Evidence-Based Software Engineering: Opinion Surveys

Why and When?

The purpose of a survey is to collect opinions about and/or experiences of, a software engineering issue from a large group of people in a standard and systematic manner, and then to look for *patterns* in the data that can be generalised to a wider population than just the sample itself.
Surveys are useful when:

- A large number of people have experience of the issue and a broad overview of their experiences/opinions will answer important questions.
- Questions about the issue can be clearly specified and understood by all potential responders.

They provide results of the form:

- 80% of 41 software practitioners identified allocating tasks as a major challenge for global software development[1].
- 96% of 102 practitioners reported that personal experience was either an extremely or very useful part of their general approach to ensuring software quality[2].

Planning & Designing Surveys

Because surveys are familiar to us, it is easy to under-estimate this task. As with all empirical studies, we should begin by writing a *protocol* for the study, identifying the following key issues:

- The *research question(s)* which are the high-level issues the survey aims to investigate.
- The *data* needed to answer the research questions. This usually includes two types of data: data related directly to the questions, and data about the person answering the questions (such as the type of organisation they work for, how long they have worked in the software industry, their current job type…).
- The *questionnaire* that will be used for data collection. This converts the required data into specific questions.
- The *sampling frame*, which defines the set of people able to answer the research questions, for example, software maintainers, users of agile methods, students who have completed an advanced module.
- The *sampling technique,* which will be used to obtain the required sample. This is one of two methods: probabilistic or non-probabilistic. Whenever possible, a probabilistic sample method should be employed, because a random sample allows for more reliable generalisations.
- *Administering* the questionnaire, which involves inviting members of the sample to participate. In principle, the individuals should be personally invited and the reason for their invitation and the purpose of the survey explained. Putting a questionnaire on a web site and making it available to anyone who finds it does not deliver trustworthy conclusions.
- The *response rate* we are seeking, and what we will do about non-responses.
- The *sample size*, which needs to be large enough to form the basis for generalisation.
- The *analyses* that will be performed to answer the research questions and summarise the characteristics of the participants.
- The way that the outcomes will be *reported.*

An essential step is to test the protocol by conducting a 'dry run' using carefully chosen participants who can provide feedback on the understandability of the individual questions in the questionnaire and the time it takes to complete it.

Sampling techniques

Probabilistic sampling methods are used where we want a sample that is a representative cross-section of the population of a sampling frame. Ways of doing this include:

---

[1] See Niazi et al. (2016) Challenges of project management in global software development: A client-vendor analysis. *Information & Software Technology*, 80, pp1-19

[2] See Stevenson & Wood (2017) Recognising object-oriented software design quality: a practitioner-based questionnaire survey. *Software Quality J*. 26, pp321-365.

random sampling from the sampling frame
systematic sampling, using a random first index then taking every *n*th person
stratified sampling which tries to keep same profile of participant attributes (e.g. experience, job type) as the population
cluster sampling, using grouping of people who have some shared interest

Non-probabilistic sampling is used where we might have little control over who participates (as might happen if we create a web site to collect data and invite people to use it), or where we do not know enough about the population to be able to define a suitable sampling frame. Ways of doing this include:

purposive sampling, with participants selected on basis of being likely to respond
snowball sampling, that begins with one person from the population, who is asked to identify others who should be consulted, and so on…
self-selection sampling, where we collect from those who respond to an advert
convenience sampling, based on ready access or willingness to help

Response Rates

One characteristic of surveys is a low *response rate*. This is often around 10% of people directly contacted, with figures such as 30% being deemed very good. Various ways to improve this include giving a good explanation of why taking part is important, as well as monitoring responses and following up non-responders.

However, for statistical analysis you need moderate to large response rates. You need at least 63 observations, in each group, if you want to compare estimated probabilities and an 80% chance of correctly detecting a medium size difference (i.e. about an 0.5 difference) at the 0.05 significance level. In addition, the more statistical tests you do, the greater the chance you have of finding spurious results. For example, if you perform 10 statistical tests with a significance level of 0.05, you have an approximately 0.4 chance of obtaining a spurious positive. So, if you are expecting a relatively small response rate, don't plan on doing too many statistical tests.

Questionnaire Design

We should assess our design in terms of:

- *content validity*: whether the questions are a well-balanced sample of issues in the domain?
- *construct validity*: whether we are measuring the 'right' issues with our questions?
- *reliability*: would the questions get the same responses if given more than once to the same people?

So, we need to keep questions brief, relevant, unambiguous, specific (one issue per question) and objective. Can use either closed or open questions, where:

- *Closed* questions restrict the respondent to choosing from one of a set of options offered, which might be yes/no, a list for them to select from, a quantity, or a scale [rating]; they may also ask the respondent to order items [ranking] or to select a group from a list.
- *Open* questions leave it to the respondent to determine the form of answer. Open questions are more difficult to analyse than closed questions and may require qualitative analysis techniques, such as thematic analysis or content analysis.

We may need more than one question to address a particular topic, although this does complicate statistical analysis. A questionnaire also needs a clear introduction, should not mix the question formats too much, should group questions that are related to a particular topic, and not be too long.

Should I use Likert Scales?

SE studies often confuse Likert *scales* with Likert style *responses*. Lickert scales are a set of questions that *together* are used to assess a single concept, (e.g. the Technology Acceptance Model uses four related questions to assess the concept of *Software Usability*). These are complicated to design and validate. Likert style responses are ordinal scales of the form: strongly agree; agree, neither agree nor disagree, disagree, strongly disagree. These are frequently used in SE surveys to standardise questions concerning personal opinions. You are very likely to use Likert responses in an opinion survey.