seconds, then would 120 seconds be good enough … or even, “task completed fairly quickly”?

Advertisers deliberately use apparently precise numbers as a way to suggest validity: “most cats prefer Fishkers” sounds unconvincing, "applying face cream with RexonolicB++ reduces wrinkles by 37%" has an aura of scientific truth. Whilst academics are not deliberately attempting to deceive their readers, they are perhaps often accidentally deceiving themselves.

Of course if you do not quote exact numbers, it is impossible for a reader to, for example, check your statistics. However, it is possible to be both precise with data and accurate in rhetoric, for example, giving precise numbers in tables, but using the most appropriate number language in the text. Also, where numbers of subjects or trials are small, it is often better to quote the exact numbers (e.g. 7 out of 9 subjects) rather than converting this into an apparently over-precise percentage (52.9%).

3.3.2 Statistics: significance and importance

People find statistics difficult. There are various reasons for this, some to do with education and some to do with the mix of mathematics and real-world understanding. For those with computing background and unlikely to have been exposed to statistics at undergraduate level, this is particularly hard … and it is not surprising that this is reflected in published work.

One problem lies in the particular meaning of the word "significance" in statistics. It seems that when we have collected our data for a range of dependent and independent variables, we put this into a statistics package and then quote its significance, but we often do not stop to think what it means. If p<0.05, we think "yes done it", but in a highly accurate experiment totally unimportant differences may show up as significant – yes visualisation X is faster than visualisation Y … but only by 3 milliseconds! What (and all) a significance test is saying is that with p<0.05 the chance of the observed data being a random occurrence is 1 in 20.

Even more worrying is the treatment of non-significant results. Often a graph, which as a visualisation community we know is hard to ignore, appears to show a difference, but the text says this is not significant or 'marginally significant' (whatever that might mean!). In other words, the graph we are seeing could just as well be the effects of random chance … like a good day at the races, but by being presented to us we are being tempted to believe otherwise … the advertisers would love us!

Often non-significant results are (erroneously) treated as meaning "no difference". This is a misapprehension that every statistics course highlights, but it is still endemic in the literature. Whilst "not significant" does not mean "no effect", in many cases, a confidence interval can allow you to say "unimportant difference". Sadly, although confidence intervals are not difficult to compute or understand, they seem to be where most statistics courses give out. Furthermore, readers are less familiar with confidence intervals and so explanation is often needed. Indeed one of the authors once received a referee’s comment saying he should use proper statistics terminology "significance of p<x" rather than "confidence" … not only having limited understanding, but confident enough in his/her statistical ignorance to critique! Again it is not enough to do evaluation correctly, but also reviewers need to be educated to appreciate it.

3.3.3 Points of comparison and control conditions

Any numerical or ordinal measure requires some gold standard, point of comparison or control in order to know what values are good. Choosing a suitable 'control' can be problematic. For example, the paper that introduced the Xerox Butterfly Browser, a 3D visualisation for following references and citations, included an empirical evaluation – against Dialog [9]. Whilst Dialog was in a sense 'industry standard', it was effectively 1960s technology designed to work over very low bandwidth (10 cps.) phone lines. Perhaps it was not surprising that the users preferred the 3D interactive interface to the command line search, but this hardly tells us much about the new visualisation.

This may seem as an extreme example, but the problem is not so much whether the evaluation in that paper was effective or not, but instead, to determine what a valid point of comparison would have been. Any other 'state of the art' comparison is bound to differ in many respects, causing the credit assignment problem noted above. A suitable alternative would be to change some specific feature and measure the effect of that alteration – more like a traditional psychological experiment. However, if there are any interactions between features (and this is usually the case), then in principle one has to test all possible interactions, which is combinatorially impossible.

4. THE BIG ISSUES

Many of the problems discussed in the previous section can be reduced to two issues, the generative nature of visualisation techniques and the lack of clarity over the purpose of evaluation.

4.1 Evaluating generative artefacts

Visualisations (like all interfaces) are generative artefacts: that is they are things that are not something of value in and of themselves, but only yield results in some context. In the case of a piece of visualisation software, this is when used by a particular type of user to visualise a particular data set for a particular purpose. To further complicate things, the visualisation software is itself typically an implementation of some more generic visualisation technique. But all we can evaluate is the success of the particular instance. In order to really produce a reliable empirical evaluation of a visualisation technique one would need to have many tasks, many data sets, many users and many implementations of the technique produced by many different designers … hardly likely in a finite time!

In search for the validation of generative artefacts …

empirical evaluation of generative artefacts is methodologically unsound

… or put in other words, you cannot evaluate a visualisation … or at least any evaluation cannot tell you, *in itself*, that the visualisation works or doesn't work.

However, whilst you cannot 'prove' a visualisation is good or correct through evaluation; you can perform useful evaluations, and may be able to *validate* the visualisation in other ways.

In some domains, particularly mathematics, it is rare to attempt to perform post hoc evaluation. Instead it is the proof, the process of reasoning from initial knowledge (or assumptions) through lemmas to theorems that give you confidence in the answer. Sometimes you may put example numbers through a theorem – but this is largely to check for 'silly' mistakes in the proof process, not to prove the theorem through the examples. However, in other domains it is hard to work beyond some point through reasoning. For example, in areas of chemistry, one can deduce that certain classes of compound are likely to have an effect, but the precise form and level of that effect may only be found by exhaustive trial of many potential compounds.

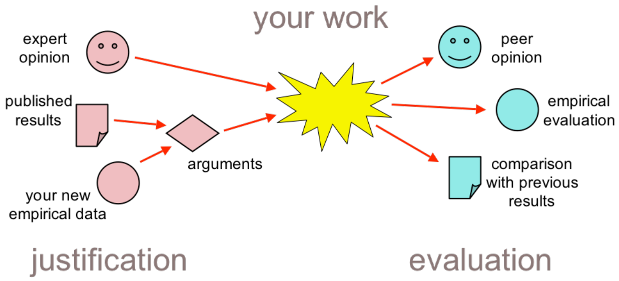

Fig1: The two sides of validation: justification and evaluation

In visualisation we can never have perfect evaluation because of the generative nature of the artefacts we build. Likewise we cannot have perfect justifications because our base knowledge of human perception and cognition is incomplete, and because our ability to reason from these to their implications is flawed. However, if empirical evaluation complements reasoned justification then it can lead to a reliable and strong validation of the visualisation.

Our justification may include (see Fig. 1):

- existing published results of experiments and analysis
- our own empirical data from experiments, studies, etc.
- expert opinion (published or otherwise) and common sense
- arguments based on the above

On the evaluation side we may use:

- empirical evaluation, user studies, timing data,. etc.
- peer reviews of our work (other people agree it is a good idea)
- comparison with previous work (do the parts that should behave the same actually do so)

If one is aware of the weaknesses and gaps in the justification, then an evaluation and subsequent analysis can be tuned to verify the questionable aspects of the justification.

Sadly again, for the researcher, there is a tension between good science and publishability. If you choose to evaluate aspects that are questionable then one is more likely to find problems in the visualisation or to have inconclusive answers. In contrast, if you evaluate the aspects that you are pretty sure are okay from the justification argument, then you are likely to get clear results with nice $p<0.05$ significance results … but learn nothing.

4.2 Purposes of evaluation: summative, formative and explorative

The other cross cutting issue is the need to have a clear idea of the purpose. The distinction between summative and formative evaluation is well known, although for usability, the techniques are very similar and hence it is easy to blur the two.

For example, head-to-head comparisons of techniques for dealing with similar data are really attempting a form of *summative* evaluation: "my visualisation is better than yours". Now this is good marketing, but not very useful science, or even design. In fact, when one looks at the discussion of such evaluations those doing it more often than not end up with some level of suggestions for improving their visualisation. In fact, the actual use of the evaluation is *formative*.

This confusion of purpose is also evident in papers that effectively give a record of an iterative interface development process. If this is an iteration of fundamental novel visualisation concepts and techniques, this is good use of evaluation in research, but if (as is often the case in iterative development) it is merely tweaking features that increase usability, but do not hit the heart of the novelty (e.g. font size), it is good product development, but not good research. Of course, such development is often needed to get a basic concept into a usable enough package to evaluate … but that is the forerunner to the evaluation, not the evaluation itself.

So for both summative and formative evaluation, we need to be constantly careful that they are what we really want and that they address the real issues. However, what is really needed in most research contexts is neither. In fact, we require explorative[1] evaluation – evaluations that help us see new things about our ideas and concepts, which are useful to us. Whilst the purpose of summative evaluation is to obtain a seal of approval and the purpose of formative evaluation is to improve a design, the purpose of explorative evaluation is to find out, to provide knowledge. The difference is sometimes just in the way one views results, but can be more fundamental, for example, for explorative purposes one may deliberately use a bad design to uncover user behaviour in extreme circumstances.

As an example of the latter, some years ago one of the authors, working with Stephen Brewster on audio feedback, deliberately created a calculator-style interface that involved large an inefficient movements if mouse and eye in order to create mis-clicking errors that only occur infrequently, but problematically, in normal 'good' interfaces. By creating the error we were able to validate our understanding of its causes and thus design appropriate feedback to ameliorate its effects [3]

Indeed, many evaluations that appear weak or problematic, when viewed as summative or formative evaluations, are far more convincing when seen as explorative. For example, the paper [12] cited in section 2.3 seems problematic if seen as the former (a form of fishing), but more appropriate for exploration: "what kind of things is Scatter/Gather good for?"

Because the techniques used for all kinds of evaluation are similar, it is often unclear which kind of evaluation authors intended to undertake. Indeed, Zhai [18] in his response to Lieberman's 'The Tyranny of Evaluation' [8], notes that the value of the best evaluation is often not in the original (summative or formative) purpose, but in accidental understandings and findings – that is the explorative aspects. How much more effective might these evaluations have been if the eventual explorative purpose had been identified explicitly in the first place!

4.3 From data to knowledge

To some extent, in even writing (or attending a workshop) on evaluation, we run the danger of subscribing to the phenomenological notion that it is the data (whether qualitative or quantitative) that is in some way the 'real' and 'objective' truth.

In fact it is impossible to generalise data per se; it is always a singular event: whether an ethnography of a particular group at a particular time, or a formal experiment with particular subjects in a particular setting. Any future use of a particular visualisation application, technique or design principle will be different. Knowledge and hence generalisation only comes through the application of reasoning informed by (interpreted) data.

Unfortunately the genre of scientific writing often serves to blur or hide this and many attempts to evaluate by adopting this genre run the risk of not making the best of their work and at worst misleading their readers.

5. DOING IT RIGHT

So effective evaluation of visualisation is hard and fraught with problems, but is also essential in order to rise beyond simply saying "I did this and it's cool". However, bad evaluation is at best useless and at worst can be plain wrong. So, how can we do it right?

think purpose – First of all it is important to know what you are hoping to gain from the evaluation. If your aim is to prove that your system is best, go get a job as an advertising executive. If your aim is simply to make your system as good as possible, then sell your product but don't write about its development. If your aim is to make your product as good as possible in order to effectively deploy it and so learn, this is essential, but not a thing to report in detail. However, if your aim is to understand whether, when and under what circumstance a technique or design principle works or is useful – yes now you are doing research.

think measures and tasks – Is the thing that you are measuring useful (see also below) and if so what is good. It is easy to measure something just because you can. For example, if you are primarily interested in a user engagement with a visualisation then time to complete a fixed task tells you little. However, in an open-ended task, users' time on task (combined with qualitative data) may tell you how much they were enjoying themselves.

think success – Ask yourself: "If this evaluation is as successful as it could be, what will I know at the end that I don't know now?" If the answer is "not much" then why do the experiment? As noted in section 4.1, use the weaknesses in your justification to drive your evaluation, make every subject hour count.

think failure – In the case of quantitative experiments or questionnaires where you plan statistical analysis ask: "If this evaluation does not give statistically significant results, will I have learnt anything?" In fact, at very least, you will have learnt enough to do a power analysis and calculate how many more subjects you would need in order to either detect an important difference or conclude (using a confidence interval) that any differences are negligible. However, this turns what you hoped to be your full evaluation into merely a pilot. This may be all you can do and you just have to do more work. However, if you also collect rich data (video, keystroke logs, talk-aloud transcripts, post-task interviews) then you are more likely to have something of value to report. This takes us to …

think qualitative and quantitative – If you speak to one person who has a particular behaviour, is it just that person? If a formal experiment or questionnaire shows that 75% of people have a particular behaviour, then it is clearly prevalent, but is it important and why does it occur? However, if you combine the two, you both know *that* the behaviour occurs and have some idea *why*.

think mechanism – If you do not understand a process then you cannot generalise. This often involves qualitative data (as above), but may include quantitative data on parts of an interaction, not just end-to-end measurements.

think understanding – How can you manipulate the visualisation itself, the data used or the task you give the user in order to find out most about the most interesting things And yes, as we noted in section 4.2, this may mean using versions of your visualisation that are not the 'best' ones.

6. SUMMARY AND CONCLUSION

An explorative analysis of the experimental details and results questioned the viability of evaluations in cases where the outcome is probably a foregone conclusion, or where inappropriate experiments are perhaps carried out, or even where the results are possibly unconvincing. It was also apparent that, in many of the studies, comments from the users contributed a great deal to

[1] Actually the proper word is exploratory, but explorative rhymes with summative and formative ☺

understanding the visualisation and reinforces the belief that ethnographic or observational techniques often provide more useful data.

The fundamental reason behind so few user studies may be due to the fact that information visualisations are very difficult to evaluate. The visualisation process is made up of a complex set of interactions and ideally, we should understand the mechanisms inherent in the process in order to assess the viability of an evaluation. End-to-end time measurements are not particularly useful when attempting to work out the critical components of a visualisation.

We showed that the choice of appropriate tasks, datasets and participants is important when determining how to evaluate a particular visualisation. In addition, when reporting results of experiments, we discussed the importance of understanding the meaning of accuracy, precision and significance of the statistical data and we also highlighted the problem of finding valid point of comparison between visualisations.

We put forward the idea that empirical evaluation of visualisations on its own is methodologically unsound due to the generative nature of visualisation techniques. However, if empirical evaluation is used in conjunction with reasoned justification then this may lead to a reliable and strong validation of the visualisation.

We also emphasise the need to apply formative and summative forms of evaluations in the appropriate context; but in many cases neither may be suitable. We therefore propose explorative evaluation as a method for helping us see new things about our ideas and concepts and revealing those that are useful to us.

In order to balance the more critical stand of the earlier parts of the paper we have tried to give some practical guidance on how to do evaluation correctly. We hope this will be valuable for those who are new to this and a reminder for those more experienced.

As we strive for publishability, experimental designs and reporting of results may be unduly influenced by the expectation of reviewers. Hence it is not enough to do evaluation correctly; reviewers also need to be educated to appreciate it!

REFERENCES

[1] Bederson, B.B., Shneiderman, B., Wattenberg, M. Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies . *ACM Transactions on Graphics*, 21(4), Oct 2002, 833-854

[2] Carroll, J.M, Rosson, M.B. Getting around the task-artifact cycle: how to make claims and design by scenario. *ACM Transactions on Information Systems*. Vol 10 No 2, April 1992, 181-212

[3] Dix A., Brewster, S. Causing Trouble with Buttons. Ancilliary Proc. HCI'94, Glasgow, Scotland, 1994. http://www.hcibook.com/alan/papers/buttons94/

[4] Dumais, S., Cutrell, E., Chen, H. Optimizing Search by Showing Results In Context. *Proc. CHI'01*, 2001, ACM Press, 277-284

[5] Ellis, G.P., Bertini, E., Dix, A. The Sampling Lens:Making Sense of Saturated Visualisations, *Proc. CHI'05 Extended Abstracts on Human Factors in Computing Systems*, Portland, USA, 2005, ACM Press, 1351-1354

[6] Fekete, J-D., Plaisant, C. Excentric Labeling: Dynamic Neighbourhood Labeling for Data Visualization. Proc. *CHI'99*, Pittsburgh, 1999, ACM Press, 512-519

[7] Kosara, R., Healey, C.G., Interrante, V., Laidlaw, D.H., Ware, C. Thoughts on User Studies: Why, How, and When. Computer *Graphics & Applications*, 23(4), July 2003, 20-25

[8] Lieberman, H. The Tyranny of Evaluation. (accessed 2006). http://web.media.mit.edu/~lieber/Misc/Tyranny-Evaluation.html

[9] Mackinlay, J. D., Rao, R., Card, S. K. An Organic User Interface For Searching Citation Links, *Proc. CHI'95*, Denver, May 1995, ACM Press, 67-73

[10] O'Donnell, R., Dix, A., Ball, L. Exploring The PieTree for Representing Numerical Hierarchical Data, Proc. HCI2006, London, Sept. 2006, Springer

[11] Paek, T., Dumais, S., Logan, R. WaveLens: A New View onto Internet Search Results. *Proc. CHI'04*, Vienna, Austria, Apr 2004, ACM Press, 727-733

[12] Pirolli, P., Schank, P., Hearst, M., Diehl, C. Scatter/Gather browsing communicates the topic structure of a very large text collection. *Proc. CHI'96*, Vancouver, May 1996, ACM Press, 213–220

[13] Plaisant, C., Milash, B., Rose, A., Widoff, S., Shneiderman, B. LifeLines: Visualizing Personal Histories. *Proc. CHI'96*, 1996, ACM Press, 221-227

[14] Plaisant, C. The Challenge of Information Visualization Evaluation. *Advanced Visual interfaces*, Italy, 2004, ACM Press

[15] Tory, M., Möller, T. Evaluating Visualizations: Do Expert Reviews Work? *IEEE Computer Graphics and Applications*, 25(5), 2005, 8-11

[16] Wong, N., Carpendale, S., Greenberg, S. EdgeLens: An Interactive Method for Managing Edge Congestion in Graphs. *IEEE Symposium on Information Visualization*, Oct 2003, 51-58

[17] Yang, J., Ward, M.O., Rundensteiner, E.A., Huang, S. Interactive hierarchical displays: a general framework for visualization and exploration of large multivariate data sets. *Computers and Graphics*, 27(2), Apr 2003, 265-283

[18] Zhai, S. Evaluation is the worst form of HCI research except all those other forms that have been tried. (accessed 2006). http://www.almaden.ibm.com/u/zhai/papers/EvaluationDemocracy.htm