Chapter 7. Experimentation

Table of Contents

Most IT research is experimental

Experiments are at the heart of the scientific method, where hypotheses are confirmed or refuted through carefully controlled artificial test situations. One of the differences between system development in industry and in research, is the quest for new knowledge and for evidence to support such claims. Therefore, with the exception of formal or theoretical studies, IT research requires conducting experiments to backup claims about new computing techniques, systems, models, architectures, etc. A major difference between a software/hardware development project and a research project is that the latter involves a hypothesis (new knowledge whose truth we want to evaluate), the design and execution of experiments to test this hypothesis, and the careful and rigorous analysis of data from these experiments to see whether they support the hypothesis. In this way new knowledge is gained by building systems or models, and observing and analysing their usage. A computer science research publication that does not include experiment results contributes ideas only, without any evaluation of these ideas. Experiments can be exploratory, or can be structured so as to answer specific question(s) or to test hypotheses. The kind of experiment most common in IT is an experiment which sets out to validate a new model/system/architecture/approach by implementing a prototype that shows it is feasible and practical, or that it behaves as was predicted. Another type of experiment in computing is that which compares a new model/system/architecture/approach proposed by the researcher with current practice; while others are designed to establish optimum parameter values or resource needs. Experiments should also be conducted when a critical decision is made, to avoid proceeding further without confirmation that a good choice was made. This is crucial particularly when working in the context of safety critical systems.

Any research should state a hypothesis at the outset and end with a conclusion as to whether the hypothesis was supported or refuted. In some cases where the research focuses on building a system, this hypothesis may be stated along the lines of "System P is good for task X" or "System P is better than rival systems Q and R for task X", or similar sentences with the word system replaced by model, technique or theory [Bundy97]. The context in which P is better also needs to be stated. It is seldom possible to test all types of task X, so the range of such tasks must be classified so that a representative sample can be used in the experiments. At the same time the meaning of "better" needs to be precisely defined (for example P may be more widely applicable, easier to use, faster, consuming less resources, or the like), and in a way which is explicit and measurable. A useful check that a hypothesis is acceptable is to imagine a scenario in which the hypothesis is refuted.

Planning the experiment

When an experiment is designed, researchers should document their goal/hypothesis and experimental procedure, and then scrutinize this from an outsiders viewpoint to see what criticisms could be leveled at their proposal. As ever, it is preferable to also seek such feedback from colleagues and friends. The experiment can then be modified to take these criticisms into account; or where this is not possible, the goal/hypothesis can be altered or limited accordingly.

A pro forma plan is a brief indication of the requirement being tested and the finding we hope to reach in an experiment. Its first sentence states the requirement and its second sentence summarizes the anticipated result. A pro forma report is written after the experiment, and also comprises a first sentence giving the requirement and a second giving the findings of the experiment. Here is an example:

Pro forma plan: It is required of the ticket machine that first-time users should be able to operate it successfully without prior training. We intend to carry out laboratory tests of the machine to confirm that at least 95 per cent of users can complete normal purchases without difficulty.

Pro forma record: It is required of the ticket machine that first-time users should be able to operate it successfully without prior training. We have carried out laboratory tests of the machine in which over 97 per cent of users completed normal purchases in less than 60 seconds. [Newman and Lamming]

References: W.M. Newman & M.G. Lamming. Interactive System Design. ISBN 0-201-63162-8. Addison-Wesley: 1995.

Criteria for successful experiments

The purpose of a laboratory experiment is to obtain precise measurements for a system (measures of performance, reliability, usability, or whatever criteria are of interest). To achieve this requires tight control over experimental conditions so that extraneous influences are factored out.

For a prototyping experiment to demonstrate an idea is valid requires some measure by which feasibility can be established (e.g. response time). To compare two systems or approaches to determine which is better requires suitable performance indicator/s (e.g. resource utilization, time, or error rate). To test a hypothesis, at least two measurable variables are needed, in order to study the relationship between them. Experimental research thus requires finding performance criteria that are convincing, verifiable, easy to measure, and measuring the right thing.

The aim of experimental techniques is to minimize the effect of bias in research and facilitate independent verification of observations. If experiments are repeatable and use well-known methods, then others can more readily accept research results. An experiment succeeds if it measures only the effect that it sets out to measure, without interference from other influences, and if its results can be generalized to other cases (or to other cases satisfying assumptions underlying the experiment).

An experiment should be designed so as to learn something that is generally applicable, and not just true for the particular instances involved in the research. To generalize from experiments requires that an adequate number of experiments is performed, which in turn depends on the number of extraneous factors that are possibly affecting results. If a researcher wants to demonstrate that an idea is feasible and implementable, only one experiment is necessary. To show that one system/approach is better than another, requires running two experiments - one with each alternative - and comparing results. It is best to use well-known benchmarks for such experiments, where these exist. Showing a relationship between variables or testing a hypothesis will require more experiments, so that any influence from extraneous factors can be counteracted.

Example

As an example, consider experiments conducted to evaluate MYCIN, the medical diagnosis expert system. Initially, experts were given MYCINs diagnoses and recommended treatments and asked to rate them. Later experiments used improved measures: experts were given diagnoses and treatments, some of which had been done by other experts and some of which were done by MYCIN, and asked to rate these without knowing which were which. This removed bias, and also showed up that expert opinions were not always reliable and needed to be averaged out.

Hypothesis testing

The hypothesis being tested in an experiment is one that predicts a relationship between two events or characteristics, called variables. The variable that is systematically manipulated by the researcher

in an experiment is called the independent variable because the conditions to test this variable are set up independently beforehand. The variable being measured to see how it is affected is called the dependent variable because its value depends on the setting of the other variable.

If there is a control and an experimental test, and only one variable is different for the two, then it can be argued that differences between the two tests are a result of the differences in that variable. But this means that extraneous variables need to be kept constant in an experiment, with only the experimental variable changing so that its effect (alone) can be determined. Experimental research involves varying one or more independent variables to assess their effect, controlling extraneous variables to prevent their influencing results, and creating different experimental conditions to study.

A graph or bar chart helps in comparing two sets of results easily. Statistical tests like the student t-test can be used to test if differences are significant, and hence if the hypothesis has been confirmed (or not).

Experiments with human subjects

If an IT experiment is concerned with system behaviour or performance, it is easier to control the effects of extraneous factors; if it involves humans (e.g. usability studies), then a number of participants will be needed to balance out the effect of the many extraneous characteristics of the individuals themselves.

If two systems compared in an experiment turn out to be very similar in performance (or error rate, or whatever is being measured), it can be due to a poor selection of test tasks or to extraneous differences between control and experimental groups.

Finding a representative sample

For an experiment to give a good indication of what will occur in the real world, we require a realistic group of users tackling realistic tasks. The small fraction of future users and tasks that are involved in an experiment is called the sample, and it is important to obtain a representative sample. If we have a sufficiently large sample that is representative of the target population, then the mean performance measured in the experiment should be very close to the real mean (for the population as a whole).

Dividing subjects into groups

To limit the effect of extraneous characteristics of human subjects, participants should be divided into a control and an experimental group, without being told the difference between the two groups. (Ideally, researchers observing them should not know which group they are looking at either. Such double-blind experiments are seldom possible in computing however). One approach is to randomly divide participants into two groups, on the assumption that the groups will thus be equivalent/ comparable. It is best to test this assumption by interviewing, questioning or testing the two groups before the experiment starts, and then forming two balanced groups accordingly. The technique known as matched-participants creates pairs of subjects having matching characteristics; the group is then divided into two equivalent halves by splitting each pair. This will also enable you to do before-vs-after comparisons on each group, by repeating the test, interview or questionnaire afterwards.

A single group repeating a task

A problem with dividing subjects into two groups is that, unless there are enough subjects available, the groups can be very small and so the effect of any individual differences can be pronounced. As an alternative, all participants can do the control case first and then do the experiment afterwards (or vice versa). This removes the possibility of extraneous factors affecting different results shown by the two groups (since it is the same group of people every time), but it is only applicable if subjects do not learn or change during the first experiment, thus affecting their performance in the second one. This approach is certainly best when systems and not humans are involved. Otherwise, the order effect can

be reduced by halving the group of subjects, so that one half does A and then B, while the other does B and then A (called counterbalancing).