Principles of Survey Research
Part 4: Questionnaire Evaluation

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

Shari Lawrence Pfleeger
RAND Corporation
Arlington VA 22202-5050
shari_pfleeger@rand.org

Abstract

This article discusses how to avoid biased questions in survey instruments, how to motivate people to complete instruments and how to evaluate instruments. In the context of survey evaluation, we discuss how to assess survey reliability i.e. how reproducible a survey's data is and survey validity i.e. how well a survey instrument measures what it sets out to measure.

Keywords: survey reliability, survey validity, researcher bias, respondent motivation

Introduction

In our previous article, we looked at what questions to ask: which ones, what type, and how many. In this article, we turn to three specific issues:

- how to motivate respondents to reply to self-administered instruments,
- how to avoid or diminish responder bias, and
- how to evaluate questionnaires and survey instruments.

As with all previous installments, we use the three surveys described in part 1 of this series ([6], [7], [8] and Appendix 1 of part 3) to illustrate how these issues are addressed in software engineering surveys.

Motivation

It is often very difficult to motivate people to answer an unsolicited survey. Survey researchers have 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 [6] 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, we attempted to motivate response to our technology survey 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 that ours.

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.

A survey we were recently asked to participate in gives a good example of researcher bias. The survey (organized by the IEEE but not yet published) was intended to evaluate the benefits of software engineer certification. So it asked questions about the potential benefits of certification. However, there is no opportunity for a respondent to contradict the view that certification is beneficial. If you do not think certification of software engineers is a good thing, the survey designers made it clear (by the choice of questions and response categories) that they do not want to hear from you.

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. For instance, there are examples in the social science literature that indicate that using a scale such as "zero, one, two, three, more than three" yields fewer answers than "zero, one to three, three to five, more than five," everything else being equal.
- 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 below.

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. 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, ambiguous questions and 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. We conducted a pilot test with our technology survey, using only one colleague to assess the questions. Our colleague did not reflect the varied backgrounds of our intended audience, so his approval of our questionnaire (as well as his rapid completion of the form) did not provide advanced warning of the difficulties in answering our questions.

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.

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, including different categories or types of reliability. The overall idea is that a survey is reliable if we administer it many times and get roughly the same distribution of results each time. But we can look at survey reliability from many different perspectives.

Test-Retest (Intra-observer) Reliability

Ideally, we would like to think that if the same person responds to a survey twice, we would get the same answers each time. This notion is the basis of test-retest reliability, and 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. For example, if we administer a questionnaire twice over a six-month period, a respondent may classify herself as "novice" on the first survey and "expert" on the second if she has used a particular technology extensively during the intervening six months.
- There is a large practice effect. For example, answering the survey the first time may make the respondent think about issues that she had not otherwise thought about. By the time the survey is administered the second time, the respondent may have different opinions or a different sense of what is happening in her organization or project.
- 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

In order to reduce the practice effect and recall problems associated with a simple test-retest reliability study, researchers sometimes give test-retest subjects different versions of the questionnaire. Questions are reworded or re-ordered in each version. This type of test-retest reliability is called alternative form reliability.

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. To see how, consider a famous example reported by Schuman and Presser [9]

of a survey undertaken in the USA that included the following two questions:

> A: If a US reporter is working in Russia, should he be allowed to report events without censorship? Yes/No
>
> B: If a Russian reporter is working in the US, should he be allowed to report events without censorship? Yes/No

If question A was asked first, the number of *Yes* responses to question B was much greater than when question B was asked first.

Lethbridge [6] and 7] attempted to cater for question order bias by having two different versions of his questionnaire, asking questions in different orders. The assignment of questionnaire to respondent was randomized. However, this was done to avoid potential bias, not for purposes of assessing questionnaire reliability.

Sometimes researchers deliberately include reworded questions in a survey instrument in order to check the internal consistencies of respondents' answers. This is usually only done in telephone surveys where the respondent does not see the questionnaire, or in surveys where the respondents are supervised and timed. In self-administered questionnaires duplicate reworded questions are likely to de-motivate respondents, since they:

- Increase the length of surveys.
- Imply that the researcher doesn't trust the respondents.
- Suggest that the researchers are trying to trick the respondents.
- Suggest that researchers think respondents are too stupid to notice that they are being tricked.

Internal consistency

Sometimes we ask groups of questions that measure different aspects of the same concept. We want to make sure that these aspects are consistent with respect to the overall concept. To test that they are, we can analyze the results of a pilot study using Cronbach's alpha statistic [2].

The alpha statistic measures the internal consistency reliability among a group of items that combine to form a single scale. It indicates how well the different items complement each other in their measurement of different aspects of the same variable or quality.

Inter-observer (inter-rater) reliability

Non-administered surveys sometimes 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. This type of consistency is called inter-observer or inter-rater reliability. 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 [3], [4]).

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 goal is not always as easy as it seems, since we often have to use surrogate variables to measure what we really want to know. For instance, just as we use the length of a column of mercury (an indirect measure) to tell us something about body temperature (what we really want to measure), we also use the number of faults found pre-release to suggest likely post-release software reliability.

There are several aspects of survey validity that we must consider when verifying that we are measuring our intended characteristics.

Face validity

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

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 valid form of preliminary validation available. As we explain below, other more formal methods of assessing validity assume that other similar survey instruments (or survey results) exist against which the new instrument can be assessed. It should be noted that we are not looking for consensus among reviewers, we are looking for potential problems. A majority opinion is not necessary correct, and a minority view may reveal critical flaws in a survey instrument.

Criterion validity

Criterion validity is a measure of how well one instrument compares with another instrument or predictor. Sometimes, we have an existing instrument we can compare with a newly devised questionnaire. This is the basis of *concurrent* criterion validity. Concurrent criterion validity compares a new instrument against one that is considered a "gold standard." For instance, a proposed new Capability Maturity Model questionnaire might be compared with the existing one to test its concurrent criterion validity. We can use correlations to indicate the extent to which the two questionnaires agreed. *Predictive* criterion validity assesses the ability of a survey to predict future phenomena. For

example, we can survey project managers about on-going projects and their characteristics in order to predict how much effort they will require. We can correlate the predictions with the outcomes to assess predictive criterion validity.

Construct validity

Construct validity concerns how an instrument "behaves" when used. We can think of this behavior in two ways. *Convergent* construct validity assesses the extent to which different data collection approaches produce similar results. *Divergent* construct validity assesses the extent to which results *do not correlate* with similar but distinct concepts. It often requires many years of experience to assess construct validity properly.

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 [5]. Of the three surveys we discussed in part 1 of this series, only the Finnish Survey [8] made a concerted effort to undertake reliability and validity studies. Our own studies of technology adoption used face validity only. Lethbridge [6] and [7] discusses the basis for his questions, but his discussion of validity is based only on a post-hoc assessment of possible responder bias.

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 [2] 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.

Survey Documentation

After the instrument is finalized, Bourke and Fielder [1] 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. In the next two articles, we discuss survey sampling and analysis issues.

References

[1] Linda Bourque and Eve Fielder, *How to conduct self-administered and mail surveys*, Sage Publications Inc., 1995.

[2] L.J. Cronbach, "Coefficient alpha and internal structure of tests," *Psychometrika*, 16(2), 1951, pp. 297-334.

[3] 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.

[4] 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.

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

[6] Timothy Lethbridge, A survey of the relevance of computer science and software engineering education, *Proceedings of the 11th International Conference on Software Engineering*, IEEE Computer Society Press, 1998.

[7] Timothy Lethbridge, What knowledge is important to a software professional, *IEEE Computer*, May 2000.

[8] J. Ropponen and K. Lyytinen, Components of software development risk: How to address them. A project manager survey, *IEEE Transactions on Software Engineering* 26(2), February 2000.

[9] Howard Schuman and Stanley Presser, *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context*. Sage Publications, Inc. 1996.