CS 598 ACK
Experimental HCI & Interactive Technologies

Text Chapter 6

Reporting Experimental HCI Research in Publications
So, now you have your results, and you want to tell everyone about them. This chapter discusses the way in which you report your research, typically in a research article for an academic conference or journal, or in a dissertation for assessment. In all cases, you need to keep in mind that someone else will be reading what you write, and that this person has not been party to your decision-making process. It is easy to leave information out because it appears obvious.
Reviewer’s Concerns and Attitudes:

- An experiment focuses on addressing specific research questions: reviewers may not believe that these questions are interesting or important.
- An experimental research question could be addressed in many different ways: everyone will have their own idea as to how best to address it.
- Different statistical methods are favoured by different people: beware the reviewers who are well versed in statistics when you have not analysed your data using their favoured method!
- No experiment can ever be perfect: reviewers can easily find faults on which to base a negative judgement if they want.
- Many reviewers have never actually designed and conducted an experiment themselves: they do not always appreciate the amount of work required for running experiments, the difficulty in making appropriate design decisions, or the constraints that apply to experimental design.
Justifying the Research Question and the Experimental Design:

First, a case needs to be made as to why the research question you are addressing is important and interesting – this is typically done with reference to previous research, experimental or otherwise. Identifying specific prior research that makes a clear case for the question to be addressed is useful.

More difficult is the case when the research question arises from ideas spread across different research areas: A clear case needs to be made as to why these areas were worth bringing together and how the research question emerged from a consideration of how they linked together.
Reporting Experimental HCI Research in Publications

Justifying the Research Question and the Experimental Design:

Typical reviewer comments on the motivation for the research include “the authors need to make clearer the case for doing this study,” “the whole context of the experiment was flawed to begin with,” “the authors need to reframe this research some other way,” “combining [two areas of research] into a body of work is not necessarily the only reason to do it,” or “I am left with one question: So what?”

Note: It is tremendously easier to respond to such questions when the research is anchored in terms of addressing a socially relevant problem or issue.

It is much, much harder to respond to these questions when the research motivation is (essentially), to get another publication.
Reporting Experimental HCI Research in Publications

Justifying the Research Method:

ALL experimental design decisions need to be justified. That is, the reader needs to know why the experiment has been designed this way and why other possible options were not chosen.

This is particularly difficult when there are restrictions on the maximum number of pages allowed for the article because you need to be both comprehensive and succinct. In the (usually rare) case that the method is similar to one that has been previously published, then it may be possible to refer to this prior work as an indirect means of justifying your method.

However, this would need to be a popular and well-known experiment, and you would need to be sure that the readers will know about it. Not explaining design decisions leaves you vulnerable to reviewers who are looking for reasons to reject your paper [REASONS include that competition for published research pages is a ZERO-SUM GAME].
Justifying the Research Method:

It is useful to remember that the experiment is a single point in a multidimensional design space, where each dimension represents a decision that has been made, and that there are many dimensions of choice, including (but not limited to) the following:
Justifying the Research Method: Research Dimensions Include:

- conditions,
- experimental objects,
- tasks,
- within- or between-participants method,
- pre- and postexperiment activities,
- location of the experiment,
- nature of participants,
- equipment used,
- number of participants,
- experimental timing,
- data collection methods,
- data analysis methods.
Justifying the Research Method:

The choice of experimental **objects and tasks** will be related to your choice of domain. It is common for reviewers to complain that the objects and tasks are too abstract and that the results thus have no “real world” applicability (e.g., “limited generalizability to typical interfaces”).

This is a difficult issue to address. In breaking new experimental ground, using an abstract domain in a preliminary experiment enables important experience to be gained before experimenting using real world scenarios, and reviewers need to be persuaded of the importance of an initial (abstract) step in guiding the design of a second (applied) experiment.

A clear case therefore needs to be made for an abstract experiment: this may be easier if it is clear that either the results are in a never before-explored area or the experiment and its results are obviously necessary to inform the design of a later “real world” experiment.
Reporting Experimental HCI Research in Publications
Reporting Experimental HCI Research in Publications

Justifying the Research Method:

Timing is also a common query, especially if you have chosen to limit the time for each trial, for example, “I have trouble understanding why the task completion time needs to be limited to 20 seconds.”

You would usually have a good reason for limiting the time for the trials: this decision may interact with other decisions and should be explained clearly.

Because one reason for limiting the time for each trial is to keep the experiment at a reasonable length, this should be stated, particularly if your timing choices have been based on experiences in pilot experiments.

Note: if decisions are practical rather then theoretical, FESS UP!
Justifying the Research Method:

The nature and number of participants should always be reported, as well as the means of recruiting them. Comments on this issue are sometimes associated with the choice of research question and the experimental objects; for example:

“The experiment did not include software engineers, which might have different results.” In other cases, there may be comments on the nature of the participants (in particular the use of computing science students); for example,

“The use of a narrow spectrum sample group, although providing an adequate depth to the results, does limit the relevance of results to a small demographic of users, and as such the results cannot be applied to the broader population.”
Review of HF-16-4827: Sequential Revision of Belief, Trust Type, and the Order Effect

In this paper the authors present an experimental investigation of the “order effect,” a phenomenon with a long history of research in the judgment and decision making (JDM) literature, and to a lesser extent, in human factors. The general idea is that the order in which information pertaining to a judgment, and more specifically, order coupled with whether successively presented items of information either confirms or disconfirms the human judge’s current belief, creates variance in the human’s ultimate judgment. A major reason for interest in this phenomenon is that this variance is non-normative, in that, assuming a static, unchanging judgment environment or task, normative models presume that an ultimate judgment should be independent of the arbitrary order in which items of information are presented to, or received by, the human judge. In contrast, in a dynamic environment or task, an order effect may in fact be normative due to the possibility that more recent information is likely to be more relevant, etc.
In keeping with decades of related research, the experimental task described (judging an “attack probability” in a fictional military scenario), was static, in that there was no reason for the first item presented to a human judge to be any less or more informative or trustworthy than the second (the study used two sequential information sources). The authors provide a substantial review of the literature in tasks of this type (e.g., explanations of the order effect related to “anchoring and adjustment” heuristics and many related models). What is novel, in the current study, is the addition of an experimental manipulation focused not solely on cognitive factors, but also by the construct of trust, which the authors take to have both cognitive and affective components, or (from line 154): “trust from the head versus trust from the heart.”
The logic of the study is clearly described in the long paragraph on page 8. To operationalize the contrast of “trust from the head” versus “trust from the heart,” experimental participants were variously informed that an item of information came from expert military intelligence officers (head) versus from a longtime trusted friend (heart). Various hypotheses are stated about how experimental participants will differently weight each presented information item as function of both its source (expert versus friend) and whether the information tends to either confirm or disconfirm the participants’ current judgment. A factorial experiment, crossing information type (source), information item presentation order, and disconfirming/confirming nature, was performed to test the authors’ various hypotheses about how these factors contribute to, and combine together, to result in the hypothesized effects on participants’ ultimate judgments.
A first pass of statistical analysis was conducted, omitting the trust (or source) (friend versus expert) information, that revealed that the pattern of results were consistent with findings from a nearly identical study on the order effect by Entin and Serfaty (1997). Notably, both the Entin and Serfaty study employed task-naïve college students as experimental participants (N=10), while the current study employed 70 task-naïve college students as participants.

Having replicated the findings of this earlier study that did not include the possible “head versus heart” trust dimension, a second analysis using a separate ANOVA was performed focusing on the hypothesized trust manipulation. This analysis revealed a mixed, yet generally supportive pattern of results given the authors’ “head versus heart” trust hypotheses.
A first pass of statistical analysis was conducted, omitting the trust (or source) (friend versus expert) information, that revealed that the pattern of results were consistent with findings from a nearly identical study on the order effect by Entin and Serfaty (1997). Notably, both the Entin and Serfaty study employed task-naïve college students as experimental participants (N=10), while the current study employed 70 task-naïve college students as participants.

Having replicated the findings of this earlier study that did not include the possible “head versus heart” trust dimension, a second analysis using a separate ANOVA was performed focusing on the hypothesized trust manipulation. This analysis revealed a mixed, yet generally supportive pattern of results given the authors’ “head versus heart” trust hypotheses.
The paper concludes with a discussion of potential implications in health care, Internet searching, and many other activities. I found this discussion to be highly speculative, as compared to the generally positive way I judged how the many aspects of experimental methodology, internal validity, and statistical analysis were handled in this study. When one turns to considering what implications, if any, this study has for contexts, tasks, and judges beyond those employed in the experiment reported here, I have some serious concerns about external validity issues, or the generalizability and robustness of the central findings.
In particular, and based not solely on this study but also in my long experience in trying to generalize experimental findings from task-naïve participants in cognitive tasks to applied or operational settings, I am highly dubious about whether these findings will be relevant to understanding or supporting human judges (e.g., in the military, healthcare, transportation, etc.) who use not merely a few sequentially presented information items (2 in the current research), but also typically a considerable amount of task-relevant knowledge gained through education, training and experience in the performance of their tasks.
In contrast, the participants in the present study were denied access to any similar or analogous knowledge due to the selection of an experimental task with which they likely had no experience (making probability of military attack judgments). I do recognize that some other cognitive researchers with less invested in practical application and robust generalization from the laboratory to meaningful rather than fictional settings may be less concerned by what I believe are serious deficiencies of this research associated with external validity. If so, I suggest that this manuscript might meet with a better reception by the readership of a JDM or psychological science journal than by the readership of *Human Factors.*
There are three aspects to presenting results:

- describing the method of analysis used,
- presenting the results of applying the method, and
- explaining the results.

For quantitative data, the statistical analysis methods that have been used should be clearly stated.

When qualitative data have been analysed, the methods used need to be described, including information about how many independent people took part in the coding analysis and what process was followed to reach agreement.
Many reviewers know a great deal about statistics (more than you and I). Some will get excited about your data and suggest other possible analyses that could be performed. Some will complain that you have not done your statistical analysis the way that they would have done it (and therefore the entire research is questionable). Conversely, many reviewers know little about statistics and will simply accept what is presented.
My approach is to focus on the research questions (rather than on the data), to let the questions guide the analysis, and not to perform any analyses that are not required for shedding light on the research questions (even if the data permit). This approach is not always agreed on, and reviewers may question why further analyses were not performed (e.g., “as there are quite a few independent variables, it would be interesting to see a complete model comprising all independent variables and all interaction effects”). If your research questions are clearly listed at the start of the paper, and your description of the analysis method is clearly related to each research question, you should be able to mitigate against comments like this by making sure your article tells a complete “research story.”
6.3.2 Presenting the data

It is common to use data visualisation techniques for presenting the quantitative data (e.g., bar charts, scatter plots, box plots, histograms) (as shown in Figure 5.3). Some simple rules will help ward against criticism:

- bar charts representing similar data should have the same scales;
- bar charts should always have a zero origin (or if it is necessary to shift the origin, then this should be clearly highlighted in the caption);
- line charts should only be used for data where the x-axis represents ordered categories (not nominal data);
- pie charts and 3D graphics should be avoided;
- all axes should be labelled, and the units used should be clear;
- don’t clutter the graphic with unnecessary clutter (as advocated by Tufte (2001)).
6.3.3 Making conclusions: separating fact from opinion

It is easy to (consciously or unconsciously) overgeneralise the experimental results when stating conclusions, and thus mislead readers (Zhai, 2003). However, if all experimental facts (tasks, experimental objects, procedure, analysis) are described completely, clearly, and unambiguously, then the readers will be able to form their own conclusions. Your conclusion is simply one possible interpretation of these facts.

The data and its analysis are therefore the objective facts, whereas the conclusions are your interpretation of the facts. It is therefore useful to clearly separate the presentation of the hard facts from your opinion of what they mean. By doing so, you acknowledge that your conclusions are your interpretation of the facts, and you therefore make it possible for other readers to interpret the facts in their own way. This means that readers who do not agree with your interpretation can clearly identify the facts and make their own judgement, independent of their disagreement with your opinion. Allowing a reader to make their own judgement of the facts is difficult if the fact is not clearly distinguished from the opinion in the report.
Acknowledging Limitations

It is important that the limitations of the experiment are faithfully acknowledged. All experiments are limited, and no experiment is perfect. The limitations may be caused by the choice of tasks, experimental objects, participants, unexpected incidents that occurred in the experimental process, confounding factors, etc. These limitations will constrain the boundaries within which your conclusions can be generalised.

Two concepts that are sometimes used in discussing limitations are reliability and validity, and particularly in the empirical software engineering literature, limitations are often discussed in terms of “threats to reliability” and “threats to validity.”
Validity (accuracy) is the extent to which the experiment correctly addresses the specified research questions. Has the experiment been conducted in a manner that allows appropriate conclusions to be made?

There are two types of validity: internal and external. They are defined as follows:

- **Internal validity** relates to the design of the experiment. Has it been designed appropriately with respect to randomisation, controls, data collection methods, and experimental process? Can the effect on the dependent variable(s) be attributed to the changes in the independent variables (and not to any other intervention)?

- **External validity** relates to the generalisation of the results. Would these results hold for other participants, or to other experimental objects and tasks? The external validity of an experiment can be improved by:
  - using large random samples from the population as the participants and employing wide advertising methods for recruiting participants;
  - increasing the number and range of tasks and experimental objects used in the experiment.
Summing up your Findings:

6.3.5 Explaining the results

Having made conclusions based on your results, it is important to relate them, as much as possible, to the world outside your particular experiment: you need to provide some explanation as to why you believe that your results are the way they are. It may be useful to explain the results by referring to existing theoretical models (typically, psychological, perceptual, or interaction models) and discussing how your experiment has reinforced them (or not). Similarly, your results may be discussed in relation to prior experimental results in the same or a similar area. Doing this may be tricky if your results do not say what you expected or appear to conflict with prior results. In these cases, addressing obvious limitations and constraints may help with explaining the results obtained. Explaining your results allows you to show how they contribute to an exploration of ideas that is wider than your own specific research question.