

6

Experimental Design

The basic experimental design in IIR evaluation examines the relationship between two or more systems or interfaces (independent variable) on some set of outcome measures (dependent variables). IIR evaluations can include other independent variables as well such as task-type, and quasi-independent variables such as sex and search experience. One important part of experimental design which will be discussed in detail is rotation and counterbalancing. Tague-Sutcliffe [261] was one of the first to write formally about this in IIR. This allows one to control aspects of the study that might otherwise introduce experimental confounds. This section also presents other issues related to experimental design including study mode, protocols, tutorials, timing and fatigue, and pilot testing.

6.1 Traditional Designs and the IIR Design

Traditional designs can be discussed in terms of pre-experimental designs and experimental designs. These are standard designs that are discussed and presented in a number of research methods textbooks (e.g., [13]). They are not a creation of IIR and do not always fit perfectly with IIR study situations, but they do provide different

Group	Time_1		Time_2
1	O	E	O
2	O	C	O
3		E	O
4		C	O

Fig. 6.1 Solomon four-group experimental design.

ways of thinking about study design and measurement. The distinction between pre-experimental and experimental designs rests on the absence of a control group and baseline measurement. Figure 6.1 presents a well-known experimental design, the Solomon four-group design [47]. The different groups in this design can be used to illustrate other types of research design, including pre-experimental designs. A pre-experimental design with no control group or baseline measurement is represented by Group 3. In this group, an experimental stimulus (*E*) (e.g., a system) is introduced and then an observation or measurement (*O*) is taken of some outcome measure (e.g., performance). One of the most common types of studies in IIR that follow the design depicted by Group 3 is the single system usability test. There is no comparison or control system. Instead, subjects use one system and some initial feedback is collected regarding its goodness. Note that this type of study does not allow for the testing of hypotheses related to the system because there is only one system being studied. No comparison is possible, except with pre-determined population parameters, which are unlikely to exist. It is important that one looks closely at one system studies before deciding they are usability studies and not experiments. Many experiments only involve a single system, but some other variable of interest is manipulated and of interest. It is possible for IIR evaluations to have independent variables that are not tied directly to a system. The system may just be used as an instrument to facilitate information search (e.g., [168]).

The other attribute that makes this (Group 3 only) a pre-experimental design is that a baseline measure of the outcome variable has not been taken. In traditional experimental models, baseline measures of the outcome variables of interest are elicited before the stimulus is introduced. This is depicted by Group 1 in Figure 6.1. For instance, if one were evaluating a new drug designed to help people lose

weight and the outcome measure was a person's weight, one would need to obtain a baseline weight for each subject to know if the drug was associated with a decrease in weight — without this measure it would not be possible to determine this. In the context of IIR, one general goal of many evaluations is to determine if a particular system helps subjects find relevant documents. Attempting to elicit a baseline measure before the system (stimulus) is introduced does not make much sense and would probably not be possible. We can also imagine that the goal of an IIR system is to help subjects learn something about their information problems. To evaluate this, we would really need to measure subjects' knowledge of their information problems before and after they used the system.

Baselines are used in IIR evaluations, but in a way that differs slightly from the classic experimental model. In the context of IIR evaluations, baselines are often introduced as an alternative to the experimental system. Instead of taking a baseline measure before a user interacts with a stimulus, the baseline is more often represented by one level of the stimulus variable. For example, if the stimulus variable is an IIR system, it might have two levels: experimental and baseline. Thus, baselines in IIR evaluations are more similar to control groups (*C*) in Figure 6.1. In the *traditional experimental model*, the stimulus variable is usually either present or absent and a control group is used along with pre-treatment measurement (Figure 6.1, Groups 1 and 2). In Figure 6.1, the *classic IIR design* is represented by Groups 3 and 4. This model ostensibly functions as the archetypical IIR evaluation design.

A baseline (or control in the traditional model) is generally defined as the status quo, which raises some interesting questions with respect to IIR evaluations. Specifically, if IIR systems are under study and baselines represent subjects' normal experiences, then in most cases this would be a commercial search engine. However, it is not possible or valid to compare an experimental IIR system to a commercial search engine.¹ For instance, a researcher may be using a closed collection of newspaper articles; if a commercial search appliance were used to access

¹This may be possible if you work for a commercial search engine company.

this collection as a baseline, it might not work optimally because of characteristics of the corpus and search algorithm. Thus, such an evaluation would not be comparing similar situations. Of course, whether a commercial search engine is a valid baseline depends greatly on the purposes of the study and the system. Even though it may not be possible or desirable to use a commercial search engine as a baseline, it is important to recognize subjects' previous search experiences and search norms will impact their interactions with, and expectations of, any experimental IIR system.

Developing a valid baseline in IIR evaluations often involves identifying and blending the status quo and the experimental system. For instance, if a researcher developed a new technique for displaying search results, then a baseline method of doing this could be modeled after methods used by commercial search engines. If the experiment was done using a proprietary system or well-established system, then the baseline could be the retrieval method currently used by that system (given that one was testing the workings of the system). Things get a bit more difficult when the experimental system or interface is something that subjects have never seen. Researchers often develop experimental IIR systems from scratch using languages such as Java. There is a good chance that the interface will look very different from a Web-based system to which subjects are accustomed. In this case, if one were comparing a new search interface feature, it would not be reasonable to compare this to a standard Web search engine since the number of differences between these two systems would be great. If differences were found, it would be difficult to relate them to the specific search interface feature of interest and to rule out the possibility that these differences were not caused by some other feature or aspect of the system.

As mentioned earlier, the design depicted in Figure 6.1 is called the *Solomon Four-Group Design* [47]. It was developed to address several major threats to the internal validity of experiments. These will not be discussed here, but suffice to say the four groups allow the researcher to control a number of threats to validity. The Solomon Four-Group Design is quite nice, but requires large numbers of subjects, since the groups are independent. Many researchers in other disciplines use the classic experimental design (Groups 1 and 2 only), while others

(IIR included) use a modified design based on Groups 3 and 4. This design is called a *Posttest-only Control Group Design* [47]. Campbell and Stanley [47] argue that these are the only two groups needed, if subjects have been randomly assigned to the groups.

All of these designs rest of the assumption that subjects comprising each of the groups are equal across a range of characteristics. Characteristics which might, if not equally distributed across groups, conspire to generate spurious results — results not caused by the stimulus, but by some other characteristic of the group. Random assignment can be used to increase the likelihood that these characteristics are distributed equally across groups. While it is usually not possible to conduct true random sampling in IIR evaluations, it is possible to randomly assign subjects to groups (or conditions).

6.2 Factorial Designs

Currently, the more common way for researchers to discuss experimental design in IIR is as *factorial designs*. This is particularly useful when studying the impact of more than one stimulus or variable. The models presented above assume a binary stimulus (experimental and control), but the researcher might also be interested in studying the impact of a number of factors² on one or more outcome variables. Factorial designs accommodate this. In the preceding example there was one factor, system type, which had two levels, experimental and baseline. If the researcher believed that there might also be differences in the outcome variable based on the sex of subjects, then sex would be an additional factor, with two levels, male and female. This is tightly coupled with the previous levels of measurement discussion; the factors in a factorial design should be discrete. The levels represent distinct categories rather than ratio level values.

There is a specific notation and language for describing factorial designs. If the relationship between the two factors mentioned above were examined (system type and sex) in relation to an outcome measure such as performance, then the experiment is described as a 2×2 factorial

² Used as a synonym for independent variable.

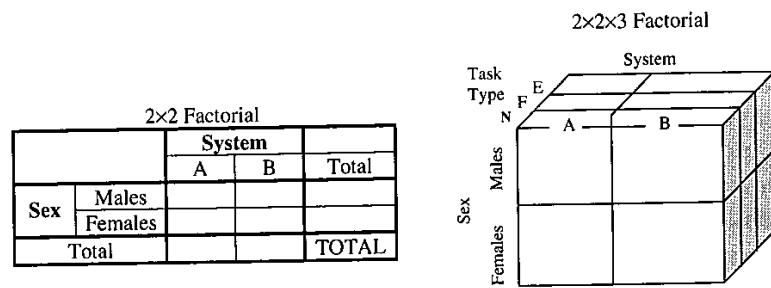


Fig. 6.2 Example factorial designs. The first example is a 2×2 design with one independent variable, system and one quasi-independent variable, sex. The second example is a $2 \times 2 \times 3$ design with one additional independent variable, task-type, which has three levels: navigational (n), fact-finding (f), and exploratory (e).

design. Both the number of digits and their magnitudes are meaningful. The number of digits describe the number of factors (system type and sex) and the magnitude of each number describes the number of levels of each factor (experimental, baseline; male, female). If another factor called task-type were added which had three levels (exploratory, fact-finding, and navigational), the experiment would be described as a $2 \times 2 \times 3$ factorial. These two designs are illustrated in Figure 6.2. Such illustrations aid in describing a study and allow researchers to understand and communicate the different types of comparisons that are available. The different combinations of levels generate different conditions (in the 2×2 there are four conditions and in the $2 \times 2 \times 3$ there are 12 conditions). Each condition will have some value on one or more outcome variables. Comparisons can be made using cell, column and row values. Each factor adds another dimension to the representation and studies with four or more factors do not lend themselves as easily to this type of representation and are not conducted that often anyway because they require large numbers of subjects and it is difficult to interpret results.

6.3 Between- and Within-Subjects Designs

Studies can also be characterized with respect to whether the independent variables are *between* or *within* subjects. This is an important

distinction which should be made in all reports. Between-subjects means that subjects experience only one level of the variable, while within-subjects means that subjects experience all levels of the variable. Studies can be mixed along this characterization: some variables can be between-subjects, while others can be within-subjects. For instance, system type might be a between-subjects variable, while task-type might be a within-subjects variable. This means that each subject would only use one system, but would have to complete all three task-types. Some variables are necessarily between- or within-subjects. For instance, values on the sex variable are completely beyond the control of the experimenter and reside outside of the study. In the classic IIR evaluation study, system type (or interface type) is typically within-subjects to facilitate comparison of the two systems. Otherwise, it is not possible to ask subjects to compare the systems since they would have only used one of them. In other cases, it might be desirable to make a variable between-subjects to avoid exposing subjects to all conditions of the experiment, which might lead to contamination.

6.4 Rotation and Counterbalancing

Rotation and counterbalancing are cornerstones of most experiments and evaluations and are often associated with systems and tasks in IIR evaluations [260]. The primary purpose of rotation and counterbalancing is to control for order effects and to increase the chance that results can be attributed to the experimental treatments and conditions. Although *treatment* typically refers to the things that a researcher tests or manipulates (e.g., interface), it also refers to the tasks and topics which subjects execute when engaging in IIR regardless of whether these items are variables of interest. In most IIR evaluations that involve searching, search tasks are necessary in order for subjects to exercise systems. Even though they may not be treated by the researcher as independent variables, they do function as variables and therefore must be controlled. This is typically achieved through rotation.

Two types of designs can be used to systematically rotate variables, the *Latin square design* and the *Graeco-Latin square design* (Graeco is

also spelled Greco). The Latin square design accommodates a single variable, while the Graeco-Latin square design can accommodate multiple variables — it is essentially a combination of two or more Latin squares. To illustrate the different designs, let us assume that we are testing three interfaces using six different search topics and that each user will complete two topics per interface. The task will be held constant and will be a document finding task.

6.4.1 A Basic Design

First, let us look at a basic design with no rotation (Figure 6.3), where rows represent subjects and columns represent interfaces. Topics are represented in the cells of the table. There are two major problems with this design — the first is related to topic order and the second is related to interface order. A Latin square can be used to control for one of these things, but not both. A Graeco-Latin square is needed to control both.

The major experimental confounds that are introduced by this design are caused by order effects. Specifically, learning and fatigue can produce results that are attributable to the experimental design rather than the treatments. As subjects complete each consecutive topic, they learn more about the experimental situation and the experimental system (assuming that the systems are similar). With each topic encountered, subjects potentially transfer what they learn by completing one topic to their interactions with the next topic, which might result in better performance on topics that are presented last, as opposed to first. Since the order of interfaces is fixed, it is also likely that the subject's

Subjects	Interface type		
	Interface 1	Interface 2	Interface 3
S1	1, 2	3, 4	5, 6
S2	1, 2	3, 4	5, 6
S3	1, 2	3, 4	5, 6
S4	1, 2	3, 4	5, 6
S5	1, 2	3, 4	5, 6
S6	1, 2	3, 4	5, 6

Fig. 6.3 A basic design with no rotation. Numbers in cells represent different topics.

earlier experiences will impact later experiences. If, on average, subjects perform best with the last interface, it may be a function of a learning effect, rather than the goodness of the last interface. Another problem with a fixed order for topics and interfaces is related to the potential interactions among the topics and the system. Some topics may be easier than others and some systems may do better with some topics than others. If, for example, it was found that Interface 2 was the best, it may be because Topics 3 and 4 were easier than the other topics. In this case, while the researcher may attribute differences to the interfaces, the differences are really caused by the topics.

Fatigue can also impact the results. As subjects engage in more and more searches, they are likely to become fatigued, especially in experiments that last over one hour. At the beginning of the study, subjects may be more motivated and attentive than at the end of the study. When subjects become fatigued they may move quickly through the experiment just to finish. They may become cognitively exhausted and just not be able to perform as well as they did at the start of the study. If, for example, it was found that Interface 1 was the best, it may be because subjects were more energized and worked harder in the beginning of the study than at the end.

6.4.2 A Latin Square Design

To improve the design in Figure 6.3, a *Latin square* can be used to control for the effects of one of the variables — either topic or interface. Latin square designs are used to rotate and control for a *single* variable and in Figure 6.4 this variable is topic. The items in the cells of a Latin square are distinct and should appear an equal number of times in each row and each column. This can be accomplished fairly easily: for each row, topics are shifted among the columns in a systematic way. Since there are six topics, we need six rows (or six subjects) to get through one topic rotation. Note that each user completes all topics. It is also important to note that these designs do not eliminate learning or fatigue, but distribute their impact equally across all treatments and conditions.

While the rotation in Figure 6.4 is balanced with respect to topics, it is problematic for other reasons. There are two important things that

Subjects	Interface type		
	Interface 1	Interface 2	Interface 3
S1	1, 2	3, 4	5, 6
S2	2, 3	4, 5	6, 1
S3	3, 4	5, 6	1, 2
S4	4, 5	6, 1	2, 3
S5	5, 6	1, 2	3, 4
S6	6, 1	2, 3	4, 5

Fig. 6.4 Basic design with Latin square rotation of topics.

Subjects	Interface type		
	Interface 1	Interface 2	Interface 3
S1	2, 4	1, 6	5, 3
S2	3, 5	2, 1	6, 4
S3	4, 6	3, 2	1, 5
S4	5, 1	4, 3	2, 6
S5	6, 2	5, 4	3, 1
S6	1, 3	6, 5	4, 2

Fig. 6.5 A basic design with Latin square rotation of topics and randomization of columns.

this type of rotation does not address. First, it is possible that there may be some interaction among the topics, such that encountering Topic 4 after Topic 3 makes completing Topic 4 easier. Note that for all rows of the table except one, Topic 4 always follows Topic 3. One can see this visually in the design via the diagonal — this indicates that there is still some order preserved in the table (it is easiest to spot this along the '6' diagonal). One way to address this problem is to randomize the order of the columns (excluding the column headings). One could assign numbers to each of the columns and then use a random number generator to determine the column orders in the rotation. Figure 6.5 illustrates the table once this has been done. The properties of the Latin square are still maintained and topics are no longer completed consecutively. Note that even after randomization of the columns (not topics) it is still the case that each topic will be completed first, second, third, etc. an equal number of times.

The second thing that a standard Latin square design does not address is the order effects introduced by the interfaces. One assumption behind a Latin square rotation is that there is no interaction

between the items represented by the rows and columns. Notice in Figures 6.4 and 6.5 that Interface 1 is always used first, Interface 2 second and Interface 3 third. The previous discussion of order effects as they relate to a fixed topic order also applies to a fixed interface order. Learning and fatigue may conspire to impact the results.

6.4.3 A Graeco-Latin Square Design

The solution to the problem described above is to rotate the order in which subjects experience the interfaces. This can be accomplished with a *Graeco-Latin square* which is a combination of two or more Latin squares. This is essentially equivalent to reproducing the Latin square in Figure 6.4 above three times, each with a different interface order. A single representation of this is displayed in Figure 6.6. In this Figure, the interfaces are now represented within the cells instead of as column headings. The column headings represent points in time (or order) and the rows represent subjects. For instance, the first user would use Interface 1 to complete Topics 1 and 2, and then Interface 2 to complete Topics 3 and 4, etc.

Subjects	Time 1	Time 2	Time 3
S1	I ₁ : 1, 2	I ₂ : 3, 4	I ₃ : 5, 6
S2	I ₁ : 2, 3	I ₂ : 4, 5	I ₃ : 6, 1
S3	I ₁ : 3, 4	I ₂ : 5, 6	I ₃ : 1, 2
S4	I ₁ : 4, 5	I ₂ : 6, 1	I ₃ : 2, 3
S5	I ₁ : 5, 6	I ₂ : 1, 2	I ₃ : 3, 4
S6	I ₁ : 6, 1	I ₂ : 2, 3	I ₃ : 4, 5
S7	I ₂ : 1, 2	I ₃ : 3, 4	I ₁ : 5, 6
S8	I ₂ : 2, 3	I ₃ : 4, 5	I ₁ : 6, 1
S9	I ₂ : 3, 4	I ₃ : 5, 6	I ₁ : 1, 2
S10	I ₂ : 4, 5	I ₃ : 6, 1	I ₁ : 2, 3
S11	I ₂ : 5, 6	I ₃ : 1, 2	I ₁ : 3, 4
S12	I ₂ : 6, 1	I ₃ : 2, 3	I ₁ : 4, 5
S13	I ₃ : 1, 2	I ₁ : 3, 4	I ₂ : 5, 6
S14	I ₃ : 2, 3	I ₁ : 4, 5	I ₂ : 6, 1
S15	I ₃ : 3, 4	I ₁ : 5, 6	I ₂ : 1, 2
S16	I ₃ : 4, 5	I ₁ : 6, 1	I ₂ : 2, 3
S17	I ₃ : 5, 6	I ₁ : 1, 2	I ₂ : 3, 4
S18	I ₃ : 6, 1	I ₁ : 2, 3	I ₂ : 4, 5

Fig. 6.6 A basic design with Graeco-Latin square rotation for topic and interface.

Note that this design has the same problem as the design in Figure 6.4: Interface 2 always follows Interface 1 except when Interface 2 is first. To address this problem, the same column randomization strategy described above can be applied. The column randomization should be applied after the Graeco-Latin square has been built, otherwise it cannot be ensured that each topic will be paired an equal number of times with each system.

Randomization should be used to assign subjects to the different rows in the table, even when the columns have been randomized. All experimental designs assume random assignment of subjects to conditions. To accomplish random assignment, numbers could be assigned to the rows in Figure 6.6 and a random number generator could be used to determine the order of the rows. Random assignment to condition controls for any potential differences that might be attributable to subjects. The assumption is that any individual differences in subjects (e.g., intelligence, search experience, and motivation) that might impact the results will be equally distributed across condition and therefore controlled as much as possible.

Notice that the rotation in Figure 6.6 provides insight into how many subjects are needed for the study. We know that we need at least 18 subjects to get through the rotation once and to keep the study completely balanced we would need to recruit subjects in batches of 18. However, this is not the only way to determine an appropriate sample size. Statistical power, representativeness and generalizability are also important factors.

6.4.4 Using the Mathematical Factorial to Construct a Design

Another method that can be used to construct an experimental design makes use of the mathematical factorial to enumerate all possible orders for topics and interfaces. However, it is important to note that this is *not* a Latin square rotation — it is a factorial rotation. It is also important to note that this type of rotation is infeasible and cannot be used to create a completely balanced design in most cases. For instance, in our example with three interfaces and six topics, we would first need

to do a factorial for interface type ($3! = 3 * 2 * 1$), which results in six possible orders. Next, we would need to do this for the six topics ($6! = 6 * 5 * 4 * 3 * 2 * 1$), which results in 720 possible orders. To make the experiment completely balanced, we would need 4320 subjects ($6 * 720$). It is unlikely that anyone would have the resources to recruit and study 4,320 subjects for a single study and besides, studying this many subjects is not really necessary since at some point statistical power plateaus. One might select a portion of these orders, but this would not result in a completely balanced design. However, there are some types of situations, where a factorial rotation is feasible. For instance, two interfaces ($2! = 2 * 1$) and four topics ($4! = 4 * 3 * 2 * 1$) results in 48 possible orders.

6.5 Randomization and User Choice

Another method that can be used to create experimental rotations is to randomly create orders. This can be done by combining different orders of the interfaces and topics, or in conjunction with a Latin Square, where the main variable of interest, interface type, is rotated using a Latin Square and topics are randomly assigned to subjects. It is often the case that researchers want to include more topics in a study to increase generalizability and randomization is selected as a way to assign topics to subjects. However, topics are unlikely to be equally represented in the data set (unless very large numbers of subjects are studied). Thus, results may be attributable to topics and/or topic interactions with other independent variables. If one can use a Latin or Graeco-Latin Square design, then it is a better choice for ensuring a more balanced experimental design.

Another approach is to give subjects a choice of topics. For instance, subjects might be presented with 10 topics and allowed to select four that they would like to research using the experimental systems. The justification for this approach is it helps increase subjects' motivation [292]. However, if one does this, one should be careful not to give subjects too many choices and have some control over how many topics are completed with particular systems. The danger in letting people choose is that the choices may naturally create a situation where topic

effects are present. There will likely be an unequal distribution in subjects' choices, resulting in some topics being overrepresented in a study and others being underrepresented. It may also be the case that some system–topic pairs occur more frequently unless extra effort is taken to prevent this.

6.6 Study Mode

The mode in which a study is administered can also vary. IIR evaluations can be administered in batch-mode, where multiple subjects complete the study at the same location and time or in single-mode, where subjects complete the study alone, with only the researcher present. The choice of mode is ultimately determined by the purpose of the study. In studies where subjects are deceived in different ways, completing different sequences of activities or will be interviewed, single-mode studies are more appropriate. If the experiment is relatively self-contained, subjects do similar things and can be directed via computer, then batch-mode is appropriate.

Batch-mode studies are very efficient — more subjects can be ran in a shorter period of time. However, it is important to note that subjects can influence one another even when they do not communicate verbally. For instance, in a batch-mode design, the first person to finish the study will likely signal to others that the end of the experiment is approaching. As a result, the remaining subjects might work faster and be less thoughtful, even if they are in different conditions that require more time. Thus, one should think carefully about non-verbal signals that are present in batch-mode studies, what these signals might communicate and how they might contaminate or change a subject's experiences and subsequent behaviors.

Studies can also be administered via the Web instead of in the laboratory. Toms et al. [272] adapted the traditional TREC Interactive Track IIR evaluation model so that it could be run on the Web. The WiRE framework provided an infrastructure where researchers could plug-in different systems or interfaces for evaluation and tailor common instruments, such as questionnaires, to their needs. Researchers are increasingly experimenting with different ways of administering

evaluations online, although the impact of this on the quality of the evaluation data is unclear. The main concern is that allowing people to login to a system and complete a study in any environment potentially introduces confounding variables that will be unknown to the researcher. For instance, one subject might complete the study while sitting in a loud environment, another might multi-task between the study and other tasks, including text messaging and emailing, while another might solicit help from others or refer to alternative resources of information while searching. Of course, these all represent real use scenarios (and subjects can be instructed about what is expected of them) and this may be what interests the researcher. However, such studies should not be treated as controlled experiments, because they are not.

6.7 Protocols

A study protocol is a step-by-step account of what will happen in a study. It is useful to have a document describing in detail exactly what should happen to guide the researcher. Check lists and other such documents can be used to ensure consistency in the administration of the study. This consistency helps maintain the integrity of the study and ensure that subjects experience the study in similar ways. Creating a detailed protocol also helps ensure that the experiment will run smoothly and that the researcher knows what to expect. In cases where multiple researchers are conducting a study, a protocol helps ensure that the same steps are followed for each subject.

6.8 Tutorials

When subjects encounter new IIR systems it is often the case that they need some instruction on how to use them. Many of the systems that IIR researchers investigate are experimental and thus, differ from the standard systems to which subjects are accustomed. In the past, researchers have created print tutorials to introduce subjects to an experimental system, while others have verbally administered tutorials. Of the two, the print option is best because it ensures that the

presentation is consistent. These days, an easy method for creating a tutorial is to record a video tutorial using screen capture software. This video can be played for each user and it can be guaranteed that what is told and how it is told is consistent. It is best to first develop a script before creating a video tutorial.

There are objections to the use of tutorials and other instructional materials on the grounds that they potentially bias subjects and that in real life people do not read instruction manuals. The issue related to bias is arguably the more important objection; the tutorial may suggest to subjects how they should interact and behave. If one is using a measure such as uptake of a new feature and the feature is prominently discussed in the tutorial, the measure may just reflect how cooperative subjects are, rather than their real interests in the feature. However, if the purpose of the experiment is to evaluate a new feature, then asking people to use the feature seems reasonable since it must be used in order to be evaluated. When it is necessary to provide a tutorial researchers should ensure consistency and balance in the presentation and consider how this experience might influence subjects' behaviors and the study results.

6.9 Timing and Fatigue

Another issue that needs to be considered is the length of time the study will last. This is a critical issue because typically subjects are performing activities that take some length of time to complete. Unlike studies in psychology, where hundreds of trials can be conducted in a single hour, very often only four search tasks can be completed in one hour. Moreover, search activities can be exhausting both mentally and physically. There are no set rules on how long one should give subjects to complete tasks; this is usually contingent on the type of task and study purpose. For instance, in an evaluation of Web search result surrogates, Käksi and Aula [158]) imposed short time limits in an attempt to simulate how people actually scan surrogates in real life. In many other evaluations, subjects are given 10–15 min to complete search tasks.

6.10 Pilot Testing

One way to get an estimate of how long a study will last is to conduct a pilot test. Pilot tests help researchers do a number of other things besides estimate time. They help researchers identify problems with instruments, instructions, and protocols; allow systems to be exercised in the same way they will be in the actual study; provide researchers with an opportunity to get detailed feedback from test subjects about the method; help researchers gain comfort with administering the study; and finally, they can be used to train inexperienced researchers. Ultimately, pilot tests help researchers identify and eliminate potential confounds and errors that might otherwise compromise the integrity of the study results.