

Using Field Experiments to Evaluate Mobile Guides

Joy Goodman, Stephen Brewster and Philip Gray
Glasgow Interactive Systems Group, Department of Computing Science
University of Glasgow, Glasgow, UK
+44 (0)141 330 3541

{joy, stephen, pdg}@dcs.gla.ac.uk
<http://www.dcs.gla.ac.uk/utopia>

ABSTRACT

This paper discusses the experimental evaluation of location-based services, such as mobile guides, in the field. It identifies advantages and disadvantages of using such field experiments over and against other evaluation methods. We then describe a specific method of running field experiments that we have found to be useful. The use of various evaluation measures, tools and scales is then discussed, based on our experiences with them. These include timings, errors, perceived workload, distance travelled and percentage preferred walking speed (PPWS).

Keywords

Evaluation Methods, Field Experiments, Mobile Devices

1. INTRODUCTION

Mobile and handheld devices, such as mobile telephones and PDAs (personal digital assistants) are becoming increasingly widespread and there is much competition among manufacturers to create new capabilities, applications and interfaces in order to increase and maintain their market share. It is essential that these new services are not only useful but useable if they are to be of real use to real people. This is particularly important for older and disabled people who often find such technology more difficult to use in the first place. Because of the ageing population, the increasing numbers in these user groups make usability an essential concern.

However, usability testing for mobile and handheld devices is not as well understood as for desktop applications. This is particularly true when location-based services, such as mobile guides, are involved. The dependence of such services on the context of use creates extra challenges for effective usability testing. In particular, lab experiments have a key place in usability testing as a whole, but they are of limited use for location-based services because of the difficulties in convincingly simulating the context of use in a laboratory. This paper suggests that field experiments form a useful alternative. Section 2 describes them in more detail and compares them to a range of other evaluation methods, while Section 3 describes a particular way in which they can be carried out, based on practical experience. In Section 4 we examine various measures, measuring tools and scales which could prove helpful in field experiments and discuss the circumstances in which they can be usefully employed.

2. FIELD EXPERIMENTS

Field experiments are quantitative experimental evaluations that are carried out in the field, drawing from aspects of both qualitative field studies and lab experiments. Although their use is

seldom reported in the HCI literature (as Kjeldskov and Graham found in their review in [6]), they have many advantages and can prove to be an important part of the repertoire of evaluation techniques, as some recent researchers have found (e.g., [10]).

To start with, field experiments have advantages over expert evaluations (such as heuristic evaluations and checklists) because they involve real users. This is particularly important when studying novel, variable and less understood situations, such as those involved in mobile devices.

Secondly, there are various reasons for preferring field situations to laboratory settings. Principally, the difference between use in a laboratory setting, even when mobile conditions are simulated, and use in the real world can be quite startling. This is particularly true for devices such as mobile guides that rely heavily on the surrounding environment. Aspects such as lighting levels, weather, the effects of walking, the appearance of landmarks in real life and the effectiveness of location-sensing systems can have unpredictable effects on the usability and effectiveness of a device. The only way to really see how the device will work in practice is to use it in practice.

In addition, field experiments have various advantages over more qualitative and ethnographic field studies. Although such field studies do have some advantages, it can be hard to use them to obtain an objective evaluation of a device, determine its performance or gain hard evidence comparing one device or method with another. Field experiments offer one way to overcome these disadvantages while still utilizing the advantages of a field setting. They are also quicker than carrying out a full-blown ethnographic field study and can gauge the responses of a greater number of participants.

3. METHOD

One possible reason for the low usage of field experiments is the lack of a clear, carefully worked out method for running such experiments. The examples available are varied, sometimes vague and can be hard to find. It is also often thought that field studies, both field experiments and more qualitative field trials, are much harder and more time-consuming than lab experiments (e.g., [7]). While field experiments can take more time and do include a number of additional challenges (see, e.g., Section 3.2), we have not found them, in general, to be substantially more difficult. The main challenges involved, such as developing prototypes and recruiting participants, apply in both cases.

This section therefore presents one method of running field experiments that we have found to be useful and effective. By laying out the main aspects of the method clearly, we hope to

illustrate that running such experiments need not be a difficult task.

We have used this method and variations on it in two linked studies of a navigation aid [4], illustrated in Figure 1. These involved a total of 56 participants and approximately 8 pilot participants from both older (aged 60 to 78) and younger (aged 19 to 34) age groups.

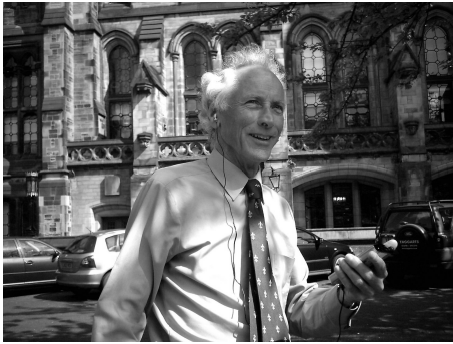


Figure 1. A participant taking part in a field experiment.

3.1 Core aspects of the method

As in any usability experiment, participants are given specific tasks to do. In our example, because we were evaluating a navigation aid, the tasks were to find one's way along specified routes. Examples of alternative tasks include uncovering information about a particular building or tourist attraction, finding the nearest Post Office and determining the location of work colleagues.

To obtain a quantitative evaluation, it is necessary to compare the results from one method with those from another. If a single version of a device is being evaluated, it can be compared with a control, such as a standard paper map of the area or the currently used work practice. If the aim is to determine which of a set of possible alternatives (e.g., alternative interfaces) is best, then these alternatives can be compared against each other.

This comparison can be done using standard between-groups, within-groups and mixed designs. However, special care should be taken when assigning tasks in within-groups and mixed designs to ensure that an individual does not do the same task for multiple methods because location-based tasks are particularly sensitive to repetition. The tasks may also need to be set in different locations, as familiarity with an area may affect performance on repeated tasks. For example, when evaluating our navigation aid, we used two distinct (but similar) routes.

As always, care needs to be taken with counter-balancing the tasks and methods, so that the same method is not always evaluated with the same task nor method one always evaluated before method two.

An experimenter follows the participants as they carry out their tasks, walking a few steps behind them to avoid influencing navigation decisions. This enables the experimenter to note observations on the participants' behaviour and use of the device and to take measurements. This also enables the experimenter to provide help where necessary to prevent excessive distress and harm (e.g., to prevent a participant from getting knocked over) in order to conform to ethical guidelines. On a more mundane

(but very important) level, the experimenter can ensure that the participants do not abscond with valuable equipment.

In addition to the quantitative data collected during the experiment (see Section 4 below), participants are questioned after trying out each method about their experience of that method, using a questionnaire or interview. At the end of the experiment, they are also asked to compare the different methods.

3.2 Controlling Variables

One of the biggest challenges in carrying out field experiments is the difficulty of controlling possibly confounding variables. This is difficult enough in a lab setting but in the field it forms an even greater challenge because of variables such as light and noise levels, weather conditions and traffic that are outside the experimenter's control.

It is possible to attempt to keep the levels of such variables consistent, for example, by carefully scheduling and re-scheduling experiments and removing data from runs that did not stay within appropriate control levels. However, this can consume much time and resources and may prove irritating to participants as well as experimenters. In addition, we believe that removing all variation would produce unrealistic results which may not mean much for real-world usage. Therefore, a more effective method is to let these variables vary across conditions, as they would in real world use.

These techniques will lead to robust results, but may "blur" the data, masking some of the smaller effects. For devices such as mobile guides, this is not generally a problem as such smaller effects are unlikely to have an effect on real-world usage. However, it can sometimes be useful to identify such effects so that they can be increased through careful design. More seriously, although varying the variables across the conditions reduces the risk of introducing bogus differences between conditions, this danger still exists. It may therefore be useful to complement field experiments with carefully chosen laboratory studies.

4. SOME EVALUATION MEASURES AND TOOLS

This section describes some of the most relevant measures that can be used when evaluating mobile guides and other location-sensitive applications. It also discusses their relevance, use, reliability and measurement. A summary of these measures is given in Table 1.

It is important to take a range of different measures (both quantitative and qualitative) to find where usability issues may occur. Participants often make tradeoffs which can only be discovered from a range of measures. For example, two interfaces may result in the same error rate, but one may cause participants to walk far slower. In this case, users put their resources into keeping errors down, but at the expense of walking. Without taking a range of measures such problems would not be uncovered.

4.1 Timings

The time taken to complete a task or set of tasks forms one of the most common usability measures. It is often used as an indicator of performance with a device or interface because it is assumed that the faster a task is performed, the better the interface is supporting it.

Timings are easy to measure. They can be noted by the experimenter using a watch or stopwatch or can be determined by the device itself by adding a few lines of code to log the timing of interesting events, or start and end times.

4.2 Errors

The number or rate of errors is also commonly used in usability tests to indicate performance and success in using an application. If the causes of these errors are noted then they can also help to identify particular aspects of a design that are causing difficulty.

The definition of an error and hence the way it is measured vary widely depending on the application. Errors can range from failure to correctly complete the task to non-fatal, recoverable errors, such as pressing a wrong button or turning the wrong way at a junction but quickly turning back. Such errors can be noted by the experimenter or identified from interaction logs or measures of the route taken, described in Section 4.4.

Some errors are fatal, in that they prevent successful completion of the task. It may sometimes be desirable for all participants to complete the task (e.g., to enable full timing comparisons to be made) and therefore to provide help to overcome these errors. In such cases, it is extremely important to note when and why the error occurred and the help that was necessary to overcome it.

4.3 Perceived workload

Hart and Staveland [5] define workload as the effort invested by the human operator into task performance. Cognitive resources are required for a task and there is a finite amount of these. As a task becomes more difficult, the same level of performance can be achieved but only by the investment of more resources. Workload is important in a mobile setting as users must monitor their surroundings and navigate, therefore fewer resources can be devoted to an interface. An interface that reduces workload is likely to be more successful in a mobile setting.

Workload is commonly measured using the TLX (task load analysis) scales developed by NASA [5]. They can be used without modification but we have found that some of the terms

used to describe aspects of workload are not relevant to mobile devices, such as mobile guides. It may therefore be worthwhile to modify the wording to relate more immediately to the device and tasks used, as in [3].

4.4 Distance travelled and route taken

The distance travelled and route taken by participants are measures with specific significance for field trials of mobile devices. They can help to identify location-based errors and to determine particular points at which difficulty arose. These measures are closely related to each other and can be measured in similar ways. A few of these ways are described below.

4.4.1 Pedometers

A pedometer can be used to measure the distance travelled (although not the route taken). It is small device attached to the waistband at the hip, which counts the number of steps taken.

In order to measure the distance walked, it needs to be calibrated at the beginning of an experiment by asking the participant to walk a short measured distance. We have found that this calibration phase can be inaccurate. It is therefore best if the participant can walk the distance several times and an average step length entered manually.

We also observed some difficulties using the pedometer in winter and wet weather. The extra layers of clothes worn can make attaching and reading the pedometer awkward. In addition, the body shapes and clothing of some older adults mean that attaching the pedometer to the waistband places it in the wrong position to register steps. Nevertheless, in the majority of cases, the pedometer worked without difficulty and participants had no objections to its use.

A pedometer measures the distance walked, which can help to identify if detours took place, but is no use in identifying the location of those detours. Together with the difficulties in using a pedometer, we therefore believe other methods to be of more use, as described below.

Table 1. A comparison of evaluation measures and tools.

Measure	What does it tell us?	How is it measured?
Timings	Performance	By the experimenter via a watch or stopwatch. Automatically by the device.
Errors	Performance. Particular sticking points in a task or design	By success in completing the tasks or answering questions correctly. Through experimenter observation, examining the route walked and the buttons pressed.
Perceived workload	Effort invested. User satisfaction	Through NASA TLX scales and other questionnaires and interviews.
Distance travelled and route taken	Depending on the application, these can be used to pinpoint errors and to indicate performance	Using a pedometer, GPS or other location-sensing system. By experimenter observation.
Percentage preferred walking speed (PPWS)	Performance	By dividing distance travelled by time to obtain walking speed, which is then compared with normal walking speed.
Comfort	User satisfaction. Device acceptability	Using the Comfort Rating Scale and other questionnaires and interviews.
User comments and preferences	User satisfaction and preferences. Particular sticking points in a task or design	Through questionnaires, interviews and think-alouds. Using open-ended questions and forced choice.
Experimenter observations	Many different aspects, depending on the experimenter and on the observations	Through observation and note-taking by the experimenter.

4.4.2 GPS

GPS and other location-sensing equipment can be used to track the route taken and hence the distance covered automatically (e.g., [1]). However, inaccuracies in GPS in built-up areas, especially at high latitudes, reduce their effectiveness. Difficulty getting a location fix can indicate detours in the route where none took place, as indicated in [1].

4.4.3 Experimenter observation

Because of these difficulties, we recommend experimenter observation as the most reliable method of noting the route taken. However, care should be taken to avoid giving the experimenter too many tasks to do as it can be difficult for him or her to manage multiple pieces of paper and equipment while on the move. It may prove helpful to provide the experimenter with a map of the area on which the route can be drawn and which can be annotated with observations at appropriate points. Alternatively, multiple observers can be used in order to share the observational load.

4.5 Percentage preferred walking speed

Percentage preferred walking speed (PPWS) [9] measures the extent to which the use of a device disrupts normal walking: the further users walk below their normal walking speed, the more negative the effect of the device. As Petrie *et al.* [9] say "... all pedestrians have a walking speed which they prefer ... The ability of any mobility aid to allow the pedestrian to walk at this preferred walking speed (PWS) is therefore argued to be a measure of its effectiveness".

To use PPWS, participants' normal preferred walking speed must be measured before the experiment starts. This can be done in a similar way to the calibration of a pedometer (described above). This can then be compared to the experimental walking speed after recording distance walked and time taken.

4.6 Comfort

It is important to know if any system developed will be acceptable to users in practice. One important aspect of this is comfort because a device needs to be comfortable and people need to be happy to wear it if it is to be accepted. Knight *et al.* [8] have developed the Comfort Rating Scale (CRS) which assesses various aspects of the comfort of a device. Using a range of rating scales similar to the NASA TLX, CRS allows users to rate factors such as emotion, harm and anxiety.

4.7 Other possible measures and tools

Comments from participants and information about user preferences can be gained through questionnaires and interviews, as well as by noting down comments made while using the device. Particularly useful may be use of the think-aloud method in which participants speak their thoughts and observations aloud, although care needs to be taken that this does not interfere with the use of the device. Forced choices where participants must choose between different versions of the device can help to throw light on their preferences.

It is also useful to have the experimenter note down observations during the experiment on particular uses of the device and routes taken, as mentioned above. Taking photographs and vid-

eos may also come in handy, although, as already mentioned, care should be taken to avoid overburdening the experimenter or giving him or her too much to carry.

5. CONCLUSIONS

We conclude that field experiments need not be as difficult as commonly thought and have a big payoff in terms of measureable results. This paper has presented a set of techniques and suggestions to make running such experiments easier.

6. ACKNOWLEDGMENTS

This work was funded by SHEFC through the UTOPIA project (grant number: HR01002), investigating the design of usable technology for older people [2].

7. REFERENCES

- [1] Bornträger, C. and Cheverst, K. Social and Technical Pitfalls Designing a Tourist Guide System. In *Proceedings of Mobile Guides*, Workshop at *Mobile HCI 2003*, 1-5.
- [2] Eisma, R., Dickinson, A., Goodman, J., Syme, A., Tiwari, L. and Newell, A. F. Early User Involvement in the Development of Information Technology-Related Products for Older People. *Universal Access in the Information Society*, 3, 2, 2004.
- [3] Fairclough, S. H. *Adapting the TLX to assess driver mental workload*. DRIVE I project V1017 BERTIE Report No. 71, HUSAT Memo Number 1156A, Loughborough, UK: HUSAT Research Institute, 1991.
- [4] Goodman, J., Gray, P., Khammampad, K. and Brewster, S. Using Landmarks to Support Older People in Navigation. In *Proceedings of Mobile HCI 2004*, Springer-Verlag, LNCS series, 2004.
- [5] Hart, S.G. and Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Hancock, P. and Meshkati, N. (eds.). North Holland B.V., Amsterdam, 1988, 139-183.
- [6] Kjeldskov, J. and Graham, C. A review of Mobile HCI Research Methods. In *Proceedings of Mobile HCI 2003*, Springer-Verlag, LNCS 2795, 2003, 317-335.
- [7] Kjeldskov, J., Skov, M. B., Als, B. S. and Høegh, R. T. Is it Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. In *Proceedings of Mobile HCI 2004*, Springer-Verlag, LNCS series, 2004
- [8] Knight, J. F., Baber, C., Schwirtz, A. and Bristow, H. W. The Comfort Assessment of Wearable Computers. In *Proceedings of the International Symposium of Wearable Computers (IEEE ISWC) 2002*, 65-72.
- [9] Petrie, H., Furner, S. and Strothotte, T. Design Lifecycles and Wearable Computers for Users with Disabilities. In *Proceedings of the first workshop on HCI mobile devices*, 1998.
- [10] Ross, D. A and Blasch, B. B. Department of a Wearable Computer Orientation System. *Personal and Ubiquitous Computing* 6(1), 2002, 49-63.

