1 Controlled experiments

ANN BLANDFORD, ANNA L. COX AND PAUL CAIRNS

1.1 Overview

Controlled experiments, an approach that has been adopted from research methods in psychology, feature large in the arsenal of HCI research methods. Controlled experiments are a widely used approach to evaluating interfaces (e.g. McGuffin and Balakrishnan, 2005) and styles of interaction (e.g. Moyle and Cockburn, 2005), and to understanding cognition in the context of interactions with systems (e.g. Li *et al*., 2006). The question they most commonly answer can be framed as: does making a change to the value of variable X have a significant effect on the value of variable Y? For example, X might be an interface or interaction feature and Y might be time to complete task, number of errors or users' subjective satisfaction from working with the interface. Controlled experiments are more widely used in HCI research than in practice, where the costs of designing and running a rigorous experiment typically outweigh the benefits.

The purpose of this chapter is to outline matters that need to be considered when designing experiments to answer questions in HCI.

1.2 The method

We have structured this section about how to design and run a controlled experiment in the order that the information is usually reported within the method section of a paper or project report; that is, first we will consider how to go about choosing the participants who will take part in the experiment before moving on to consider designing the experiment itself, assembling the materials and apparatus required, and finally the procedure. We hope that this approach will help you to find your way around papers that are written up in this way and also help you when you are considering designing your own experiments.

1.2.1 Participants

For any experiment it is necessary to consider what the appropriate user population is. For example, if the experiment is designed to test the effect of a changed display structure for a specialist task, for instance a new air traffic control system,

https://doi.org/10.1017/CBO9780511814570.002 Published online by Cambridge University Press

it is important to recruit participants who are familiar with that task, namely experienced air traffic controllers. Similarly, if the concern is with an interface for older users, it is important to recruit such users to the study. Ideally, for any experiment, a representative sample of the user population is recruited as participants; pragmatically, this is not always feasible (also, it is so much easier to recruit friends, students or members of a psychology department participant database). If a non-representative sample of users is involved in the study then the consequences of this for the findings should be carefully considered. For example, how meaningful is it to have run an experiment on an interface intended for air traffic controllers with undergraduate psychology students? Probably not at all. Sometimes the target population is hard to define. Who, for example, is the audience for a government benefits website? In that case, undergraduate psychology students might not be a bad starting point to begin to study the website.

Having decided on the user population, decisions need to be made on how many participants to recruit, depending on factors such as the power of the statistical tests to be used, the time available for the study, the ease of recruiting participants, funds or other incentives available as participant rewards and so on. Participants can then be recruited through direct approach or by advertising in suitable places.

1.2.2 Ethical considerations

Although not usually reported explicitly, one important consideration is the ethical dimensions of any study. Most professional bodies (e.g. British Psychological Society, 2006) publish codes of practice. Less formally, Blandford *et al*. (2008) have proposed that the three important elements of ethical consideration can be summarised by the mnemonic 'VIP':

- Vulnerable participants
- Informed consent
- Privacy, confidentiality and maintaining trust

Examples of vulnerable participants will include obviously vulnerable groups (such as the young, old or infirm), but may also include less obvious people such as those with whom the investigator has a power relationship (e.g. students may feel obliged to participate in a study for their professor), or who otherwise feel unable to refuse to participate for any reason, or who might feel upset or threatened by some aspect of the study. Some concerns can be addressed simply by making it very clear to participants that it is the system that is being assessed and not them.

It is now recognised as good practice to ensure all participants in any study are informed of the purpose of the study and of what will be done with the data. In particular, the data should normally be made as anonymous as possible (e.g. by using codes in place of names) and individuals' privacy and confidentiality need to be respected. It is now common practice to provide a (short) written information sheet about the experiment and to have a consent form on which participants can indicate that they understand what is expected of them,

that they are participating voluntarily and that they are free to withdraw at any time without giving a reason. This is informed consent – a person agrees to take part knowing what they are getting into.

Usually it is possible to offer participants the opportunity to talk about the experiment in a debriefing session after they have finished the tasks they were set. Not only does this help to make the participants feel valued, but sometimes it can be a source of informal feedback that can lead to a better design of experiment or even new ideas for experiments. All data should be stored in accordance with legislation; for example, in the UK the Data Protection Act specifies what information can be held and for what reasons, and it is necessary to register with the government if the data being stored allows individuals to be identified.

1.2.3 Design: dependent and independent variables

A controlled experiment tests a hypothesis – typically about the effects of a designed change on some measurable performance indicator. For example, a hypothesis could be that a particular combination of speech and keypress input will greatly enhance the speed and accuracy of people sending text messages on their mobile. The aim of a classical experiment is, formally, to fail to prove the null hypothesis. That is, for the texting example, you should design an experiment which in all fairness ought not to make any difference to the speed and accuracy of texting. The assumption that there will be no difference between designs is the null hypothesis. By failing to show this, you provide evidence that actually the design is having an effect in the way that you predicted it would.

Put more generally: the study is designed to show that the intervention has no effect, within the bounds of probability. It is by failing to prove that the intervention has had no effect – that the probability of getting this result if the intervention has no effect is very small indeed – that one is led to the conclusion that the intervention did indeed have an effect. More formally, using the terminology defined below, the failure to prove the null hypothesis provides evidence that there is a causal relationship between the independent and dependent variables. This idea is discussed at greater length in Chapter 6.

In an HCI context, the changes to be made might be to interaction design, interface features, participant knowledge, and so on. The variable that is intentionally varied is referred to as the *independent variable* and that which is measured is the *dependent variable*. One way to try to remember which way round these are is to think that the value of the dependent variable *depends* on the value of the independent variable. There may be multiple dependent variables (e.g. time to complete task, error rate) within one experiment, but – at least for simple experiments – there should normally only be one independent variable.

One of the challenges of experimental design is to minimise the chances of there being *confounding variables* – variables that are unintentionally varied between conditions of the experiment and which affect the measured values of

https://doi.org/10.1017/CBO9780511814570.002 Published online by Cambridge University Press

the dependent variable. For example, in testing people with different interfaces for text message entry, it could be that you use different messages for people to enter in the different interfaces. The length of message clearly has an effect on how long it takes to enter a message, regardless of the interface. Thus, the message length is a possible confounding variable. Another one might be the complexity of entering certain words. In designing the experiment, you would need to do something to make sure that any differences in text message entry time were solely due to the interface and not to the messages people had to enter. The simplest thing would be to make sure that every message was entered on all the interfaces. This way, even if the messages did take different times, over the course of the whole experiment the effect of the different messages would be evenly spread out across all of the interfaces. This is called counterbalancing.

In designing an experiment, then, the aims are to vary the independent variable in a known manner, to measure the dependent variable(s) and to minimise the effects of confounds on the outcomes of the study.

Within HCI, there are various techniques for minimising the effects of possible confounds. An important starting point is simply to eliminate as many confounds as possible from the experiment, such as those relating to different rooms or different computers. Another approach is to randomise variables wherever possible; for example, if time of day is likely to have an effect on results then either run all experiments at the same time of day (which is likely to be impractical) or randomise the timing of trials so that participants under each condition take part at different times of day.

One particular set of confounds to be aware of is individual differences. This is a general name for how people differ from each other. There are obvious things like the physical differences between men and women but the term also covers a huge range of things such as differences in personality, aesthetic sensibilities, cognitive skills and so on. It is clearly not possible to control for all individual differences in an experiment, but it is advisable to control for the most likely factors that might influence performance or attitudes, such as age, sex and education level. To do this, we must avoid putting all the men in one group and all the women in the other, or having all the older students in one group and all the younger ones in the other. Of course, there might be experiments in which such a difference is an independent variable, for example testing a hypothesis that there will be performance differences with a given interface between women and men or older and younger users. In such cases, a particular difference will be the independent variable – but more on that later!

1.2.4 Design: 'within subjects' or 'between subjects'

Some experiments are most appropriately conducted 'within subjects' and others 'between subjects'. A within-subject experiment involves each participant performing under all sets of conditions, whereas a between-subject experiment

involves each participant only performing under one condition. So in a study to compare three websites we might choose to have everybody use all three websites, and that would be a within-subject design; or each participant might use only one website, so that would be a between-subject design. A third design that is common in HCI research is a 'mixed factorial' design in which one independent variable is within subjects and another between subjects. This would not mean that some participants use only one or two of the websites! Instead, there are two independent variables (also called factors in this case) and one factor is within and one factor is between. So if we were comparing differences between the three websites but also between how men and women use the websites, we could have the website as a within factor, so that each person used all three websites. But obviously, the other factor would be a between factor with each person being either a man or a woman but not both.

Note that the terms 'within-subject' and 'between-subject' are a throw-back to the days when the people who took part in experiments were referred to as 'subjects', a term that is now considered inappropriate as it implies a power relationship between experimenter and experimentees and an objectification of the people who take part in experiments. (Indeed, in even earlier times, people who took part in experiments were sometimes called experimental material!) In this enlightened age, the people who take part in an experiment are 'participants', not 'subjects'.

When is it best to choose a within-subject design and when a between-subject one? This is a difficult question to answer. It depends on whether participants are required to compare interfaces (in which case a within-subject design is essential); whether there are likely to be unwelcome learning or interference effects across conditions (in which case a between-subject design is essential); what statistical tests are planned; how long each participant is likely to take in completing each part of the experiment (the longer each task, the less attractive a within-subject design is); and how easy it will be to recruit participants to the study (the more people can be recruited, the more feasible a between-subject design is). One advantage of within-subject designs is that individual differences are less likely to influence results; disadvantages include possible learning effects and more complicated statistics.

In a within-subject design, participants will typically be required to repeat a very similar procedure multiple times with different values of the independent variable. In many cases, it would be inappropriate to repeat literally the same task for each condition, so it is advisable to generate multiple tasks, one for each condition, for each participant to perform. The task becomes, in effect, another independent variable, but one that is of no direct interest in the analysis. The different values are sometimes referred to as 'levels'; for example, in the experiment described in section 1.3 there are two independent variables, the mode of input for message entry (which has three levels) and the mode of input for text entry (which has two levels). Each combination of levels that a participant engages with is referred to as a 'condition', so in the experiment described there are six conditions.

1.2.5 Apparatus/Materials

Every experiment is conducted using some 'instruments'; most commonly (in HCI experiments) the core instrument will be a piece of computer software. For some experiments it is possible to make use of existing software. For others it is necessary to create your own. Many experimental instruments are computer simulations of systems. It is sometimes possible to make use of an experiment generator to simplify the process of creating a suitable instrument. E-prime is one such system commonly used in psychology experiments. Sometimes it is necessary to exploit or acquire programming skills in order to create prototypes that can measure and record task completion times, keystrokes, etc. and that enable you to manipulate the variables of interest.

1.2.6 Procedure

You should create a formal procedure that describes what the participants do during the experiment. This has two purposes. First, it enables you to make sure that every participant in your experiment has the same experience because it is possible that if you use different procedures between different participants, this could be a confounding variable. For instance, if a study was done to see how older users performed compared with younger, and the older users were treated more deferentially (as would be polite), it is possible that the older people would perform better not because of the user interface but because they were more relaxed and felt happier having met this charming experimenter.

The second purpose of a formal procedure is that it allows other people to replicate your experiment. This is the basis of good science. If other people can replicate a study's findings then collectively science can feel confident that the findings are sound and can build on this in future work. The formal procedure removes confounding variables not only within the one experiment but also between separate attempts at the same experiment.

There are a number of issues that we need to consider when designing the experimental procedure:

- Minimising the effect of confounds by controlling the order in which we test the interfaces, the tasks we ask participants to perform and the context in which the study is run
- Making the experiment robust through careful design of instructions, piloting and careful collection and management of data
- Building up to a bigger series of experiments that probes the phenomenon of interest more deeply.

Minimising the effects of confounds

This typically involves considering in which order interfaces are tested, how tasks are assigned to interfaces and the broader context within which trials take place.

https://doi.org/10.1017/CBO9780511814570.002 Published online by Cambridge University Press

If all participants in a study experience interfaces in the same *order*, there are likely to be performance effects due to learning (for example, improving performance with the second interface if learning transfers from one to the other); there may also be changes in attitude due to novelty effects (with the first interface perceived as more novel than the second).

As noted above, in many experiments it is necessary to give participants well-defined *tasks* to complete with the interfaces; it is usually important to devise different tasks to reduce learning effects and also reduce boredom. However, if particular tasks are always associated with the same conditions, this may affect performance in unexpected (or even undetected) ways.

Sometimes in experiments it is more convenient to run one condition (i.e. one value of the independent variable) in a different *context* from the other, for example, using a different computer or a different room, or conducting the experiment at a different time of day. Such apparently innocuous changes can influence results. For example, the first author was once involved in an experiment to study the effects of a training intervention on students' performance. For various reasons, a within-subject design was used (that is, one in which all participants performed under both sets of conditions). In this case it was clearly not possible to control the order of presentation, since it is impossible to undo any effects of training, so all participants had to do the before-training task first, then undergo the training before the second task set. We controlled for task-set variations by allocating half the participants to each task set for the first test then swapping the task sets over for the second set of tests. However, one variable we could not control for was the fact that the participants had a big party in the evening between the first and second tests, so they were all slightly hung over during the second test. Obviously, this should be avoided if possible.

For some variables, a systematic approach to variation is appropriate. One example is to use a 'Latin square' design; a second is to administer the test in every possible sequence to different participants. A Latin square is a square grid in which every element appears precisely once in each row and each column, where a row represents the order in which test elements are administered to a participant and a column represents the sequence of participants in the study. If only two conditions are being considered, and there is no variation in the task participants are being given, then this randomisation is easily achieved by allocating participants to two groups, balanced as far as possible for age, gender, education level, relevant prior experience and any other variables considered relevant to the experiment. If more than two possible conditions are being considered then a Latin square design might look something like that shown in Table 1.1.

If two different tasks are being administered then the design is usually run as 'mixed factorial' and might be organised as shown in Table 1.2. Here, both the order of presentation of the tasks and the order of presentation of the interfaces are being systematically varied to eliminate possible order effects. It should be noted, however, that the statistical test for analysing the resulting data is a three-factor

Table 1.1 *Example Latin square with four different tasks*

Group	First task	Second task	Third task	Fourth task
i	Task A	Task B	Task C	Task D
ii	Task B	Task C	Task D	Task A
iii	Task C	Task D	Task A	Task B
iv	Task D	Task A	Task B	Task C

Table 1.2 *Organising tests for comparing two interfaces with two different tasks*

Group	First task	Second task
i	Interface 1	Interface 2
	Task A	Task B
ii	Interface 1	Interface 2
	Task B	Task A
iii	Interface 2	Interface 1
	Task A	Task B
iv	Interface 2	Interface 1
	Task B	Task A

repeated measures ANOVA (assuming the data is normal), which is not for the faint-hearted.

Making the experiment robust

This includes ensuring clear and consistent task instructions, piloting the experiment to ensure that people behave (roughly) as anticipated and making sure all recording equipment is working properly.

It is important to decide how to *describe tasks* to participants. Some tasks – such as those described by Moyle and Cockburn (2005), which are small tasks involving simple gestures – are easily described to participants. Others require more extensive task descriptions to be given to participants. It is important to consider what level of detail is to be given: if the focus is on whether participants can make sense of interface features then minimal instructions are most appropriate, whereas if the concern is with how actions are performed then greater detail is likely to be required. It is important that tasks are kept to a reasonable time limit, and that they are interesting and engaging enough to keep participants' attention (so that what is actually being measured is that which is intended and not some result of boredom).

It is usually advisable to *pilot* any new experimental design (i.e. to run it with a very small number of participants) to check the design. For example, it is essential to check that the instructions to participants are clear and that participants can complete the experiment in a reasonable time (typically no longer than an hour,

and even then there should be opportunities for breaks at realistic intervals). It is also helpful to discover early on if any participants behave in unexpected ways that will result in the experiment not delivering the data intended. These early participants can be asked to give feedback on the experimental design to improve it and their data should be discarded. You might use the data to pilot test the statistical tests you intend to use, recognising that data from a very small number of participants are not going to yield statistically reliable results.

However well designed the experiment, there are always things that might go wrong. These include failures in *recording and retaining data*. Common difficulties are: essential data not being captured (hence the need for piloting the analysis as well as the data recording); recording being lost due to equipment or software failure; and data being subsequently lost. Care should be taken over making sure that you know exactly how equipment works, have sufficient recording media (e.g. tapes, disks or memory cards), have everything fully charged up (or with enough new batteries) and are alert to other kinds of equipment failure. Once gathered, data should be stored securely and systematically so that they can be easily retrieved for analysis and review purposes.

Bigger investigations

To develop a good understanding of a phenomenon, it is usually necessary to investigate it in more than one way, leading to a series of linked experiments each of which involves a single controlled manipulation. It is almost invariably more reliable and easier (if more time consuming) to conduct a series of linked experiments than to increase the complexity of one individual experiment.

1.2.7 Analysis

In controlled experiments the focus is on quantitative data (see Chapters 2 and 7 for approaches to gathering qualitative data). In quantitative analysis, dependent measures might include some or all of: time; errors; particular action types; and user satisfaction ratings. These may be measured by one or more of: automatic logging of user actions; external (video or audio) recording; questionnaires (typically giving numerical ratings); and (less usually) interviews.

Before running any experiment, it is important to decide, at least provisionally, what statistical tests will be performed on the data. Plans may subsequently need to change, for example if data that were expected to have a normal distribution turn out to have a surprisingly different distribution (see Chapter 6 for more details). The choice of statistical test will influence both the detailed design of the experiment and the decision about how many participants to recruit.

1.2.8 Are you addressing the question?

It may seem odd to raise this issue at this point; however, it is an important question. Having gone through all the fine details of designing the experiment, it

is easy to lose sight of the purpose of the experiment. And in adapting the original idea in order to avoid confounds, addressing issues from piloting, setting up the apparatus and so on, it is possible the original design has changed in subtle ways that actually mean that it is not addressing the question intended.

This is best understood with an example. Suppose a researcher was interested, as Dearden and Finlay (2006) were, in studying the effects of using interface patterns in design processes. This sort of question could be answered by having many groups of people doing a design project, some using patterns and some not. To measure the value of patterns, the design produced by each group might be evaluated by a set of users who were not involved in the design. So far so good. But in practice the researcher could not use real designers, partly because they are hard to get hold of and partly because it is unrealistic to expect real designers to waste time producing different systems intended for the same purpose. Instead, then, the researcher might settle for using different groups of students on an HCI course.

For the evaluation, the researcher would want to make sure that there are clear quantitative measures because that is what experiments are all about, so the evaluation might involve timing users in how long they take to achieve various tasks using the different interfaces that the student groups produced.

This all sounds very plausible but now think again about the research question. Suppose the researcher found a significant difference between how users performed with the different designs. Would this tell you anything about the use of patterns in the design process? Actually, all it really tells you is that different groups of students produced different designs that were differently usable! With some careful argument, the researcher may be able to attribute some of the design differences to the patterns, but that would be more of a qualitative argument rather than the conclusion from the experiment. The difference in performance of the users could be attributed to the differences between the groups, the lack of experience of the students in design, the difficulty of learning patterns for inexperienced designers and so on. These are all confounds that undermine the point of the experiment.

One way to avoid losing the plot of the experiment could be to write up the report before devising the experiment. This is the approach advocated in Chapter 10 on writing. That way, as the details of the experiment become clearer, you can check to see if it still fits with the story that you set out to write in the first place.

1.3 Applying the method

The study reported here (Cox *et al.*, in press) is also used as the example in Chapter 6 on statistics. You will therefore be able to use the two chapters together to see how the experimental design and statistical analysis support each other in ensuring a good quality result.

As the mobile phone develops and its set of capabilities grows richer, there is a constant pressure to evolve the user interface and develop more efficient,

https://doi.org/10.1017/CBO9780511814570.002 Published online by Cambridge University Press

convenient input methods. Recent attention has focused on changing handset layout (Hirotaka, 2003) or improving word-prediction software (MacKenzie *et al.*, 2001; Butts and Cockburn, 2001; Silfverberg, MacKenzie and Korhonen, 2000). Although these proposed alternatives ease the use of such an interface, they do not solve the problem of shrinking user interfaces driven by consumer demand, nor do they address the issue of truly mobile usage, that is, hands-busy, eyes-busy situations.

The investigation explores the viability of speech recognition as an alternative method of text entry. The intention is not to replace the traditional keypress mode of interaction altogether, but instead to add functionality to the existing user interface, thus addressing the limitations of keypress input. This is motivated by previous research suggesting that user performance and satisfaction could be enhanced by the addition of a separate input channel to provide multimodal functionality (Cohen, 1992). Specifically, it is posited that in the same way as a speech user interface (SUI) and a graphical user interface (GUI) can provide a complementary multimodal interface, there exists a similar relationship between keypress and speech recognition at the text-message interface.

1.3.1 Participants

We decided to recruit 36 undergraduate participants who were all regular mobile phone users. In order to verify this, we asked our participants how long they had been using a mobile phone and how many text messages they sent per month. The users had an average of 3.2 years of mobile phone use and, from the information on their most recent phone bills, they sent on average 72 messages a month.

1.3.2 Design

In order to test whether or not speech recognition would be a useful addition to the mobile phone interface, a number of mock-up interfaces were created which enabled interaction using different combinations of speech and keypress. Speech (S) could be used at two different points in entering text messages: it could be used to actually enter the words of the message and it could be used to allow the user to navigate through the mobile phone menus both to start entering the message and to send it off. For entering the message itself, the two most common ways are predictive text (P) and multi-tap entry (M). Navigating round a phone menu is usually done via the keys on the phone (K). Thus, there were three ways of entering the message, S, P and M, and two ways of doing the navigation, S and K. The more formal way of saying this is that there were two factors (or independent variables) in the design: one for message entry with three levels and one for navigation with two levels; giving rise to six experimental conditions. These six conditions are summarised in Table 1.3, where each condition is describing the mode of navigating (to select message entry), then the mode of message entry, and finally the mode of navigating (to send the message).

https://doi.org/10.1017/CBO9780511814570.002 Published online by Cambridge University Press

Table 1.3 *The design of the experiment showing two factors and six conditions*

		Message entry		
		Speech	Predictive text	Multi-tap
Navigation	Speech	SSS	SPS	SMS
	Keypress	KSK	KPK	KMK

Table 1.4 *Predicted time (in seconds) for each task condition based on the GOMS model*

		Message entry		
		Speech	Predictive text	Multi-tap
Navigation	Speech	10.09	16.89	26.02
	Keypress	9.21	16.01	25.14

Before conducting an experiment a GOMS model (John and Kieras, 1996b) was used to create predictions for the time it would take to complete a task of navigating through the menu, composing a message and then using the menu to send the message, using each of the six different methods (see Chapter 4 for information about GOMS models). These are shown in Table 1.4.

Just looking at the predicted times made us hopeful that we were right in thinking that using speech would improve the interaction. Although for navigation speech was always slightly slower than keypress, for message entry the speech times were quicker than both the predictive text entry and the multi-tap entry.

In order to determine whether our predictions were correct it was important to carry out a user study. We decided to use a within-subject design (so that all participants completed tasks on all interfaces).

We have already described the main dependent variable for this study – the task completion times. However, time alone is not the only thing that can tell us something interesting about whether or not one interaction method is better than another. If we only looked at task completion times we might find that people were very quick using one particular method but not realise that this might be because they were rushing and thus making many mistakes (this is known as a speed/accuracy trade-off). This meant it was important for us to look at number of errors too. The NASA-TLX workload questionnaire enables us to look at how difficult people find it to use each method. There would be little point in proposing the introduction of a particular interaction method, no matter how quick it might be, particularly in the context of mobile phones where interaction often occurs on the move, if people found it more difficult to use than other methods. Having three dependent variables (times, errors, workload) enabled us to be more confident that we would find out whether or not speech interaction would really be of benefit to people using mobile phones.

1.3.3 Apparatus/Materials

For this experiment we needed to:

- develop the prototype mobile phones which made use of the different interaction methods
- develop the tasks that participants would complete
- identify and obtain any questionnaires we wanted to include: in our case, the NASA-TLX and the questionnaire that asked participants about their age, mobile phone usage, SMS usage, and so on.

The experiment was conducted on a PC, running a mobile phone prototype. The prototype supported both keypress and spoken input, and the output was displayed on a 17in colour monitor. The numerical keypad of the keyboard was relabelled to mirror the standard mobile phone keypad interface. Dragon Naturally Speaking Professional 7 was used for the voice-recognition system. The environment supported speaker-dependent, continuous recognition of natural speech. A head-mounted microphone was provided to allow dictation to the system.

This level of detail may seem over the top. However, it is what would be needed if someone else were to repeat this work and to be sure that they could compare their results with ours. For example, if the voice-recognition software was not specified, another person might find very different results due to the difference in quality of the voice recognition that they used. Thus, it would be impossible to build up a secure knowledge of the issues in this area across different groups of researchers.

The ideal experimental procedure would be to test participants' usage and acceptance of a multimodal mobile phone interface. However, the mobile phone handsets currently available are not yet powerful enough to support speech recognition in the text-message interface. Moreover, mobile phones are now so pervasive that participants in the experiment are likely to be experts at text entry on their own handset model. Hence, a study using a Nokia phone, say, would be biased by the prior experience of any owners of Nokia handsets. These two factors therefore led to the decision to carry out the experiment on a mobile phone prototype, running on a desktop PC. In addition, the lack of a familiar interface would produce more consistent results across the population. Finally, such a prototype can be instrumented to accurately measure and record task completion times.

Six experimental text messages, derived from a logging study, were printed in lower case, 14 pt bold type and mounted individually onto A5 card. We used this method of creating the messages so that we could be reasonably sure that we were including words that were commonly used in SMS messages. Four training text messages were also printed in lower case 14 pt bold type and presented as a group, on plain A4 paper. The messages were formulated to mirror the conversational text-sending habits among undergraduate students, utilising their most frequently used words. Each text message consisted of 8 words, averaging 28 characters in length. All punctuation and capital letters were omitted from the messages.

A one-page pre-experiment questionnaire was constructed to elicit background information on the users and their mobile phone usage. This A4 document consisted of nine short questions. An A4 booklet was also provided to the participants consisting of six NASA-TLX task workload evaluation forms. Each page displayed the six workload factors and an unmarked 20-point, equal-interval scale, with endpoints labelled 'low' and 'high'. The final page of the booklet consisted of one question requiring the user to state their preferred method of interaction and give a brief explanation of their choice.

1.3.4 Procedure

All participants were tested individually, seated at a desk in front of the computer monitor and the keyboard. It was explained that the aim of the investigation was to evaluate user satisfaction and performance with different mobile phone navigation and text-entry techniques. These would consist of keypress navigation, multi-tap and predictive text entry, and spoken navigation and text entry. Each participant was informed that the experimental test procedure would consist of using a mobile phone emulator to create and send six different text messages via six different methods of interaction (KMK, KPK, KSK, SMS, SPS, SSS). It was then explained that the experimental procedure would consist of four stages: pre-experiment questionnaire; training the voice-recognition software; practice session using the prototype; experimental testing procedure.

After the participants had consented to continue with the experiment, they were asked to complete the pre-experiment questionnaire and then underwent the training program required by the Dragon Naturally Speaking system. The head-mounted microphone was positioned approximately one inch from the participant's mouth and they were instructed to read aloud the instructional text from the computer screen. Participants were encouraged to speak naturally and clearly. Training continued until the system had successfully completed the enrolment program – this process took approximately 10 minutes, depending on the system and the participant's ability in reading the instructional text. On completion of the speech training, participants were allowed a short break while the system analysed the speech data and constructed a speech model for the participant.

Following speech training, a brief practice session using the prototype was carried out. This was necessary to accustom each participant to the modified numerical keypad and the on-screen interface of the prototype. The keyboard was placed on the desk in front of the monitor and moved by the participant into a comfortable 'texting' position. This practice session consisted of the six experimental conditions used in the main experiment, but the participant was allowed to choose the order in which conditions were practised, with a set of training messages supplied by the experimenter. Participants were allowed to complete each condition once only.

After the practice session, the main experiment was summarised. Each participant was informed that task completion times were being recorded and was

encouraged to complete each task as quickly and accurately as possible. They were informed that timing would begin when the experimenter pressed the 'Start' button on the emulator and would stop when the final menu command option was executed. Participants were instructed not to correct any spelling mistakes, although it was explained that any mistakes made would be recorded. Each participant was told that they would be required to complete a task workload form on completion of each task. In addition to an oral explanation of the form, a hard copy explanation for each workload factor was provided. It was then explained that the experiment would conclude with one final question. Once the participants understood and were comfortable with the experimental testing procedure, they were encouraged to ask any questions.

The experiment began when the participants told the experimenter they were ready to start. A Latin square was used to determine the order of task completion (thus minimising any order or practice effects) and at the start of each task the proposed method of interaction was fully explained to the participants and they were presented with a new, unseen text message to enter. They were encouraged to familiarise themselves with the text message by reading through it two or three times. Each test began when the experimenter said 'Go' and pressed the 'Start' button on the prototype. Each test was completed when the last menu command option had been executed. Following completion of each condition, participants were required to rate their experience of the interaction method by circling one of the unmarked values on each task workload scale. On completion of all six conditions, participants were asked to answer the final question on the last page of the booklet. In total, the experiment lasted approximately 45 minutes.

The analysis of the data collected is discussed in Chapter 6. However, the data does show that the differences between the experimental conditions were as predicted by the GOMS model. This suggests that speech-keypress combination could be very useful as a new form of multimodal interface for text messaging. It is, of course, not the last word on the matter. Specifically, this was a formal experiment, and a company interested in such an interface might like to see how it would work for someone sitting on a bus or walking down a street before committing to this as the new design for their latest mobile phone.

1.4 Critique

A well-designed and executed controlled experiment, or series of experiments, can give confidence in the findings. Controlled experiments are well suited to studying details of cognition or interactive behaviour, such as those presented in the preceding example. Other examples include the work of Cockburn, Gutwin and Alexander (2006), who ran a series of controlled experiments on ways of navigating documents, and of Brewster, McGookin and Miller (2006), who studied the use of smell for searching large photographic collections.

The strength of controlled studies can also be their weakness: the causes of success or failure of new interactive systems are commonly to be found in the broader context of activity rather than the details. Controlled experiments are poorly suited to analysing these situations because it is not possible to isolate and control the variables that are pertinent to the interactive behaviour, and it is difficult to design experiments to eliminate all confounds. It is also possible that the experiment measures something other than that which the experimenter believes it is measuring, so that data can be misinterpreted.

Despite the fact that controlled experiments are common within HCI research, we are not aware of any thorough accounts of how to design and conduct such experiments written from an HCI perspective. Nevertheless, there are many text books on the subject of designing and reporting on experiments in the behavioural sciences – for example, Cochran and Cox (1992), Field and Hole (2002) and Harris (2002) all give good overviews of the subject, though with different emphases on the details.