Chapter 3
Personal Opinion Surveys

Barbara A. Kitchenham and Shari L. Pfleeger

Abstract Although surveys are an extremely common research method, survey-based research is not an easy option. In this chapter, we use examples of three software engineering surveys to illustrate the advantages and pitfalls of using surveys. We discuss the six most important stages in survey-based research: setting the survey's objectives; selecting the most appropriate survey design; constructing the survey instrument (concentrating on self-administered questionnaires); assessing the reliability and validity of the survey instrument; administering the instrument; and, finally, analysing the collected data. This chapter provides only an introduction to survey-based research; readers should consult the referenced literature for more detailed advice.

1. Introduction

Surveys are probably the most commonly used research method worldwide. Survey work is visible because we are often asked to participate in surveys in our private capacity, as electors, consumers, or service users. This widespread use of surveys may give the impression that survey-based research is straightforward, an easy option for researchers to gather important information about products, context, processes, workers and more. However, in our experience this is not the case. In this chapter, we will use actual survey examples to illustrate the attractions and pitfalls of the survey technique.

The three surveys we will use as our examples will be discussed in the next section. After that we will define what we mean by a survey. Then we will discuss the main activities that need to be considered when you undertake a survey:

- Setting the objectives
- Survey design
- Developing the survey instrument (i.e. the questionnaire)
- Evaluating the survey instrument
- Obtaining valid data
- Analysing the data

F. Shull et al. (eds.), *Guide to Advanced Empirical Software Engineering*.
© Springer 2008

2. Example Surveys

In this section we describe three software engineering surveys that will be used as examples throughout this chapter.

2.1. Technology Evaluation Survey

Recently we were involved in far from successful survey. A few years ago, Zelkowitz et al. (1998) surveyed practitioners to determine their confidence in different types of empirical evaluations as the basis for technology adoption decisions. Their findings indicated that the evidence produced by the research community to support technology adoption is not the kind of evidence being sought by practitioners. To build on Zelkowitz et al.'s work, a group of researchers, including ourselves, wanted to do a follow-up survey of managers, to find out what kinds of evaluations they make of proposed technologies, and what kinds of evidence they rely on for their technology decisions.

We had noticed that many newsletters often include reader survey forms, some of whose questions and answers could provide useful insight into managers' decision-making processes. We approached the publisher of *Applied Software Development*; he was eager to cooperate with the research community, and he agreed to insert a one-page survey in the newsletter and gather the responses. As a result, we took the following steps:

1. We designed a survey form and asked several of colleagues to critique it. The survey asked respondents to examine a list of technologies and tell us if the technology had been evaluated and if it had been used. If it had been evaluated, the respondents were asked to distinguish between a "soft" evaluation, such as a survey or feature analysis, and a "hard" evaluation, such as formal experiment or case study.
2. We "tested" the resulting survey form on a colleague at Lucent Technologies. We asked him to fill out the survey form and give feedback on the clarity of the questions and responses, and on the time it took him to complete the form. Based on his very positive reaction to the questionnaire, we submitted a slightly revised survey to the newsletter publisher.
3. The publisher then revised the survey, subject to our approval, so that it would fit on one page of his newsletter. The questionnaire was formatted as a table with four questions for each of 23 different software technologies (see Table 1).
4. The survey form was included in all copies of a summer 1999 issue of Applied Software Development.

Of the several thousand possible recipients of *Applied Software Development*, only 171 responded by sending their survey form back; thus, the response rate was low, which is typical in this type of survey. The staff at *Applied Software Development*

Table 1 Format of technology survey questionnaire

Technology/ technique	Did your company evaluate this technology?	Soft Evaluation techniques: read case studies, articles, talking with peers, lessons learned, or other more anecdotal evidence?	Hard Evaluation techniques: feature comparisons, performance benchmark, or other more quantitative evidence?	Are you now using the technique in some production work or most production work?
Specific software technology	Yes/No	Yes/No	Yes/No	Some/Most/None

transferred the data from the survey sheets to a spreadsheet. However, when the results of the survey were analyzed, it appeared that we had made errors in survey design, construction, administration and analysis that rendered any results inconclusive at best.

2.2. Software Education Survey

Lethbridge (1998, 2000) conducted surveys to help him understand those areas where practitioners feel they need more or better education. The goal of the surveys was to provide information to educational institutions and companies as they plan curricula and training programs. A secondary goal involved providing data that will assist educators and practitioners in evaluating existing and proposed curricula.

Lethbridge and his team recruited participants for the surveys in two ways: by approaching companies directly and asking them to participate, and by advertising for participants on the Web. To determine the effects of formal education, Lethbridge presented the respondents with a list of topics related to computer science, mathematics and business. For each topic, the respondent was asked "How much did you learn about this in your formal education?" The choices for answers ranged on a six-point ordinal scale from "learned nothing" to "learned in depth." Other questions included

- What is your current knowledge about this considering what you have learned on the job as well as forgotten?
- How useful has this specific material been to you in your career?
- How useful would it be (or have been) to learn more about this (e.g. additional courses)? (This question appeared in the first version of the survey.)
- How much influence has learning the material had on your thinking (i.e. your approach to problems and your general maturity), whether or not you have directly used the details of the material? Please consider influence on both your

career and other aspects of you life. (This question appeared in the second version of the survey.)

2.3. Software Risk Management Survey

Ropponen and Lyytinen (2000) described an examination of risk management practices. They administered a survey addressing two overall questions:

- What are the components of software development risk?
- What risk management practices and environmental contingencies help to address these components?

To find out the answers, the researchers mailed a questionnaire to each of a preselected sample of members of the Finnish Information Processing Association whose job title was "manager" or equivalent. They sent the questionnaire to at most two managers in the same company.

Ropponen and Lyytinen asked twenty questions about risk by presenting scenarios and asking the respondents to rate their occurrence with a five-point ordinal scale, ranging from "hardly ever" to "almost always." For example, the scenarios included:

> Your project is cancelled before completing it
> and
> Subcontracted tasks in the project are performed as expected.

The researchers posed additional questions relating to organizational characteristics, such as the organization's size, industry, type of systems developed, and contractual arrangement. They also sought technology characteristics, such as the newness of the technology, the complexity and novelty of technological solutions, and the process technologies used. Finally, they asked questions about the respondents themselves: their experience with different sizes of projects, their education, their experience with project management, and the software used.

3. What is a Survey?

To begin, let us review exactly what a survey is. A survey is not just the instrument (the questionnaire or checklist) for gathering information. It is a comprehensive research method for collecting information to describe, compare or explain knowledge, attitudes and behavior (Fink, 1995). Fowler (2002) defines a quantitative survey in the following way:

- The purpose of a survey is to produce statistics, that is, quantitative or numerical descriptions of some aspects of the study population.

- The main way of collecting information is by asking questions; their answers constitute the data to be analysed.
- Generally information is to be collected from only a fraction of the population, that is a sample, rather than from every member of the population.

In this chapter we will concentrate on surveys of this type where data is collected by means of a questionnaire completed by the subject. This excludes surveys that use a semi-structured interview schedule administered by the researcher. We will also exclude surveys using mainly open-ended questions, surveys based on observing participant behaviour and data mining exercises. Thus, we restrict ourselves to surveys that collect quantitative but subjective data (concerning individual's opinions, attitudes and preferences) and objective data such as demographic information for example a subject's age and educational level.

4. Setting Objectives

The first step in any survey research (or any research, for that matter!) is setting objectives otherwise referred to as problem definition. Each objective is simply a statement of the survey's expected outcomes or a question that the survey is intended to answer. For instance, a survey may hope to identify the most useful features of a front-end development tool, or the most common training needs for new hires.

There are three common type of objective:

- To evaluate the rate or frequency of some characteristic that occurs in a population, for example, we might be interested in the frequency of failing projects (Standish Group, 2003).
- To assess the severity of some characteristic or condition that occurs in a population, for example, we might be interested in the average overrun of software projects (Moløkken-Østvold et al., 2004).
- To identify factors that influence a characteristic or condition, for example, we might be interested in factors that predispose a process improvement activity towards failure or towards success Dybå (2005).

The first two types of survey objective are descriptive: they describe some condition or factor found in a population in terms of its frequency and impact. The second type of survey looks at the relationship existing among factors and conditions within a population.

As the objectives are defined in more detail, you should be able to specify:

- The hypotheses to be tested
- What alterative explanations are to be investigated or excluded
- What scope of survey project is appropriate to address the objectives
- What resources are necessary to achieve the objectives

At this stage it is important to decide whether a survey is an appropriate research method to address the stated objectives. You need to be able to answer questions of the type:

- Is it clear what population can answer the survey questions reliably?
- Is there a method of obtaining a representative sample of that population?
- Does the project have sufficient the resources to collect a sample large enough to answer the study questions?
- Is it clear what variables need to be measured?
- Is it clear how to measure the variables?

If you cannot answer all these questions positively, you need to consider whether a survey is an appropriate means to address your research objectives.

5. Survey Design

Two common types of survey design are:

- *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 10 AM 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.
- *Longitudinal:* This type of study is forward-looking, providing information about changes in a specific population over time. There are two main variants of longitudinal designs, you can survey the same people at each time period or you can survey different people.

Recall the three survey examples we introduced in Sect. 2. The Lethbridge survey asked respondents about their levels of training and education (see Lethbridge, 1998, 2000). The Ropponen and Lyytinen (2000) study requested information about risk management practices from Finnish software projects. 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 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.

There are other more complex forms of survey design, for example designs that compare different populations, or designs that aim to assess the impact of a change. For information on such designs see, for example, Shaddish et al. 2002).

The other issue to decide is the way in which the survey will be administered. Options include:

- Self-administered questionnaires (usually postal but increasingly Internet).
- Telephone surveys.
- One-to-one interviews.

The questions that can be addressed are influenced by this factor. In addition, strategies for obtaining reliable data such as question ordering and wording differ according to the administration method. Fowler provides a detailed examination of the pros and cons of different administration methods (Fowler, 2002). In this chapter we concentrate primarily on self-administered questionnaires.

6. Developing a Survey Instrument

In this section, we turn to how to develop a survey instrument. Survey instruments, which are usually questionnaires, are developed using the following steps:

- Search the relevant literature.
- Construct an instrument.
- Evaluate the instrument.
- Document the instrument.

We discuss instrument construction in this section and instrument validation and documentation in Sect. 7, using the three surveys described in Sect. 2 to illustrate good and bad practice.

6.1. Searching the Literature

As with any good investigative study, we must begin our work by looking through the literature. We need such searches to:

- Identify what other studies have been done on the topic.
- Determine how the previous studies' researchers collected their data. In particular, we want to find out what questionnaires or other data collection mechanisms were used.

There are many reasons for knowing what has come before. First, we do not want *unknowingly* to duplicate someone else's research. Second, we want to learn from and improve upon previous studies. For example, if previous studies have developed relevant validated instruments or questions that we can adopt, it makes our own survey easier to administer and validate. Similarly, if other researchers had problems with response rates, we will be aware of the need to adopt measures to address this problem. Finally, other studies may give us ideas about variables and issues we need to consider in designing our own studies.

6.2. Creating or Re-Using an Instrument

In software engineering, we often start from scratch, building models of a problem and designing survey instruments specifically for the problem at hand. However, in other disciplines, it is rare to develop a new survey instrument. Researchers usually

rely on using existing instruments, perhaps tailored slightly to accommodate variations on a common theme. This reliance on standard instrumentation has two important advantages.

1. The existing instruments have already been assessed for validity and reliability.
2. By using common instruments, it is easy to compare new results with the results of other studies.

When researchers in other disciplines cannot use an existing instrument, they are often able to amend existing instruments. An instrument might be amended if:

- It is too long to be used in entirety.
- A different population is being studied from the one for which the original instrument was designed.
- It needs to be translated.
- The data collection method is different in some way from the original instrument's data collection.

However, we must take care when considering amending an instrument. Our changes may introduce complications that make the research more difficult. For example:

- If the original instrument is copyrighted, we may need permission to change it.
- We must repeat pilot testing of the instrument.
- The new instrument must be assessed for validity and reliability.

Unfortunately, because most survey instruments in software engineering research are developed from scratch, we introduce many practical problems. In particular, software engineering research instruments are seldom properly validated.

6.3. Creating a New Questionnaire

A survey asks the respondents to answer questions for a reason, so the starting point in designing the survey instrument should always be the survey's purpose and objectives. However, simply converting a list of objectives into a set of questions seldom leads to a successful survey instrument. The type of question and wording of the questions and answers need to be carefully designed.

6.3.1. Question Types

When formulating questions for a survey instrument, you can express them in one of two ways: open or closed. A question is o*pen* when the respondents are asked to frame their own reply. Conversely, a question is c*losed* when the respondents are asked to select an answer from a list of predefined choices.

There are advantages and disadvantages to each type of question. Open questions avoid imposing any restrictions on the respondent. However, there are many different ways respondents may choose to answer a question. Moreover, no matter how carefully we word the question, open questions may leave room for misinterpretation and provision of an irrelevant or confusing answer. Thus, open questions can be difficult to code and analyze.

6.3.2. Designing Questions

Once we have an idea of what we want to ask, we must give some thought to how we want to pose the questions. Questions need to be precise, unambiguous and understandable to respondents. In order to achieve that we need to ensure that:

- The language used is appropriate for the intended respondents and any possibly ambiguous terms are fully defined.
- We use standard grammar, punctuation and spelling.
- Each question expresses one and only one concept so we need to keep questions short but complete and avoid double-barrelled questions.
- Questions do not included vague or ambiguous qualifiers.
- Colloquialisms and jargon are avoided.
- We use negative as well as positive questions but avoid simply negating a question or using a double negative.
- We avoid asking question about events that occurred a long time in the past.
- We avoid asking sensitive questions that respondents may not be willing to answer in a self-administered questionnaire.

It is also important to make sure that respondents have sufficient knowledge to answer the questions. It can be extremely frustrating to be asked questions you are not in a position to answer. For example, of the three surveys described in Sect. 2, two of the surveys (Lethbridge's survey and the Finnish survey) asked respondents about their personal experiences. In contrast, the survey of technology adoption asked respondents to answer questions such as

> Did your company evaluate this technology? Yes/No
> Are you now using the technique in some production work or most production work? Yes/No

In this case, we were asking people to answer questions on behalf of their company. The questions may have caused difficulties for respondents working in large companies or respondents who had worked for the company only for a relatively short period of time.

To see how wording can affect results, consider the two Lethbridge surveys. Each was on the same topic, but he changed the wording of his last question. In the first survey Lethbridge, 1998, question 4 was:

> How useful would it be (or have been) to learn more about this (e.g. additional courses)?

In his second survey (Lethbridge, 2000), question 4 was:

> How much influence has learning the material had on your thinking (i.e. your approach to problems and your general maturity), whether or not you have directly used the details of the material? Please consider influence on both your career and other aspects of your life.

The first version of the question is considerably better than the second version, because the second version is more complex and thus more difficult to interpret and understand. In particular, the second version appears to be two-edged (referring both to approach to problems and to general maturity) and rather imprecise (since it may not be clear what "general maturity" really means). However, further reflection indicates that even the first version of the question is ambiguous. Is the respondent supposed to answer in terms of whether (s)he would have benefited from more courses at university, or in terms of whether (s)he would benefit from industrial courses at the present time?

The survey of technologies posed questions about evaluation procedures in terms of how the respondent's company performed its evaluation studies. In particular, it asked questions about soft and hard evaluation techniques by defining them at the top of two of the columns:

> Soft evaluation techniques: Read case studies, articles, talking with peers, lessons learned or other more anecdotal evidence? Yes/No
> Hard evaluation techniques: feature comparison, performance benchmark, or other more quantitative evidence? Yes/No

These questions include jargon terms related to evaluation that may not be well understood by the potential respondents. Similarly, the researchers used jargon when defining the technology types as well: CASE tools, Rapid Application Development, 4GLs, and more. Were the questions to be redesigned, they should spell out each technology and include a glossary to describe each one. Such information ensures that the respondents have a common understanding of the terminology.

6.3.3. Designing Answers to Questions

Answers are usually of one of four types:

1. Numerical values (e.g. Age)
2. Response categories (e.g. Job type)
3. Yes/No answers
4. Ordinal scales.

Numerical values are usually straightforward but other types of answer may cause difficulties.

Response categories require all respondents to choose from a set of possible categories. They should be:

- Exhaustive but not too long
- Mutually exclusive

- Allow for multiple selections if required
- Include an "Other" category if the categories are not known to be exhaustive

Yes/No answers are particularly problematic. They suffer from acquiescence bias (Krosnick, 1990) as well as problems with lack of reliability (because people do not give the same answer on different occasions), imprecision (because the restrict measurement to only two levels) and many characteristics are broad in scope and not easily expressed as a single question (Spector 1992). Consider the question in the technology evaluation survey:

> Are you now using the technique in some production work or most production work?

In this case our question about technology use doesn't suit a two point Yes/No scale very well. The question needs an ordinal scale answer.

Generally it is better to use an ordinal scale for attitudes and preferences. There are three types of scale:

1. Agreement scales e.g. a response choice of the form: Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly Agree.
2. Frequency scales e.g. a response choice of the form: Never, Rarely, Seldom, Sometimes, Occasionally, Most of the time.
3 Evaluation scales e.g. a response choice of the form: Terrible, Inferior, Passable, Good, Excellent.

Like response categories, ordinal scales need to be exhaustive but not too long. Researchers usually restrict them to seven points. In addition, Krosnick recommended points on a scale be labeled with words (to assist reliability and validity) but not numbered (because numbers can be interpreted in unanticipated ways by respondents) (Krosnick, 1990).

However, understanding (and hence reliability) may also be increased if we define each point on a scale. For example, Lethbridge gives some indication of the detail needed to define an ordinal scale in his survey. Each of his four main questions has its own associated ordinal scale with responses defined in the context of the question. For instance, the question "How much did you learn about this at university or college" had the following scale:

Score	Definition
1	Learned nothing at all
2	Became vaguely familiar
3	Learned the basics
4	Became functional (moderate working knowledge)
5	Learned a lot
6	Learned in depth, became expert (learned almost everything)

Although the intermediate points on the scale are a little vague, the end points are clear and unambiguous. Lethbridge's scale conforms to the normal standard of

using between 5 and 7 choices along an ordinal scale. Lethbridge's scale is also a reasonably balanced one. A scale is *balanced* when the two endpoints mean the opposite of one another *and* the intervals between the scale points appear to be about equal. Creating equal distances between the scale points is called anchoring the instrument. It is difficult to create an anchored scale and even more difficult to validate that a scale is properly anchored.

A final issue that applies to ordinal scale categories is whether to include a "Don't know" category. There is some disagreement in the social science community about this issue. Some researchers feel that such choices allow respondents to avoid answering a question. However, it may be counter-productive to force people to answer questions they don't want to, or to force them to make a choice about which they feel ambivalent. The usual approach is to consider whether the respondents have been selected because they are in a position to answer the question. If that is the case a "Don't Know" category is usually not permitted.

6.3.4. Measuring Complex Concepts

Spector points out some concepts are difficult to map to single self-standing questions (Spector 1992). This may result in one (or both) of two type of unreliability

1. If people answer in different ways at different time
2. If people make mistakes in their responses.

He proposes measures based on *summated rating scales* to address this problem. A summated rating scale is a set of two or more items (i.e. questions) that address a specific topic or aspect of interest. Having multiple items improves reliability by reducing the chance of respondents making an error in their response and increases the precision with which a concept is measured.

6.4. Questionnaire Format

For self-administered questionnaires, it is important to consider both the format of the questionnaire and the questionnaire instructions. For formatting printed questionnaires, use the following checklist (much of which applies to Web-based questionnaires, too):

- Leave a space for the respondents to comment on the questionnaire.
- Use space between questions.
- Use vertical format, spaces, boxes, arrows, etc. to maximize the clarity of questions. However, do not overwhelm the respondent with "clever" formatting techniques (particularly for Web Questionnaires).
- Consider the use of simple grids.
- Consider the use of a booklet format.

- Have a good contrast between print and paper.
- Stick to a font size of 10–12.
- Use a font that is easy to read.
- Avoid italics.
- Use bolding, underlining or capitals judiciously and consistently for emphasis and instructions.
- Do not split instructions, questions and associated responses between pages.

The order in which questions are placed is also be important. Bourque and Fielder (1995) recommend questions be asked in a logical order, starting with easy questions first. However, although most questionnaires include demographic questions (that is, questions that describe the respondent) at the front of the questionnaire, Bourque and Fielder suggest putting them at the end instead. They point out that demographic details may be off-putting at the start of the questionnaire and so may discourage respondents.

The questionnaire must be accompanied by various administrative information including:

- An explanation of the purpose of the study.
- A description of who is sponsoring the study (and perhaps why).
- A cover letter using letterhead paper, dated to be consistent with the mail shot, providing a contact name and phone number. Personalize the salutation if possible.
- An explanation of how the respondents were chosen and why.
- An explanation of how to return the questionnaire.
- A realistic estimate of the time required to complete the questionnaire. Note that an unrealistic estimate will be counter-productive.

6.5. Response Rates and Motivation

It is often very difficult to motivate people to answer an unsolicited survey. Survey researchers can use inducements such as small monetary rewards or gifts, but these are not usually very successful. In general, people will be more motivated to provide complete and accurate responses if they can see that the results of the study are likely to be useful to them. For this reason, we should be sure that the survey instrument is accompanied by several key pieces of information supplied to participants:

- What the purpose of the study is.
- Why it should be of relevance to them.
- Why each individual's participation is important.
- How and why each participant was chosen.
- How confidentiality will be preserved.

Lethbridge (1998) attempted to motivate response with the following statement:

> The questionnaire is designed to discover what aspects of your educational background have been useful to you in your career. The results of the survey will be used to help improve curricula. All the information you provide will be kept confidential. In particular we have no intention of judging you as a person–we are merely interested in learning about the relevance of certain topics to your work.

By contrast, the technology adoption survey attempted to motivate response with the statement:

> Dear Executive, We are sponsoring a study for the University of X, and Professors Y and Z. It is only through our cooperative efforts with the academic community that we bring our commercial experiences to the classroom. Thank you for your help.

It fairly clear that Lethbridge's statement is likely to be more motivating although neither is compelling.

6.6. Questionnaire Length

Although we all know that we should strive for the shortest questionnaire that will answer our research questions, there is always a temptation to add a few extra questions "while we are going to all the trouble of organising a survey". This is usually a mistake. You should use pre-tests (see Sect. 7) to assess how long it takes to answer your questionnaire and whether the length (in time and number of questions) will de-motivate respondents.

If you have too many questions, you may need to remove some. Questions can usually be grouped together into topics, where each topic addresses a specific objective. One way to prune questions is to identify a topic that is addressed by many questions, and then remove some of the less vital ones. Another way is to remove some groups of questions. Keep in mind, though, that such pruning sometimes means reducing the objectives that the questionnaire addresses. In other words, you must maintain a balance between what you want to accomplish and what the respondents are willing to tell you. Validity and reliability assessments undertaken during pre-tests can help you decide which questions can be omitted with least impact on your survey objectives.

One way to reduce the time taken to complete a survey is to have standardized response formats. For example, in attitude surveys, responses are usually standardized to an ordinal scale of the form:

> Strongly Agree, Agree, Disagree, Strongly Disagree.

If all responses are standardized, respondents know their choices for each question and do not have to take time to read the choices carefully, question by question. Thus, respondents can usually answer more standard-format questions in a given time than non-standard ones.

6.7. Researcher Bias

An important consideration throughout questionnaire construction is the impact of our own bias. We often have some idea of what we are seeking, and the way we build the survey instrument can inadvertently reveal our biases. For example, if we create a new tool and distribute it free to a variety of users, we may decide to send out a follow-up questionnaire to see if the users find the tool helpful. If we do not take great care in the way we design our survey, we may word our questions in a way that is sure to confirm our desired result. For instance, we can influence replies by:

- The way a question is asked.
- The number of questions asked.
- The range and type of response categories.
- The instructions to respondents.

To avoid bias, we need to:

- Develop neutral questions. In other words, take care to use wording that does not influence the way the respondent thinks about the problem.
- Ask enough questions to adequately cover the topic.
- Pay attention to the order of questions (so that the answer to one does not influence the response to the next).
- Provide exhaustive, unbiased and mutually exclusive response categories.
- Write clear, unbiased instructions.

We need to consider the impact of our own prejudices throughout questionnaire construction. However, we also need to evaluate our questionnaire more formally, using methods discussed in Sect. 7.

7. Survey Instrument Evaluation

We often think that once we have defined the questions for our survey, we can administer it and gather the resulting data. But we tend to forget that creating a set of questions is only the start of instrument construction. Once we have created the instrument, it is essential that we evaluate it (Litwin, 1995). Evaluation is often called *pre-testing*, and it has several different goals:

- To check that the questions are understandable.
- To assess the likely response rate and the effectiveness of the follow-up procedures.
- To evaluate the reliability and validity of the instrument.
- To ensure that our data analysis techniques match our expected responses.

The two most common ways to organize an evaluation are focus groups and pilot studies. Focus groups are mediated discussion groups. We assemble a group of

people representing either those who will use the results of the survey or those who will be asked to complete the survey (or perhaps a mixture of the two groups). The group members are asked to fill in the questionnaire and to identify any potential problems. Thus, focus groups are expected to help identify missing or unnecessary questions, and ambiguous questions or instructions. As we will see below, focus groups also contribute to the evaluation of instrument validity.

Pilot studies of surveys are performed using the same procedures as the survey, but the survey instrument is administered to a smaller sample. Pilot studies are intended to identify any problems with the questionnaire itself, as well as with the response rate and follow-up procedures. They may also contribute to reliability assessment.

The most important goal of pre-testing is to assess the reliability and validity of the instrument. Reliability is concerned with how well we can reproduce the survey data, as well as the extent of measurement error. That is, a survey is reliable if we get the same kinds and distribution of answers when we administer the survey to two similar groups of respondents. By contrast, validity is concerned with how well the instrument measures what it is supposed to measure. The various types of validity and reliability are described below.

Instrument evaluation is extremely important and can absorb a large amount of time and effort. Straub presents a demonstration exercise for instrument validation in MIS that included a Pretest, Technical Validation and Pilot Project (Straub, 1989). The Pretest involved 37 participants, the Technical Validation involved 44 people using a paper and pencil instrument and an equal number of people being interviewed; finally the Pilot test analysed 170 questionnaires. All this took place before the questionnaire was administered to the target population.

7.1. Types of Reliability

In software, we tend to think of reliability in terms of lack of failure; software is reliable if it runs for a very long time without failing. But survey reliability has a very different meaning. The basic idea is that a survey is reliable if we administer it many times and get roughly the same distribution of results each time.

Test-Retest (Intra-observer) Reliability is based on the idea that if the same person responds to a survey twice, we would like to get the same answers each time. We can evaluate this kind of reliability by asking the same respondents to complete the survey questions at different times. If the correlation between the first set of answers and the second is greater than 0.7, we can assume that test-retest reliability is good. However, test-retest will not work well if:

- Variables naturally change over time.
- Answering the questionnaire may change the respondents' attitudes and hence their answers.
- Respondents remember what they said previously, so they answer the same way in an effort to be consistent (even if new information in the intervening time makes a second, different answer more correct).

Alternate form reliability is based on rewording or re-ordering questions in different versions of the questionnaire. This reduces the practice effect and recall problems associated with a simple test-retest reliability study. However, alternative form reliability has its own problems. Rewording is difficult because it is important to ensure that the meaning of the questions is not changed and that the questions are not made more difficult to understand. For example, changing questions into a negative format is usually inappropriate because negatively framed questions are more difficult to understand than positively framed questions. In addition, re-ordering results can be problematic, because some responses may be affected by previous questions.

Inter-observer (inter-rater) reliability is used to assess the reliability of non-administered surveys that involve a trained person completing a survey instrument based on their own observations. In this case, we need to check whether or not different observers give similar answers when they assess the same situation. Clearly inter-rater reliability cannot be used for self-administered surveys that measure personal behaviors or attitudes. It is used where there is a subjective component in the measurement of an external variable, such as with process or tool evaluation. There are standard statistical techniques available to measure how well two or more evaluators agree. To obtain more information about inter-rater reliability, you should review papers by El Emam and his colleagues who were responsible for assessing ISO/IEC 15504 Software Process Capability Scale, also known as SPICE (see for example El Emam et al., 1996, 1998).

Two reliability measures are particularly important for summated rating scales: the Cronbach alpha coefficient (Cronbach, 1951) and the Item-remainder coefficient. These measures assess the *internal consistency* of a set of items (questions) that are intended to measure a single concept. The item-remainder coefficient is the correlation between the answer for one item and sum of the answers of the other items. Items with the highest item-remainder are important to the consistency of the scale. The Cronbach alpha is calculated as

$$\alpha = \frac{k}{k-1} \times \frac{s_T^2 - \sum s_I^2}{s_T^2} \tag{1}$$

Where S_T^2 is the total variance of the sum of all the items for a specific construct and S_I^2 is the variance of an individual item and k is the number of items.

If variables are independent the variance of their sum is equal to the sum of each individual variance. If variables are not independent the variance of their sum is inflated by the covariance among the variables. Thus if the Cronbach alpha is small we would assume that the variables were independent and did not together contribute to the measurement of a single construct. If the Cronbach alpha is large (conventionally >0.7), we assume that the items are highly inter-correlated and together measure a single construct.

7.2. Types of Validity

As noted above, we also want to make sure that our survey instrument is measuring what we want it to measure. This called survey validity. Four types of validity are discussed below.

Face validity is a cursory review of items by untrained judges. It hardly counts as a measure of validity at all, because it is so subjective and ill-defined.

Content validity is a subjective assessment of how appropriate the instrument seems to a group of reviewers (i.e. a focus group) with knowledge of the subject matter. It typically involves a systematic review of the survey's contents to ensure that it includes everything it should and nothing that it shouldn't. The focus group should include subject domain experts as well as members of the target population.

There is no content validity statistic. Thus, it is not a scientific measure of a survey instrument's validity. Nonetheless, it provides a good foundation on which to base a rigorous assessment of validity. Furthermore if we are developing a new survey instrument in a topic area that has not previously been researched, it is the only form of preliminary validation available.

Criterion validity is the ability of a measurement instrument to distinguish respondents belonging to different groups. This requires a theoretical framework to determine which groups an instrument is intended to distinguish. Criterion validity is similar to *concurrent* validity and *predictive* validity. Concurrent validity is based on confirming that an instrument is highly correlated to an already validated measure or instrument that it is meant to be related to. Predictive validity is based on confirming that the instruments predicts a future measure or outcome that it is intended to predict.

Construct validity concerns how well an instrument measures the construct it is designed to measure. This form of validity is very important for validating summated measurement scales (Spector 1992). *Convergent* construct validity assesses the extent to which different questions which are intended to measure the same concept give similar results. *Divergent* construct validity assesses the extent to which concepts *do not correlate* with similar but distinct concepts. Like criterion validity, divergent and convergent construct validity can be assessed by correlating a new instrument with an already validated instrument. Dybå (2000) presents a software engineering example of the validation process for a software survey using summated measurement scales.

7.3. Validity and Reliability in Software Engineering Surveys

Generally, software engineering surveys are weak in the area of validity and reliability. For example, for many years, in the extensive literature relating to the CMM, there was only one reference to a reliability coefficient (the Cronbach's alpha) and that concerned the 1987 version of the Maturity Questionnaire (Humphrey, 1991).

Of the three surveys we discussed in Sect. 1.2, only the Finnish Survey (Ropponen and Lyytinen, 2000) made a concerted effort to undertake reliability and validity studies. The technology adoption survey used face validity only. Lethbridge discusses the basis for his questions, but his discussion of validity is based only on a post-hoc assessment of possible responder bias (Lethbridge, 1998, 2000). In contrast, the Finnish researchers used a panel of experts to judge the content validity of the questions. They also attempted to assess the internal reliability of their instrument. Unfortunately, they did not perform an independent pilot study. They analyzed their survey responses using principal components to identify strategies for managing risks. They then derived Cronbach alpha statistics (Cronbach, 1951) from the same responses. They found high values and concluded that their survey instrument had good reliability. However, Cronbach alpha values were bound to be high, because they measure the structure already detected by the principal component analysis.

7.4. Survey Documentation

After the instrument is finalized, Bourque and Fielder (1995) recommend starting to document the survey. If the survey is self-administered, you should consider writing an initial descriptive document, called a *questionnaire specification*. It should include:

- The objective(s) of the study.
- A description the rationale for each question.
- The rationale for any questions adopted or adapted from other sources, with appropriate citations.
- A description of the evaluation process.

Furthermore, once the questionnaire is administered, the documentation should be updated to record information about:

- Who the respondents were.
- How it was administered.
- How the follow-up procedure was conducted.
- How completed questionnaires were processed.

One of the major reasons for preparing documentation during the survey is that surveys can take a long time. It may be many months between first distributing a questionnaire and when we are able to analyze results. It takes time for respondents to reply and for the researchers to undertake all necessary follow-up procedures. This time lag means that it is easy to forget the details of instrument creation and administration, especially if documentation is left to the end of the study. In general, it is good research practice to keep an experimental diary or log book for any type of empirical studies.

When questionnaires are administered by interview, specifications are referred to as *interviewer specifications* and can be used to train interviewers as well as for reference in the field.

Once all possible responses have been received and all follow-up actions have been completed, we are in a position to analyze the survey data. This is discussed in the following sections. However before tackling analysis we look at the problem of obtaining a data set that is suitable for statistical analysis.

8. Obtaining Valid Data

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 three aspects of survey design: 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.

In this section, we describe how to obtain a valid survey sample from a target population. We discuss why a proper approach to sampling is necessary and how to obtain a valid sample. We also identify some of the sampling problems that affect software engineering surveys.

The main point to understand is that a valid sample is not simply the set of responses we get when we administer a questionnaire. A set of responses is only a valid sample, in statistical terms, if has been obtained by a random sampling process.

8.1. Samples and Populations

To obtain a sample, you must begin by defining a *target population*. The target population is the group or the individuals to whom the survey applies. In other words, you seek those groups or individuals who are in a position to answer the questions and to whom the results of the survey apply. Ideally, a target population should be represented as a finite list of all its members called a *sampling frame*. For example, when pollsters survey members of the public about their voting preferences, they use the electoral list as their sampling frame.

A valid sample is a *representative subset* of the target population. The critical word in our definition of a sample is the word “representative.” If we do not have a representative sample, we cannot claim that our results generalize to the target

population. If our results do not generalize, they have little more value than a personal anecdote. Thus, a major concern when we sample a population is to ensure that our sample is representative.

Before we discuss how to obtain a valid sample, let us consider our three survey examples. In Lethbridge's case, he had no defined target population. He might have meant his target population to be every working software developer in the world, but this is simply another way of saying the population was undefined. Furthermore, he had no concept of sampling even his notional population. He merely obtained a set of responses from the group of people motivated to respond. Thus, Lethbridge's target population was vague and his sampling method non-existent. So although he described the demographic properties of his respondents (age, highest education qualification, nationality etc.), no generalization of his results is possible.

With respect to the Pfleeger-Kitchenham survey, we noted previously that we were probably targeting the wrong population because we were asking individuals to answer questions on behalf of their companies. However, even if our target population was all readers of *Applied Software Development*, we did not have any sampling method, so our responses could not be said to constitute a valid sample.

In contrast, in the Finnish survey, Ropponen and Lyytinen had a list of all members of the Finnish Information Processing Association whose title was manager. Thus, they had a defined sampling frame. Then, they sent their questionnaires to a pre-selected subset of the target population. If their subset was obtained by a valid sampling method (surprisingly, no sampling method is reported in their article), their subset constituted a valid sample. As we will see later, this situation is not sufficient to claim that the actual responses were a valid sample, but it is a good starting point.

8.2. Obtaining a Valid Sample

We begin by understanding the target population. We cannot sample a population if we cannot specify what that population is. Our initial assessment of the target population should arise from the survey objectives, not from a sense of who is available to answer our questions. The more precisely the objectives are stated, the easier it will be to define the target population. The specific target population may itself be a subset of a larger population. It may be specified by the use of *inclusion* or *exclusion* criteria.

It is often instructive to consider the target population and sampling procedure from the viewpoint of data analysis. We can do this during questionnaire design but we should also re-assess the situation after any pretests or pilot tests of the survey instrument. At this point we will have some actual responses, so we can try out our analysis procedures. We need to consider whether the analyses will lead to any meaningful conclusions, in particular:

- Will the analysis results address the study objectives?
- Can the target population answer our research questions?

Considering the first question, Lethbridge's objectives were to provide information to educational institutions and companies as they plan curricula and training programs. This goal raises obvious questions: which educational institutions and which companies? Lethbridge's target population was poorly defined but can be characterized as any practising software engineer. Thus, we must ask ourselves whether replies from software engineers who would have attended different education institutions, worked in different companies or had different roles and responsibilities would indicate clearly how curricula and training courses could be improved. At the very least, general conclusions may be difficult. The results would need to be interpreted by people responsible for curricula or training courses in the light of their specific situation.

The next question concerns the target population. Will the target population provide useful answers? Lethbridge did not apply any inclusion or exclusion criteria to his respondents. Thus, the respondents may include people who graduated a very long time ago or graduated in non-computer science-related disciplines and migrated to software engineering. It seems unlikely that such respondents could offer useful information about current computer science- related curricula or training programs.

Consider now the survey of technology adoption practices. We have already pointed that the Pfleeger-Kitchenham target population was the set of organizations (or organizational decision-makers) making decisions about technology adoption. However, our sample population solicits information from individuals. Thus, our *sampling unit* (i.e. an individual) did not match their *experimental unit* (i.e. an organization). This mismatch between the population sampled and the true target population is a common problem in many surveys, not just in software engineering. If the problem is not spotted, it can result in spurious positive results, since the number of responses may be unfairly inflated by having many responses from organizations instead of one per organization. Furthermore if there are a disproportionate number of responses from one company or one type of company, results will also be biased.

The general target population of the Finnish survey of project risk was Finnish IT project managers. The actual sampling frame was specified as members of Finnish Information Processing Association whose job title was "manager" or equivalent. People were asked about their personal experiences as project managers. In general, it would seem that the sample adequately represents the target population, and the target population should be in a position to answer the survey's questions.

The only weakness is that the Finnish survey did not have any experience-related exclusion criteria. For instance, respondents were asked questions about how frequently they faced different types of project problems. It may be that respondents with very limited management experience cannot give very reliable answers to such questions. Ropponen and Lyytinen did consider experience (in terms of the number of projects managed) in their analysis of the how well different risks were managed. However, they did not consider the effect of lack of experience on the initial analysis of risk factors.

8.3. Sampling Methods

Once we are confident that our target population is appropriate, we must use a rigorous sampling method. If we want to make strong inferences to the target population, we need a probabilistic sampling method. We describe below a variety of sampling methods, both probabilistic and non-probabilistic.

8.3.1. Probabilistic Sampling Methods

A probabilistic sample is one in which every member of a target population has a *known, non-zero probability* of being included in the sample. The aim of a probabilistic sample is to eliminate subjectivity and obtain a sample that is both unbiased and representative of the target population. It is important to remember that we cannot make any statistical inferences from our data unless we have a probabilistic sample.

A *simple random sample* is one in which every member of the target population has the *same* probability of being included in the sample. There are a variety of ways of selecting a random sample from a population list. One way is to use a random number generator to assign a random number to each member of the target population, order the members on the list according to the random number and choose the first n members on the list, where n is the required sample size.

A stratified random sample is obtained by dividing the target population into subgroups called strata. Each stratum is sampled separately. Strata are used when we expect different sections of the target population to respond differently to our questions, or when we expect different sections of the target population to be of different sizes. For example, we may stratify a target population on the basis of sex, because men and women often respond differently to questionnaires. The number of members selected from each stratum is usually proportional to the size of the stratum. In a software engineering survey, we often have far fewer women than men in our target population, so we may want to sample within strata to ensure we have an appropriate number of responses from women. Stratified random samples are useful for non-homogeneous populations, but they are more complicated to analyze than simple random samples.

Systematic sampling involves selecting every nth member of the sampling frame. If the list is random, then selecting every nth member is another method of obtaining a simple random sample. However, if the list is not random, this procedure can introduce bias. Non-random order would include alphabetical order or date of birth order.

8.3.2. Cluster-Based Sampling

Cluster–based sampling is the term given to surveying individuals that belong to defined groups. For example, we may want to survey all members of a family group, or all patients at specific hospitals. Randomization procedures are based on

the cluster, not the individual. We would expect members of each cluster to give more similar answers than we would expect from members of different clusters. That is, answers are expected to be correlated within a cluster. There are well-defined methods for analyzing cluster data, but the analysis is more complex than that of a simple random sample (for example, see Levy and Lemeshow, 1999).

8.3.3. Non-Probabilistic Sampling Methods

Non-probability samples are created when respondents are chosen because the are easily accessible or the researchers have some justification for believing that they are representative of the population. This type of sample runs the risk of being biased (that is, not being representative of the target population), so it is dangerous to draw any strong inferences from them. Certainly it is not possible to draw any statistical inferences from such samples.

Nevertheless, there are three reasons for using non-probability samples:

- The target population is hard to identify. For example, if we want to survey software hackers, they may be difficult to find.
- The target population is very specific and of limited availability. For example if we want to survey senior executives in companies employing more than 5000 software engineers, it may not be possible to rely on a random sample. We may be forced to survey only those executives who are willing to participate.
- The sample is a pilot study, not the final survey, and a non-random group is readily available. For example, participants in a training program might be surveyed to investigate whether a formal trial of the training program is worthwhile.

Three methods of non-probabilistic sampling are discussed below.

Convenience sampling involves obtaining responses from those people who are available and willing to take part. The main problem with this approach is that the people who are willing to participate may differ in important ways from those who are not willing. For example, people who have complaints are more likely to provide feedback than those who are satisfied with a product or service We often see this kind of sampling in software engineering surveys.

Snowball sampling involves asking people who have participated in a survey to nominate other people they believe would be willing to take part. Sampling continues until the required number of responses is obtained. This technique is often used when the population is difficult for the researchers to identify. For example, we might expect software hackers to be known to one another, so if we found one to take part in our survey, we could ask him/her to identify other possible participants.

Quota sampling is the non-probabilistic version of stratified random sampling. The target population is spit into appropriate strata based on know subgroups (e.g. sex, educational achievement, company size etc.). Each stratum is sampled (using convenience or snowball techniques) so that number of respondents in each subgroup is proportional to the proportion in the population.

8.4. Sample Size

A major issue of concern when sampling is determining the appropriate sample size. There are two reasons why sample size is important. First, an inadequate sample size may lead to results that are not significant statistically. In other words, if the sample size is not big enough, we cannot come to a reasonable conclusion, and we cannot generalize to the target population. Second, inadequate sampling of clusters or strata disables our ability to compare and contrast different subsets of the population.

However, Fowler points out that there is no simple equation that can tell you exactly how large your sample ought to be (Fowler, 2002). In particular, he rejects sample size strategies based on a proportion of the population, typical sizes found in other studies, or statistical methods based on expected error levels. His suggestion is to consider your analysis plan and ensure that you have adequate sample sizes of the smallest important subgroups in your population.

8.5. 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 did not know exactly how many people subscribed to *Applied Software Development*, so we could not calculate response rate. 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 may have compromised the validity of the surveys.

It is not obvious what a sort of response rate we should expect. Baruch (1999) reviewed 175 IS surveys and found a median response rate was 60%, but it may be that conditions are different in SE than in IS. Currently, we have relatively few surveys in SE and many of those do not publish response rates.

There are several strategies that can be used to improve response rates. Some were discussed in Sect. 6.5, others include:

- 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.
- We should have follow-up plans to send reminders to participants.
- We should approach 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.
- It may be possible to perform statistical adjustments to correct for non-response.

However, recent research has suggested that achieving higher response rates do not necessarily mean more accurate results (Krosnick, 1990). If we have used probability sampling, low response rates may not imply lower representativeness.

9. Analysing Survey Data

In this section, we assume that you have designed and administered your survey, and now you are ready to analyze the data you have collected. If you have designed your survey properly, you should have already identified the main analysis procedures. Furthermore, if you have undertaken any pre-tests or pilot studies, you should have already tested the analysis procedures.

We discuss some general issues involved in analyzing survey data. However, we cannot describe in detail how to analyze all types of survey data, so we concentrate on discussing some of the most common analysis issues.

9.1. Data Validation

Before undertaking any detailed analysis, responses should be vetted for consistency and completeness. It is important to have a policy for handling inconsistent and or incomplete questionnaires. If we find that most respondents answered all questions, we may decide to reject incomplete questionnaires. However, we must investigate the characteristics of rejected questionnaires in the same way that we investigate non-response to ensure that we do not introduce any systematic bias. Alternatively, we may find that most respondents have omitted a few specific questions. In this case, it is more appropriate to remove those questions from the analysis.

Sometimes we can use all the questionnaires, even if some are incomplete. In this case we will have different sample sizes for each question we analyze and we must remember to report that actual sample size for each sample statistic. This approach is

suitable for analyses such as calculating sample statistics or comparing mean values, but not for correlation or regression studies. Whenever analysis involves two or more questions you need an agreed procedure for handling missing values.

In some cases, it is possible to use statistical techniques to "impute" the values of missing data (Little and Rubin, 1987). However, such techniques are usually inappropriate when the amount of missing data is excessive and/or the values are categorical rather than numerical.

It is important to reduce the chance of incomplete questionnaires when we design and test our instruments. A very strong justification for pilot surveys is that misleading questions and/or poor instructions may be detected before the main survey takes place.

The questionnaire related to the technology adoption survey (shown in Appendix 1) suffered badly in terms of incomplete answers. A review of the instructions to respondents made it clear why this had happened. The instructions said:

> If you are not sure or don't know an answer just leave the line blank; otherwise it is important to answer YES or NO to the first section of every Technique/Technology section.

With these instructions, perhaps it is not surprising that most of the questionnaires had missing values. However, replies were not just incomplete; they were also inconsistent. For example, some respondents left blank question 1 (Did your company evaluate this technology?) while replying YES to question 2, about the type of evaluation undertaken. Thus, blanks did not just mean "Don't know"; sometimes they also meant YES. Ambiguities of this sort make data analysis extremely difficult and the results dubious.

9.2. Partitioning the Responses

We often need to partition our responses into more homogeneous sub-groups before analysis. Partitioning is usually done on the basis of demographic information. We may want to compare the responses obtained from different subgroups or simply report the results for different subgroup separately. In some cases, partitioning can be used to alleviate some initial design errors. Partitioning the responses is related to data validation since it may lead to some replies being omitted from the analysis.

For example, we noted that Lethbridge did not exclude graduates from non-IT related subjects from his population nor did he exclude people who graduated many years previously. However, he knew a considerable amount about his respondents, because he obtained demographic information from them. In his first paper, he reported that 50% of the respondents had degrees in computer science or software engineering, 30% had degrees in computer engineering or electrical engineering, and 20% had degrees in other disciplines. He also noted that the average time since the first degree was awarded was 11.7 years and 9.6 years since the last degree. Thus, he was in a position to partition the replies and concentrate his analysis on recent IT graduates. However, since he did not partition his data, his results are extremely difficult to interpret.

9.3. Analyzing Ordinal and Nominal Data

Analyzing numerical data is relatively straightforward. However, there are additional problems if your data is ordinal or nominal.

A large number of surveys ask people to respond to questions on an ordinal scale, such a five-point agreement scale. The Finnish survey and Lethbridge's survey both requested answers of this sort. It is common practice to convert the ordinal scale to its numerical equivalent (e.g. the numbers 1–5) and to analyze the data as if they were simple numerical data. There are occasions when this approach is reasonable, but it violates the mathematical rules for analyzing ordinal data. Using a conversion from ordinal to numerical entails a risk that subsequent analysis will give misleading results.

In general, if our data are single peaked and approximately Normal, our risks of misanalysis are low if we convert to numerical values. However, we should also consider whether such a conversion is necessary. There are three approaches that can be used if we want to avoid scale violations:

1. We can use the properties of the multinomial distribution to estimate the proportion of the population in each category and then determine the standard error of the estimate. For example, Moses uses a Bayesian probability model of the multinomial distribution to assess the consistency of subjective ratings of ordinal scale cohesion measures (Moses, 2000).
2. We may be able to convert an ordinal scale to a dichotomous variable. For example, if we are interested in comparing whether the proportion who agree or strongly agree is greater in one group than another, we can re-code our responses into a dichotomous variable (for example, we can code "strongly agree" or "agree" as 1 and all other responses as 0) and use the properties of the binomial distribution. This technique is also useful if we want to assess the impact of other variables on an ordinal scale variable. If we can convert to a dichotomous scale, we can use logistic regression.
3. We can use Spearman's rank correlation or Kendall's tau (Siegel and Castellan, 1998) to measure association among ordinal scale variables.

There are two occasions where there is no real alternative to scale violations:

1. If we want to assess the reliability of our survey instrument using Cronbach's alpha statistic (Cronbach, 1951)..
2. If we want to add together ordinal scale measures of related variables to give overall scores for a concept.

The second case is not a major problem since the central limit theory confirms that the sum of a number of random variables will be approximately Normal even if the individual variables are not themselves Normal.

However, we believe it is important to understand the scale type of our data and analyze it appropriately. Thus, we do not agree with Lethbridge's request for respondents to interpolate between his scale points as they saw fit (e.g. to give a reply of 3.4 if they wanted to).

10. Conclusions

This chapter has discussed the issues involved in undertaking survey-based research, in particular surveys based on self-administered questionnaires. The main message of this chapter is that, in spite of its ubiquity, survey-based research is not a simple research method. It requires time and effort to understand the basic methodology as well as time and effort to create, validate and administer a survey instrument.

We have only scratched the surface of survey methodology in this chapter. We hope this chapter provides a useful starting point but we strongly advise that you consult the text books and research referenced in this chapter before undertaking a survey for the first time.

References

Bourque, L. and Fielder, E. *How to Conduct Self-administered and Mail Surveys*, Sage Publications, Thousand Oaks, CA, 1995.

Baruch, Y. Response rate in academic studies – a comparative analysis. *Human Relations*, 52(4), 1999, pp. 412–438.

Cronbach, L.J. Coefficient alpha and internal structure of tests. *Psychometrika*, 16(3), 1951, pp. 297–334.

Dybå, T. An empirical investigation of the key factors for success in software process improvement. *IEEE Transactions on Software Engineering*, 31(5), 2005, pp. 410–424.

Dybå, T. An instrument for measuring the key factors of success in software process improvement. *Empirical Software Engineering*, 5(4), 2000, pp. 357–390.

El Emam, K., Goldenson, D., Briand, L., and Marshall, P. Interrater Agreement in SPICE Based Assessments. *Proceedings 4th International Software Metrics Conference*, IEEE Computer Society Press, 1996, pp. 149–156.

El Emam, K., Simon, J.-M., Rousseau, S., and Jacquet. E. Cost Implications of Interrater Agreement for Software Process Assignments. *Proceedings 5th International Software Metrics Conference*, IEEE Computer Society Press, 1998, pp. 38–51.

Fowler, F.J. Jr. *Survey Research Methods*, Third Edition, Sage Publications, Thousand Oaks, CA, 2002.

Fink, A. *The Survey Handbook*, Sage Publications, Thousand Oaks, CA, 1995.

Humphrey, W. and Curtis, B. Comments on 'a critical look', *IEEE Software*, 8:4, July, 1991, pp. 42–46.

Krosnick, J.A. Survey research. *Annual Review of Psychology*, 50, 1990, pp. 537–567.

Lethbridge, T. A Survey of the Relevance of Computer Science and Software Engineering Education. *Proceedings of the 11th International Conference on Software Engineering Education*, IEEE Computer Society Press, 1998.

Levy, P.S. and Lemeshow, S. *Sampling of Populations: Methods and Applications*, Third Edition, Wiley Series in Probability and Statistics, Wiley, New York, 1999.

Lethbridge, T. What knowledge is important to a software professional. *IEEE Computer*, 33(5), 2000, pp. 44–50.

Little, R.J.A. and Rubin, D.B. *Statistical Analysis with Missing Data*, Wiley, New York, 1987.

Litwin, M. *How to Measure Survey Reliability and Validity*, Sage Publications, Thousand Oaks, CA, 1995.

Moses, J. Bayesian probability distributions for assessing measurement of subjective software attributes. *Information and Software Technology*, 42(8), 2000, pp. 533–546.

Moløkken-Østvold, K., Jørgensen, M., Tanilkan, S.S., Gallis, H., Lien, A. and Hove, S. A Survey on Software Estimation in the Norwegian Industry. *Proceedings 10th International Symposium on Software metrics. Metrics 2004*, IEEE Computer Society, 2004, pp. 208–219.

Ropponen, J. and Lyytinen, K. Components of software development risk: how to address them. A project manager survey. *IEEE Transactions on Software Engineering*, 26(2), 2000, pp. 98–112.

Shaddish, W.R., Cook, T.D., and Campbell, D.T. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin Company, New York, 2002.

Siegel, S. and Castellan, N.J. *Nonparametric Statistics for the Behavioral Sciences*, Second Edition, McGraw-Hill Book Company, New York, 1998.

Spector, P.E. *Summated Rating Scale Construction. An Introduction*, Sage Publications, Thousand Oaks, CA, 1992.

Standish Group. *Chaos Chronicles*, Version 3.0, West Yarmouth, MA, 2003.

Straub, D.W. Validating instruments in MIS research. *MIS Quarterly*, 13 (2), 1989, pp. 147–169.

Zelkowitz, M.V., Dolores, R.W., and Binkley, D. Understanding the culture clash in software engineering technology transfer. University of Maryland technical report, 2 June 1998.