Principles of Survey Research Part 2: Designing a Survey

Barbara A. Kitchenham
Department of Computer Science
Keele University, Staffordshire, UK
barbara@cs.keele.ac.uk

Shari Lawrence Pfleeger
Systems/Software, Inc.
Washington, DC, USA
s.pfleeger@ieee.org

Introduction

This second article of our series looks at the process of designing a survey. The design process begins with reviewing the objectives, examining the target population identified by the objectives, and deciding how best to obtain the information needed to address those objectives. However, we also need to consider factors such as determining the appropriate sample size and ensuring the largest possible response rate.

To illustrate our ideas, we use the three surveys described in Part 1 of this series to suggest good and bad practice in software engineering survey research.

Survey design

Designing a survey is very similar to designing an experiment (see [1]), in that the design must match the objectives so that the survey data and analysis answer the questions we are posing. Usually, a survey has one of two basic goals. In the first case, a survey may be attempting to describe a phenomenon of interest. That is, the survey wants to make more concrete a fuzzy picture of a product, process or population. For example, we may administer a survey of the population of software practitioners to find out just which kinds of projects use software configuration management. Or we may want to survey requirements analysts to determine which requirements elicitation techniques are the most popular. In this case, our survey usually has what is called a descriptive design, where we capture descriptors of the phenomenon.

In the second case, our survey aims to assess the impact of some intervention. For instance, we may want to survey project managers using different types of process assessment techniques to determine the extent to which each technique is regarded favorably by it users. Or we may want to determine if penetration testing is more effective in finding a particular kind of security flaw than an alternative testing method. These intervention-based cases require an experimental design to support hypothesis testing.

As with any experimental method, our survey designs can range from simple to complex; we must decide which of the candidate designs is best for the given situation. In other words, we want to select a design to provide the most effective means of obtaining the information needed to address our objectives. Here, "effective" means three things:

1. *resilient to bias*: We want a design that is likely not to be unduly swayed by a particular faction, aspect or opinion. That is, we want the results of the survey to representatively reflect the reality of the situation.
2. *appropriate*: We want a design that makes sense in the context of the population. It should be complex enough to address the issues raised by the study's objectives, and no more complex than it needs to be.
3. *cost-effective*: We want a design whose administration and analysis are within the means of the resources allocated to the survey. This cost-effectiveness applies to the survey participants, too; the results of the survey should be so useful to them that it is worth their time to complete the survey.

Descriptive designs

Suppose we decide, given our objectives and resources, to do an observational study. We have three design types from which to choose:

1. *Cross sectional:* In this type of study, participants are asked for information at one fixed point in time. For example, we may poll all the members of a software development organization at 10AM on a particular Monday, to find out what activities they are working on that morning. This information gives us a snapshot of what is going on in the organization.
2. *Cohort:* This type of study is forward-looking, providing information about changes in a specific population. It is the descriptive equivalent of a longitudinal study. For example, a cohort study may view the change in software developers' activity over a day, week, month or year. Or it may track the activities of a group of testers to determine in what order different kinds of tests are administered.
3. *Case control:* This study is retrospective, asking participants about their previous circumstances to help explain a current phenomenon. For example, suppose we are interested in whether successful software designers in our company could be identified early in their careers. We might survey our current lead designers to try to identify features common to their backgrounds, such as their level of educational achievement or their breadth of experience.

Recall the three survey examples we introduced in Part 1 of this series. The Lethbridge survey asked respondents about their levels of training and education. The Ropponen and Lyytinen study requested information about risk management practices from Finnish software projects. And the Pfleeger-Kitchenham study sought to determine what kinds of evidence were used to support technology adoption decisions. All three surveys were all

cross-sectional, case control studies, in which participants were asked about their past experiences at a particular fixed point in time. It is not simply coincidence that all our examples are of this type; in our experience, most surveys in software engineering have this kind of design.

Experimental designs

If our objectives suggest an experimental design, we have five main design options from which to choose. Each design type allows us to test a hypothesis formally.

1. *Concurrent control studies in which participants are randomly assigned to groups.* Concurrent control studies assemble participants and assign them to experimental groups at the same point in time. In such studies, participants are assigned to the experimental groups at random. Since proper randomization procedures are applied, the design corresponds to a true experiment. For instance, suppose the objective of our study is to determine whether training can change a manager's attitude toward quality assurance techniques. We can divide the company's population of managers into two groups: one that undergoes quality assurance training and one that does not. Each manager is randomly assigned to one group or the other, and an attitude survey is administered to all participants once the training is complete.
2. *Concurrent control studies in which participants are not randomly assigned to groups.* There are occasions where the experimental groups exist naturally, so participants cannot be assigned at random. For example, if we were assessing the impact of hiring practices on staff retention, we might want to know whether people with a masters degree have a more favorable view of their prospects in the company than people with doctorates. In this case, the design compares the masters degree group with the doctoral degree group, a natural split of the subjects.
3. *Self-control studies.* These studies are based on pre- and post-treatment measures, such as longitudinal studies or studies in which participants are asked for information before and after some intervention. For example, suppose we want to evaluate the value of a training course. Participants in the course might be asked to take part in a survey before the course is offered (the pre-course survey) and again after the course is completed (the post-course survey).
4. *Historical control studies.* In this type of design, comparisons between or among groups are based on data collected in other previous surveys. For example, universities may survey graduates five years after graduation about their job experiences. Results from one survey can be compared with results of another survey to see whether job prospects of graduates are changing over time.
5. *Studies using a combination of techniques.* The above designs may be combined in different ways. For example, we could incorporate pre- and post-treatment surveys into concurrent control studies. This design could be used if we want to compare two staff training methods.

Sample size

When we administer a survey, it is not usually cost-effective (and sometimes not even possible) to survey the entire population. Instead, we survey a subset of the population, called a *sample*, in the hope that the responses of the smaller group represent what would have been the responses of the entire group. When choosing the sample to survey, we must keep in mind the three aspects of design we mentioned above: avoidance of bias, appropriateness, and cost-effectiveness. That is, we want to select a sample that is truly representative of the larger population, is appropriate to involve in our survey, and is not prohibitively expensive to query. If we take these sample characteristics into account, we are more likely to get precise and reliable findings. We will explain how to obtain a valid sample in a future article; for now, suffice it to say that a sample is *not* the same thing as the set of responses obtained when we send questionnaires to all members of a population.

To assure precision and reliability, we must have previous information about the phenomena we are hoping to study. That is, we need to know the magnitude and variance of the effects of the survey's target activities. This information allows us to calculate the sample size required to detect any effect should it exist. We can obtain such information from previous studies (if they exist) or from a pilot study. For example, if we are surveying employee satisfaction in our company, it is useful to know the results of previous employee satisfaction surveys. If no such surveys exist, we may want to administer a pilot study to determine baseline employee satisfaction ratings before we design and administer a large-scale study throughout the company.

Cost is one of the key issues in survey research, especially when researchers assist in the survey administration in some way (such as by telephoning or working one-on-one with respondents). Clearly, the larger the sample the larger the cost, so there a strong incentive to determine how small a sample we can use and still have effective survey results. This cost consideration is not as pressing for self-administered surveys, particularly if we have no mailing or telephone costs. However, the front-end administrative expenses are not the only costs involved. We must also remember the effort and time involved in examining each returned survey and in analyzing the data. As we will see in subsequent articles in this series, it is important to screen responses for correctness and consistency, a process that can be time- and resource-consuming.

Cost can sometimes be reduced by taking advantage of existing technology or situations. For example, our technology survey was inserted in an existing mail-shot, so that no extra expenses were incurred to send out the survey to its target population. Similarly, Lethbridge solicited respondents over the Web, avoiding mailing costs too.

For non-administered surveys, it is often assumed that the larger the sample the better the survey will be. However, this assumption is not true. If we have a well-defined sample of an appropriate size, we are in a position to concentrate some of our effort on

follow-up activities, both to increase our sample size and to understand the reasons for non-response. That is, as with any study, we should take care to understand not only the group of people who supplied answers to our questions but also the group that did not take the time to respond. In doing so, we may find ways to improve the survey design, the survey instrument, and even the process by which the survey is administered. (Did we send out surveys during a vacation period? During a busy time for the target population? Was our survey too long or too difficult to understand? and so on.)

Response rates

It is not enough to decide how many people to survey. We must also take steps to be sure that enough people return the survey to yield meaningful results. Thus, any reliable survey should measure and report its *response rate*, that is, the proportion of participants who responded compared to the number who were approached.

The validity of survey results is severely compromised if there is a significant level of non-response. If we have a large amount of non-response but we can understand why and can still be sure that our pool of respondents is representative of the larger population, we can proceed with our analysis. But if there is large non-response and we have no idea why people have not responded, we have no way of being sure that our sample truly represents the target population. It is even worse to have no idea what the response rate is. For example, we had 171 responses to our survey, but we do not know exactly how many people subscribed to *Applied Software Development*. Similarly, because Lethbridge solicited responses from companies via the Web, the size of the target population was unknown; therefore, he could not calculate the response rate. Thus in both these cases the cost savings obtained by avoiding a direct mailing have compromised the validity of the surveys.

There are ways to plan the survey so that particular activities increase the response rate. For example, it is generally assumed that unsolicited mail surveys get poor response rates, because many people do not want to be bothered unless the benefits to them are very clear and worth their time. In some cases, we consider ourselves to be extremely lucky to get a 20% response on a first mailing. This result leads some researchers to declare "We had a 20% response rate, which is typical, so therefore the response rate was acceptable." However, 20% is still a poor response rate, and we can learn from the non-responses how to follow-up and increase the rate. Moreover, if we expect an initial low response rate, we can plan for *over-sampling*. That is, when we identify the sample size we require, we then sample more than the minimum required to allow for the expected non-response.

Follow-up plans include sending reminders to participants, and approaching individuals personally, if necessary. One-to-one approaches are particularly important if we want to assess the reason for non-response. For example, the researchers in Finland phoned a random sample of people who did not reply to their survey to ask them why they did not respond. This activity allowed them to confirm that non-response was not likely to have a systematic bias on their results.

There are several other steps we can take to improve response rate. In particular, we can ensure that people are:

- Able to answer the questions: That is, we can test the questions first to make sure that they are simple, unambiguous and written in a language our target population will understand.
- Are willing to answer the questions: That is, we should avoid asking intrusive or impertinent questions.
- Are motivated to answer the questions: That is, the respondents should see some clear benefit to answering the questions.

These issues are addressed in more detail in subsequent articles in this series.

References

[1] Shari Lawrence Pfleeger, "Experimental design and analysis in software engineering, Parts 1 to 5," *Software Engineering Notes*, 1995 and 1996.