1
Introduction

This book describes the process that takes a researcher from identifying a human–computer interaction (HCI) research idea that needs to be tested, to designing and conducting a test, and then analysing and reporting the results. This first chapter introduces the notion of an "HCI idea" and different approaches to testing.

1.1 Assessing the worth of an HCI idea

Imagine that you have an HCI idea, for example, a novel interaction method, a new way of visualising data, an innovative device for moving a cursor, or a new interactive system for building games. You can implement it, demonstrate it to a wide range of people, and even deploy it for use – but is it a "good" idea? Will the interaction method assist users with their tasks? Will the visualisation make it easier to spot data trends? Will the new device make cursor movement quicker? Will users like the new game building system?

It is your idea, so of course you believe that it is wonderful; however, your subjective judgement (or even the views of your friends in the research laboratory) is not sufficient to prove its general worth. An objective evaluation of the idea (using people not involved in the research) is required. As Zhai (2003) says in his controversial article, "Evaluation is the worst form of HCI research except all those other forms that have been tried," the true value of the idea cannot be determined simply by "subjective opinion, authority, intimidation, fashion or fad."

If your idea were, for example, a new constraint satisfaction algorithm that you believe is faster than others, you could deploy vast amounts of computing power to run computational tests on different data sets to prove your point. Unfortunately, ideas that necessarily include human activity cannot be tested

so easily – there is no "ISO standard human" (Lieberman, 2003) against which the idea can be tested to prove its worth for all humanity. Because we do not have robust and complete models or theories of human behaviour against which we can test the idea, the participation of other humans is needed to assist us in checking our own intuitions (Zhai, 2003).

The phrase "HCI idea" is used here very generally to mean any idea that, once implemented and used, involves interaction between humans and technology. Possible categories of HCI ideas, with examples, include the following:

1. A type of interaction device associated with new hardware:
   - stylus input, vibratory output, olfactory output.
2. A method of perception using devices that target particular senses:
   - feedback mechanisms (e.g., vibration, sound, peripheral vision);
   - visualisation of data (e.g., different types of data charts, colour coding methods).
3. A method for system interaction usually embedded within other software:
   - direction of text scrolling (e.g., horizontal, vertical, page turning);
   - navigation method (e.g., site map, tabs, hierarchy diagram).
4. An interactive system designed to support a complex task:
   - a system to manage the proposal, preference collection, and allocation of student projects;
   - a sketch-based system for drawing project management scheduling charts.

Occasionally, new HCI ideas are quickly adopted into everyday use by a large number of people, and this wide-scale adoption is sufficient proof of their worth. For example, it seems superfluous to try to prove that mobile phones are more convenient than pay phones, that Google's interface is sufficient for online search tasks, or that Facebook is a good way for people to keep in touch.

Unfortunately, most HCI researchers (and, in particular, PhD students) do not have the luxury of time to wait and see if their novel idea takes off and is adopted at large. Proving that their HCI idea is a "good" one will need to be done in the context of a test that involves other people who try out the idea.

There are two different types of HCI test: formal comparative *experiments* and exploratory usability *evaluations*.

Experiments are objective tests that aim to demonstrate that the idea produces better results than an existing idea that performs the same function. Experiments are more appropriate for ideas in categories 1–3 in the preceding list (i.e., small, specific ideas for which alternatives can readily be found and that are usually associated with well-defined tasks). The outcome of an experiment is a conclusion indicating which idea results in better user performance. As such, experiments can be considered summative, although in many cases the

experience of running the experiment reveals useful improvements, and so also contributes to a formative process (Ellis and Dix, 2006).[1]

Evaluations are exploratory tests that aim to show that the idea works in practice, in the context of typical uses. Such evaluations (also called "usability studies") are more appropriate for category 4 in the preceding list (i.e., larger, more complex pieces of interactive software for which it is difficult or impossible to find alternatives that fulfil identical functions and that typically support a wide range of user tasks). The outcome of an evaluation is a list of suggestions for system improvement (as part of a formative test), or it can be confirmation that the system performs its function sufficiently well for it to be deployed.

1.2 Experiments: Assessing worth by comparison

It is not uncommon for papers to be written that report that an idea is "good" because, for example, experimental participants reached a "high" level of performance. An experiment that investigated the parameters that could be used for representing information in vibrotactile devices (Brown, Brewster, and Purchase, 2005) reported an overall 71% recognition rate, with tactile "rhythms" being correctly identified more than 90% of the time and tactile "roughness" identified 80% of the time. Although these are impressive and interesting results, they do not tell us whether presenting information in a tactile manner is better or worse than any other medium, or whether the 80% recognition rate of roughness is sufficient for practical use – it may be the case that this rate is actually too low to be useful. In experiments like these, unless you get a result of 100%, it is difficult to make definite claims about the worth of the idea.

HCI experiments are therefore typically about comparison. They aim to prove that one HCI research idea is better than another that fulfils the same function. Note that it may be the case that neither (or both) of the ideas may be the experimenter's own, and neither (or both) may be new. What is key is the idea of comparing the "goodness" of one idea with another by measuring their relative performance.

This is not to say that experiments should be entirely focussed on a single conclusion; indeed, the value of running an experiment often arises from the process of conducting it – defining its rationale and motivation, deciding between the appropriateness of different experimental methods and activities, and investigating the different types of data collected.

[1] Summative tests are those whose only aim is to produce a conclusion; formative tests are those whose intention is to make recommendations for improvement.

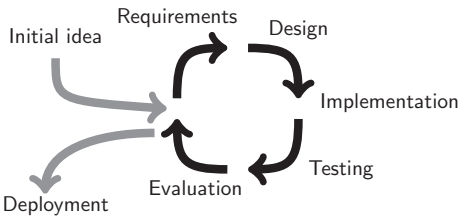

Figure 1.1: Iterative system development cycle.

In some cases, tests that do not entail comparison are the only type that can be run, simply because the technology used is so advanced that there are no viable alternatives, the HCI idea is extremely revolutionary, or the alternatives are so different that any comparison would be meaningless (e.g., comparing speech recognition technology with keyboard typing). In such cases, proving the worth of the idea may rather be done through descriptions of the design rationale, existence proofs, detailed scenarios of use or case studies, or participatory critique (Greenberg and Buxton, 2008).

An experiment is here defined in terms of comparison: a test that pits one or more alternatives against each other.

1.3 Evaluations: Assessing worth by use

Evaluations typically focus on determining the usability of a new interactive system. "Usability" is defined by the International Organization for Standardization (ISO) standard for the "ergonomics of human system interaction" (ISO 9241-11:1998; ISO 1998) as the "Extent to which a product can be used by specified users to achieve specified goals with *effectiveness*, *efficiency* and *satisfaction* in a specified context of use" (my emphasis: p 2). Evaluations entail potential users using the system, and the recording of their activities and comments.

Evaluations are an important stage of the iterative system development cycle (Figure 1.1). They not only produce useful feedback and suggestions for improvement that feed into the next requirements and design stages, but they are also crucial in determining when sufficient iterations have been completed and the system is ready to be deployed. In some cases, the system is deployed after testing as part of an external evaluation process; in this case, beta versions of software are released for a limited period of time so as to obtain feedback from a more extensive set of external potential users.

It is not usual for evaluations to be comparative as they tend to focus on one system and the experience of its potential users, their purpose being to provide feedback to improve a system within a development cycle. Often, they may not be considered "research" and may not be publishable in the serious research literature. However, evaluations are often part of a larger research project where once the system has been shown to be ready for deployment, it may subsequently be used in an experiment to demonstrate its worth in comparison against its competitors – this is more likely to be considered a research activity.

1.4 Focus of this book

It is tempting to associate other common terms with experiments and evaluations as a way of distinguishing them: experiments are sometimes considered "formal," producing "quantitative," "objective" results, and evaluations are "informal," producing "qualitative," "subjective" results.

Using these terms in this way can be misleading because evaluations can be formal and produce quantitative and objective results, whereas experiments may collect subjective data (although experiments are unlikely to be "informal"). The key distinguishing features of these two types of HCI test are as follows:

- Experiments are comparative and focus on producing data to demonstrate the worth of an HCI idea;
- Evaluations are not comparative and focus on producing feedback to either improve a system or confirm its readiness for deployment.

This book considers both experiment and evaluation tests as it follows the process from designing the test, conducting it, and collecting and analysing the data, to reporting the results.

Getting the design of an experiment right the first time is more important than doing so for an evaluation. An evaluation is a stage within the iterative development cycle, and its outcome (the feedback on system usability) feeds into the next design stage. This feedback helps designers in making design choices within given constraints. During the development of the system, several evaluations may be performed, perhaps with only a few potential users[2] and tasks in the first iteration (when it is likely that substantial changes will be suggested), an increased number in the second, and a much larger group when

[2] We distinguish here between "participants" (who take part in experiments) and "potential users" (who take part in evaluations).

the development team believes that the system is nearly ready for deployment, and only needs tweaking. Even if the design of an evaluation is flawed (e.g., by forgetting to ask potential users to use a particular contentious interface feature or to comment on a novel data entry method), useful feedback will still undoubtedly be produced from this evaluation, and the next round of evaluations can correct the previous flaws.

In contrast, an experiment is a one-off activity. It requires the cooperation of a large number of participants, all following the same experimental process. Once it has started, it cannot be interrupted to correct any flaws in its design – not unless the whole experiment is to be redesigned and run again. It is therefore important that an experiment be very carefully designed because errors can cost a great deal of time and effort.

The primary focus of this book is therefore on experiments because their design is more complex and more risky than evaluations. In addition, good experimental design requires knowledge and skills typically not taught in a computer science or software engineering curriculum in the same way that usability evaluation is often covered as part of the iterative design cycle. All chapters in the book, however, conclude with a relevant section on evaluations.

1.5 Structure of this book

The chapters of this book follow the process of designing, conducting, analysing, and reporting experiments and evaluations, with the focus on experiments. The book is intended to be read from start to finish (perhaps with the exception of Chapter 5 on statistics), and preferably prior to designing an experiment, rather than used only as a reference book. Throughout the book, details are given of example experiments: most of these have been conducted by the author, and so provide a personal insight into the actual processes that led to their design and implementation – such insight is not typically available from secondary source published papers.

Chapter 2 highlights the importance of defining a clear research question. It describes the process of designing the experiment and includes a discussion on the generalisability of the results (in particular, with reference to the choice of experimental objects and tasks).

Chapter 3 describes the practicalities of conducting the experiment, based around a discussion of the participant experience. It includes hints on recruiting and managing participants, conducting pilot tests, adhering to ethical requirements, and performing pre- and postexperiment activities.

Chapter 4 deals with the collection of data, describing the range of different data types that can be collected, and focuses on methods for collecting and analysing qualitative data.

Chapter 5 addresses the analysis of quantitative data and presents a selection of statistical tests useful for analysing data produced by comparative experiments.

Chapter 6 gives advice on how best to report the results of the experiment for publication, proposing an appropriate structure for the report and accounting for the limits of the generalisability of the results. It also includes common reviewers' comments.

Chapter 7 discusses possible problems and pitfalls in running HCI experiments, and how to address them.

Chapter 8 concludes the book by presenting "six key principles" for conducting HCI experiments, as well as a model of HCI experimentation.

2

Defining the research

The first step in running an experiment is defining what you want to discover and how you will do so. This chapter presents an approach to experiments that begins by first defining a research question, and then basing the definition of the conditions, experimental objects, and tasks on that question. These elements will ultimately define the form of the experiment.

Several key concepts used throughout the book are introduced and defined in this chapter:

- *The research question:* a clear question that succinctly states the aim of the research;
- *Conditions:* the ideas of interest – these will be compared against each other;
- *The independent variable:* the set of conditions to be used in the experiment – there will always be more than one condition;
- *The population:* all the people who might use the idea; *the sample:* the set of people who will take part in the experiment;
- *Generalisability:* the extent to which experimental results can apply to situations not explicitly included in the experiment itself;
- *Experimental objects:* the way in which the ideas are presented to the participants – experimental objects embody the conditions so that they can be perceived;
- *Experimental stimulus:* the combination of an experimental object and a condition;
- *Experimental tasks:* what the participants will actually do with the experimental objects;
- *Experimental trial:* the combination of a condition, an experimental object, and a task.

2.1 The research question

Experiments are often run within the context of wider research projects.[1] Although these projects may have broad aims, for example, "Investigating the use of head-mounted eye gaze equipment" (as in San Augustin et al., 2009) or "Designing alternative methods for menu design," it is important that each individual experiment be clearly defined by a research question – with a clear "?" at the end. Defining a clear research question upfront is crucial to focussing the study, and it is the first step to ensuring that your experiment is designed to discover what you actually want to find out! A useful side effect of expressing your experiment aim as a clearly defined question is that it makes it much easier to explain your research interests to outsiders.

Examples of inappropriately phrased research questions are as follows:

- "To investigate the use of a visual mouse in a text reading task";
- "Asking people to draw graphs using a visual mouse and seeing if they like it";
- "Seeing if the visual mouse works."

These could be better stated as follows:

- "Is reading a piece of text using a visual mouse more efficient than when using a physical mouse?"
- "Do users prefer a visual mouse to a physical mouse when drawing graphs?"
- "How accurate is the use of a visual mouse when performing fine-grained interaction tasks?"

You can see that the latter three examples, expressed as questions, are much more focussed and include details (e.g., "more efficient," "drawing graphs") that will become important features of the experimental design.

Some researchers, especially those with a psychology background, prefer to express their experiment aims in terms of a null hypothesis statement that they will ultimately try to reject as being false (e.g., "There will be no efficiency difference when reading a piece of text between a visual mouse and a physical mouse"). Although this is a valid approach, I find that starting off with a clear, focussed research question is a better (and often less confusing) starting point.

[1] Ideally, of course, the experiment or evaluation should be conducted by someone (or a team) who has not been associated with developing the HCI idea, although this is seldom the case for academic research projects. Lieberman (2003) points out that it would be unthinkable for a new medical technique to be evaluated by the person who developed it.

Thus, before commencing the design phase of an HCI experiment, you need two things:

- A clearly defined HCI research idea – this may be a technique, method, technology, system, etc.;
- A clear research question that defines how the worth of this idea will be investigated.

2.2 Conditions for comparison

As mentioned in Chapter 1, the key to HCI experiments is the notion of comparison: we compare the performance of one HCI idea against another. One or more alternative ideas need to be identified. Importantly, the alternatives must offer the same functionality as the idea you want to test; otherwise, a comparison is unfair (as in "comparing apples to oranges").

The different alternatives (including the idea to be tested) are called the *conditions*. In many cases, alternatives will be easy to identify, especially if the new idea was devised as an improvement to an existing system. For example, if the idea is a new touch-based interaction method for turning pages when reading text on a screen, then an obvious alternative will be the common existing method of vertical scrolling. The experimenter might also want to include horizontal scrolling as one of the conditions, even though this is not so common.

In other cases, alternatives may need to be contrived. For example, if the idea is the use of olfactory output to present information about the people and situations in a set of photographs to the visually impaired, then an alternative, more common method for presenting this information (e.g., voice recordings) may need to be devised to enable appropriate comparison.

An *independent variable* comprises a set of at least two conditions; in the previous page turning example, the independent variable is "the method of turning pages", and its three conditions are "touch", "horizontal", and "vertical". These three methods allow for all text to be accessed, and so have equivalent functionality.

The experimenter has control over the conditions that comprise the independent variable, and they must be defined in advance of the experiment. They must

- be clearly defined;
- permit equivalent functionality.

Looking forward to the nature of our results (and it is always a good idea to look forward to the required form of the results at the end of the experiment), what