CHAPMAN & HALL/CRC INNOVATIONS IN
SOFTWARE ENGINEERING AND SOFTWARE DEVELOPMENT

EVIDENCE-BASED SOFTWARE ENGINEERING AND SYSTEMATIC REVIEWS

Barbara Ann Kitchenham
David Budgen
Pearl Brereton

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

EVIDENCE-BASED SOFTWARE ENGINEERING AND SYSTEMATIC REVIEWS

Chapman & Hall/CRC Innovations in Software Engineering and Software Development

Series Editor
Richard LeBlanc
Chair, Department of Computer Science and Software Engineering, Seattle University

AIMS AND SCOPE

This series covers all aspects of software engineering and software development. Books in the series will be innovative reference books, research monographs, and textbooks at the undergraduate and graduate level. Coverage will include traditional subject matter, cutting-edge research, and current industry practice, such as agile software development methods and service-oriented architectures. We also welcome proposals for books that capture the latest results on the domains and conditions in which practices are most effective.

PUBLISHED TITLES

Computer Games and Software Engineering
Kendra M. L. Cooper and Walt Scacchi

Software Essentials: Design and Construction
Adair Dingle

Software Metrics: A Rigorous and Practical Approach, Third Edition
Norman Fenton and James Bieman

Software Test Attacks to Break Mobile and Embedded Devices
Jon Duncan Hagar

Software Designers in Action: A Human-Centric Look at Design Work
André van der Hoek and Marian Petre

Evidence-Based Software Engineering and Systematic Reviews
Barbara Ann Kitchenham, David Budgen, and Pearl Brereton

Fundamentals of Dependable Computing for Software Engineers
John Knight

Introduction to Combinatorial Testing
D. Richard Kuhn, Raghu N. Kacker, and Yu Lei

Building Enterprise Systems with ODP: An Introduction to Open Distributed Processing
Peter F. Linington, Zoran Milosevic, Akira Tanaka, and Antonio Vallecillo

Software Engineering: The Current Practice
Václav Rajlich

Software Development: An Open Source Approach
Allen Tucker, Ralph Morelli, and Chamindra de Silva

CHAPMAN & HALL/CRC INNOVATIONS IN
SOFTWARE ENGINEERING AND SOFTWARE DEVELOPMENT

EVIDENCE-BASED SOFTWARE ENGINEERING AND SYSTEMATIC REVIEWS

Barbara Ann Kitchenham
Keele University, Staffordshire, UK

David Budgen
Durham University, UK

Pearl Brereton
Keele University, Staffordshire, UK

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20151022

International Standard Book Number-13: 978-1-4822-2866-3 (eBook - PDF)

Contents

This page intentionally left blank

List of Figures

List of Tables

Preface

As a relatively young (and as we will later argue, still somewhat immature) discipline, *software engineering* is in an emergent[1] state for many purposes. Its foundations as a distinct sub-discipline of computing are widely considered to have been laid down at the 1968 NATO conference, although the term was probably in fairly regular use before that. Since then, ideas have ebbed and flowed, along with the incredibly rapid expansion and evolution of computing from an activity largely concerned with 'crunching numbers' in support of scientific research, to something that forms a pervasive element of everyday life. While this has helped to drive the development of software engineering as a discipline, the headlong pace has also meant that there has often been little opportunity to appraise and reflect upon our experiences of how software systems can be developed, how well the different approaches work, and under what conditions they are likely to be most effective.

The emergence of the concept of *evidence-based software engineering* (EBSE) can certainly be assigned a clear starting point, with the seminal paper being presented at the 2004 International Conference on Software Engineering (ICSE). In the decade that has followed, ideas about EBSE, and about its key tool, the systematic review, have evolved and matured; it has taken its place in the empirical software engineer's toolbox; and has helped to categorise and consolidate our knowledge about many aspects of software engineering research and practice. While few commercial software development activities can as yet even be described as 'evidence-informed', the philosophy of EBSE is beginning to be widely recognised and appreciated. As such then, this seems to be a suitable time to bring this knowledge together in a single volume, not least to help focus thinking about what we as a community might usefully do with that knowledge in the future.

Like Gaul, our book is divided into three parts[2]. In the first part we discuss the nature of evidence and the evidence-based practices centred around the systematic review, both in general and also as applying to software engineering. The second part examines the different elements that provide inputs to a systematic review (usually considered as forming a *secondary* study), especially the main forms of primary empirical study currently used in software

[1] An emergent process is one that is 'in a state of continual process change, never arriving, always in transition' (Truex, Baskerville & Klein 1999).

[2] Those with a classical education will remember that this was the first observation in Julius Caesar's *The Conquest of Gaul*, and quite possibly, that is the only thing that many of us remember from that work!

engineering. Lastly, the third part provides a practical guide to conducting systematic reviews (the *guidelines*), drawing together accumulated experiences to guide researchers and students when they are planning and conducting their own studies. In support of these we also include an extensive *glossary*, and an appendix that provides a *catalogue* of reviews that may be useful for practice and teaching.

This raises the question of who we perceive to be the audience for this book. We would like to think that almost anyone with any involvement in software engineering (in the broadest sense) can find something of use within it, given that our focus is upon seeking to identify what works in software engineering, when it works, and why. For the researcher, it provides guidance on how to make his or her own contribution to the corpus of knowledge, and how to determine where the research efforts might be directed to best effect. For practitioners, the book both explains the foundations of evidence-based knowledge related to software engineering practices, and also identifies useful examples of this. Finally, for teachers and students, it provides an introduction to the nature and role of empirical software engineering and explains what empirical studies can tell us about our subject.

So, how should the aspiring empiricist, or even the merely curious, approach all of this material, assuming that he or she might be reluctant to attempt to devour each chapter in turn, in the way that they would read a novel? We would suggest that the first few chapters provide a background to EBSE that should be relevant to anyone. These chapters explain the basic thinking about evidence-based studies and concepts, and show how they can be applied within a software engineering context.

The researcher, including of course, all PhD students, should additionally read the rest of Part I, so as to understand how to plan a secondary study. Armed with this understanding they can then turn to Part III, which provides essential practical guidance on the conduct of such a study, and which can then lead them through the steps of putting their plan into action. And, should any researcher determine that the ground is not yet solid enough for a secondary study, they can turn to Part II to learn something about how to conduct and report on a primary study in such a way as to make it a useful input to a future secondary study. Indeed, even when undertaking a secondary study, Part II should also be useful to the systematic reviewer when he or she is facing the tasks of data extraction and synthesis, by explaining something of the context behind the different forms of empirical study that provide the inputs to their analysis.

Practitioners and others who want to know more about EBSE and the use of secondary studies may find that Part I provides much of what they need in order to understand (and use) the outcomes from secondary studies. Likewise, teachers will, we hope, find much useful material in both Part I and Part II, in the latter case because an understanding of secondary studies is best founded upon a solid appreciation of the roles and forms of primary studies. Both

of these groups should also find material that is of direct usefulness in the catalogue of reviews provided in the appendix.

We are teachers as well as researchers, and should observe here that teaching the practices used in performing secondary studies to advanced undergraduates can be beneficial too. Students usually need to undertake a literature review as part of their individual 'capstone' projects, and adopting a systematic and objective approach to this can add valuable rigour to the outcomes.

In writing this book, we have drawn upon our own experiences with conducting systematic reviews and primary studies, and so our material and its organisation build upon the lessons that we have learned through these. These experiences have included both designing our own studies and reviewing the designs of others, and with conducting both methodological studies as well as ones that examine some established software engineering practices. Wherever possible we have tried to illustrate our points by drawing upon these experiences, as well as learning from those of many others, whose contribution to EBSE and its development we gratefully acknowledge.

This leads to an issue that always presents something of a challenge for evidence-based researchers such as ourselves, namely that of how to handle *citation*. As evidence-based software engineering researchers we usually feel it necessary to justify everything we possibly can by pointing to relevant evidence—but equally as authors, we are aware that this risks present the reader with a solid wall of reference material, which itself can form a distraction from gaining an understanding of key concepts. We have therefore tried to find a balance, providing citations whenever we think that the reader may possibly wish to confirm or clarify ideas for themselves. At the same time we have tried to avoid a compulsive need for justification at every opportunity, and especially when this is not really essential to enjoying the text—and of course, a sense of interest and enjoyment is exactly what we sincerely hope others will be able to experience from learning about EBSE and how the use of systematic reviews can help to inform software engineering as a discipline.

Finally, as a related point, since all the chapters of Part I relate to different aspects of secondary studies, we have provided a single set of suggestions for *further reading* at the end of this part, in order to avoid undue repetition. In Part II, where we address different forms of primary study in each chapter, we have reverted to the more conventional approach of providing recommendations for further reading at the end of each chapter.

This page intentionally left blank

Glossary

The vocabulary used in this book has been derived from a variety of sources and disciplines, which is not unreasonable, as that is how the ideas of empirical software engineering have themselves been derived. Our glossary does not purport to be definitive, the aim is to convey the relevant concepts quickly, so that when consulting it, the reader does not have to stray far from the flow of what they are reading.

absolute (measurement scale): This is the most restrictive of the measurement scales and simply uses counts of the elements in a set of entities. The only operation that can be performed is a test for equality. (See also *measurement scales*.)

accuracy: The accuracy of a measurement is an assessment of the degree of conformity of a measured or calculated value to its actual or specified value.

accuracy range: The accuracy range tells us how close a sample is to the true population of interest, and is usually expressed as a plus/minus margin. (See also *confidence interval*.)

aggregation: The process of gathering together knowledge of a particular type and form (for example, in a table).

attribute: An attribute is a measurable (or at least, identifiable) characteristic of an entity, and as such provides a mapping between the abstract idea of a *property* of the entity and something that we can actually measure in some way.

between-subject: (Also known as *between-groups* or *parallel experiment*.) In this form of study, participants are assigned to different treatment (intervention) groups and each participant only receives one treatment.

bias: A tendency to produce results that depart systematically from the true results.

blinding: A process of concealing some aspect of an experiment from researchers and participants. In single-blind experiments, participants do not know which treatment they have been assigned to. In double-blind

experiments, neither participants nor experimenters know which treatment the participants have been assigned to. In software engineering we sometimes use blind-marking, where the marker does not know which treatment the participants adopted to arrive at their answers or responses.

case study: A form of *primary study*, which is an investigation of some phenomenon in a real-life setting. Case studies are typically used for *explanatory*, *exploratory* and *descriptive* purposes. The main two forms are *single-case* studies which may be appropriate when studying a representative case or a special case, but will be less trustworthy than *multiple-case* forms, where replication is employed to see how far different cases predict the same outcomes. (Note that the term *case study* is sometimes used in other disciplines to mean a narrative describing an example of interest.) Case study research is covered in detail in Yin (2014) and for software engineering, in Runeson, Höst, Rainer & Regnell (2012).

causality: The link between a stimulus and a response, in that one *causes* the other to occur (also termed cause and effect). The notion of some form of causality usually underpins *hypotheses*.

central tendency: The 'typical value' for a probability distribution. The three most common measures used for this are the *mean*, the *median* and the *mode*. (See the separate definitions of these.)

closed question: (As used in a questionnaire.) Such a question constrains respondents by requiring them to select from a pre-determined list of answers. This list may optionally include 'other' or 'don't know' options. (See also *open question*.)

conclusion validity: (See *validity*.)

confidence interval: This is an assessment of how sure we are that the region within the stated interval around our measured mean does contain the true mean. This is expressed as a percentage, for example, a confidence interval of 95% (which corresponds to two standard deviations either side of the mean) means that there is a 95% likelihood that the true population mean lies within two standard deviations of our sample mean. So, for this value of the confidence interval, if we did many independent experiments and calculated confidence intervals for each of these, the true mean of the population being studied would be within the confidence limits in 95% of these.

confounding factor: An undesirable element in an experimental study that produces an effect that is indistinguishable from that of one of the treatments.

construct validity: (See *validity*.)

content validity: (As used in a survey.) Concerned with whether the questions are a well-balanced sample for the domain we are addressing.

control group: For laboratory experiments we can divide the participants into two groups—with the *treatment group* receiving the experimental treatment being investigated, and the experimental context of the *control group* involving no manipulation of the independent variable(s). It is then possible to attribute any differences between the outcomes for the two groups as arising from the treatment.

controlled experiment: (See *laboratory experiment*, *field experiment* and *quasi-experiment*.)

convenience (sample): A form of *non-probabilistic sampling* in which participants are selected simply because it is easy to get access to them or they are willing to help. (See *sampling technique*.)

cross-over: (See *within-subject*.)

dependent variable: (Also termed *response variable* or *outcome variable*.) This changes as a result of changes to the independent variable(s) and is associated with an *effect*. The outcomes of a study are based upon measurement of the dependent variable.

descriptive (survey): (See *survey*.)

direct measurement: Assignment of values to an attribute of an entity by some form of counting.

divergence: A divergence occurs when a study is not performed as specified in the *experimental protocol*, and all divergences should be both recorded during the study and reported at the end.

double blinding: (See *blinding*.)

dry run: For an experiment, this involves applying the experimental treatment to (usually) a single recipient, in order to test the experimental procedures (which may include training, study tasks, data collection and analysis). May sometimes be termed a *pilot experiment*. A similar activity may be performed for a survey instrument.

effect size: The effect size provides a measure of the strength of a phenomenon. There are many measures of effect size to cater to different types of treatment outcome measures, including the standardized mean differences, the log odds ratio, and the Pearson correlation coefficient.

empirical: Relying on observation and experiment rather than theory (*Collins English Dictionary*).

ethics: The study of standards of conduct and moral judgement (*Collins English Dictionary*). Codes of ethics for software engineering are published by the British Computer Society and the ACM/IEEE. Any empirical study that involves human participants should be vetted by the researcher's local *ethics committee* to ensure that it does not disadvantage any of the participants in any way.

ethnography: A form of observational study that is purely observational, and hence without any form of intervention or participation by the observer.

evidence-based: An approach to empirical studies by which the researcher seeks to identify and integrate the best available research evidence with domain expertise in order to inform practice and policy-making. The normal mechanism for identifying and aggregating research evidence is the *systematic review*.

exclusion criteria: After performing a search for papers (primary studies) when performing a systematic review, the exclusion criteria are used to help determine which ones will not be used in the study. (See also *inclusion criteria*.)

experiment: A study in which an intervention (i.e. a treatment) is deliberately controlled to observe its effects (Shadish, Cook & Campbell 2002).

external attribute: An external attribute is one that can be measured only with respect to how an element relates to other elements (such as reliability, productivity, etc.).

field experiment: An experiment or quasi-experiment performed in a natural setting. A field experiment usually has a more realistic setting than a laboratory experiment, and so has greater external validity.

field study: A generic term for an empirical study undertaken in real-life conditions.

hypothesis: A testable *prediction* of a cause–effect link. Associated with a hypothesis is a *null hypothesis* which states that there are no underlying trends or dependencies and that any differences observed are coincidental. A statistical test is normally used to determine the probability that the null hypothesis can or cannot be rejected.

inclusion criteria: After performing a search for papers (primary studies) when performing a systematic review, the inclusion criteria are used to help determine which ones contain relevant data and hence will be used in the study. (See also *exclusion criteria*.)

independent variable: An independent variable (also known as a *stimulus* variable or an *input* variable) is associated with *cause* and is changed as a

result of the activities of the investigator and not of changes in any other variables.

indirect measurement: Assigning values to an attribute of an entity by measuring other attributes and using these with some form of 'measurement model' to obtain a value for the attribute of interest.

input variable: (See *independent variable*.)

instrument: The 'vehicle' or mechanism used in an empirical study as the means of data collection (for the example of a survey, the instrument might be a questionnaire).

internal attribute: A term used in software metrics to refer to a measurable attribute that can be extracted directly from a software document or program without reference to other software project or process attributes.

interpretivism: In information systems research and computing in general, interpretive research is 'concerned with understanding the social context of an information system: the social processes by which it is developed and construed by people and through which it influences, and is influenced by, its social setting' (Oates 2006). (See also *positivism*.)

interval scale: An interval scale is one whereby we have a well-defined ratio of intervals, but have no absolute zero point on the scale, so that we cannot speak of something being 'twice as large'. Operations on interval values include testing for equivalence, greater and less than, and for a known ratio. (See also *measurement scales*.)

interview: A mechanism used for collecting data from participants for surveys and other forms of empirical study. The forms usually encountered are *structured*, *semi-structured* and *unstructured*. The data collected are primarily subjective in form.

laboratory experiment: Sometimes referred to as a *controlled laboratory experiment*, this involves the identification of precise relationships between experimental variables by means of a study that takes place in a controlled environment (the 'laboratory') involving human participants and supported by quantitative techniques for data collection and analysis.

longitudinal: Refers to a form of study that involves repeated observations of the same items over long periods of time.

mapping study: A form of secondary study intended to identify and classify the set of publications on a topic. May be used to identify 'evidence gaps' where more primary studies are needed as well as 'evidence clusters' where it may be practical to perform a systematic review.

mean: Often referred to as the *average*, and one of the three most common measures of the *central tendency*. Computed by adding the data values and dividing by the number of elements in the dataset. It is only meaningful for data forms that have genuinely numerical values (as opposed to codes).

measurement: The process by which numbers or symbols are assigned to attributes of real-world entities using a well-defined set of rules. Measurement may be direct (for example, length) or indirect, whereby we measure one or more other attributes in order to obtain the value (such as measuring the length of a column of mercury on a thermometer in order to measure temperature).

measurement scales: The set of scales usually used by statisticians are absolute, nominal, ordinal, interval and ratio. (See the separate definitions of these for details). A good discussion of the scales and their applicability is provided in Fenton & Pfleeger 1997.

median: (Also known as the 50th percentile.) One of the three most common measures of the *central tendency*. This is the value that separates the upper half of a set of values from the lower half, and is computed by ordering the values and taking the middle one (or the average of two middle ones if there is an even number of elements). Then half of the elements have values above the median and half have values below.

meta-analysis: The process of statistical pooling of similar quantitative studies.

mode: One of the three most common measures of the *central tendency*. This is the value that occurs most frequently in a dataset.

nominal measurement scale: A nominal scale consists of a number of categories, with no sense of ordering. So the only operation that is meaningful is a test for equality (or inequality). An example of a nominal scale might be programming languages. (See also *measurement scales*.)

null hypothesis: (See *hypothesis*.)

objective: Objective measures are those that are independent of the observer's own views or opinions, and so are repeatable by others. Hence they tend to be quantitative in form.

observational scale: An observational scale seeks simply to record the actions and outcomes of a study, usually in terms of a pre-defined set of factors, and there is no attempt to use this to confirm or refute any form of hypothesis. Observational scales are commonly used for diagnosis or making comparison between subjects or between subjects and a benchmark. For research, they may be used to explore an issue and to determine whether more rigorous forms might then be employed.

open question: (As used in a questionnaire.) An open question is one that leaves the respondent free to provide whatever answer they wish, without any constraint on the number of possible answers. See also *closed question.*

ordinal scale: An ordinal scale is one that *ranks* the elements, but without there being any sense of a well-defined interval between the different elements. An example of such a scale might be *cohesion*, where we have the idea that particular forms are better than others, but no measure of how much. Operations are equality (inequality) and greater than/less than. (See also *measurement scales.*)

outcome variable: (See *dependent* variable.)

participant: Someone who takes part (participates) in a study, sometimes termed a *subject.* Participant is the better term in a software engineering context because involvement nearly always has an active element, whereas subject implies a passive recipient.

population: A group of individuals or items that share one or more characteristics from which data can be extracted and analysed. (See *sampling frame.*)

positivism: The philosophical paradigm that underlies what is usually termed the 'scientific method'. It assumes that the 'world' we are investigating is ordered and regular, rather than random, and that we can investigate it in an objective manner. It therefore forms the basis for hypothesis-driven research. For a fuller discussion, see (Oates 2006).

power: (See *statistical power.*)

precision: (See also *recall.*) In the context of information retrieval, the *precision* of the outcomes of a search is a measure of the proportion of studies found that are *relevant.* (Note that this makes no assumptions about whether or not all possible relevant documents were found.) If the number of relevant documents N_{rel} is defined as

$$N_{rel} = N_{retr} - \overline{N_{rel}}$$

where N_{retr} is the number retrieved and $\overline{N_{rel}}$ is the number that is classified as not relevant, then

$$precision = \frac{N_{rel}}{N_{retr}}$$

Hence if we retrieve 20 documents, of which 8 are not relevant, the value for precision will be $(20 - 8)/20$ or 0.6. So a value of 1.0 for precision indicates that all of the documents found were relevant, but says nothing about whether every relevant document was found.

primary study: This is an empirical study in which we directly make measurements about the objects of interest, whether by surveys, experiments, case studies, etc. (See also *secondary study*.)

proposition: (In the context of a case study.) This is a more detailed element derived from a *research question* and performs a role broadly similar to that of a *hypothesis* (and like a hypothesis can be derived from a theory). Propositions usually form the basis of a case study and help to guide the organisation of data collection (Yin 2014). However, an *exploratory* case study would not be expected to involve the use of any propositions.

protocol: In the context of empirical studies, this term is used in two similar (but different) ways.

- For empirical studies in general, the *experimental protocol* is a document that describes the way that a study is to be performed. It should be written before the study begins and evaluated and tested through a 'dry run'. During the actual study, any *divergences* from the protocol should be recorded. It is this interpretation that is used throughout this book.
- The practice of *protocol analysis* can be used for qualitative studies, forming a data analysis technique that is based upon the use of *think-aloud*. In this, the protocol provides a categorisation of possible utterances that can be used to analyse the particular sequence of words produced by a participant while performing a task, as well as to strip out irrelevant material (Ericsson & Simon 1993).

qualitative: A measurement form that (typically) involves some form of human judgement or assessment in assigning values to an attribute, and hence which may use an ordinal scale or a nominal scale. Qualitative data is also referred to as *subjective data*, but such data can be quantitative, such as responses to questions in survey instruments.

quantitative: A measurement form that involves assigning values to an attribute using an interval scale or (more typically) a ratio scale. Quantitative data is also referred to as *objective data*, however this is incorrect, since is it possible to have quantitative subjective data.

quasi-experiment: An experiment in which units are not assigned at random to the interventions (Shadish et al. 2002).

questionnaire: A data collection mechanism commonly used for surveys (but also in other forms of empirical study). It involves participants in answering a series of questions (which may be 'open' or 'closed').

randomised controlled trial (RCT): A form of large-scale controlled experiments performed in the field using a random sample from the population of interest and (ideally) *double blinding*. In clinical medicine this is

regarded as the 'gold standard' in terms of experimental forms, but there is little scope to perform RCTs in disciplines (such as software engineering) where individual participant skill levels are involved in the treatment.

randomised experiment: An experiment in which units are assigned to receive the treatment or alternative condition by a random process such as a coin toss or a table of random numbers.

ratio scale: This is a scale with well-defined intervals and also an absolute zero point. Operations include equality, greater than, less than, and ratio—such as 'twice the size'. (See also *measurement scales.*)

reactivity: This refers to a change in the participant's behaviour arising from being tested as part of the study, or from trying to help the experimenter (hypothesis guessing). It may also arise because of the influence of the experimenter (forming a source of bias).

recall: (See also *precision.*) In the context of information retrieval, the *recall* of the outcomes of a search (also termed *sensitivity*) is a measure of the proportion of all relevant studies found in the search. If the number of relevant documents N_{rel} is defined as

$$N_{rel} = N_{retr} - \overline{N_{rel}}$$

where N_{retr} is the number retrieved and $\overline{N_{rel}}$ is the number that is classified as not relevant, then

$$recall = \frac{N_{rel}}{N_{rel}^{tot}}$$

where N_{rel}^{tot} is the total number of documents that are relevant (if you know it). Hence if we retrieve 20 documents of which 8 are not relevant, and we know that there are no other relevant ones, then the value for recall will be $(20 - 8)/12$ or 1.0. So while a value of 1.0 for recall indicates that all relevant documents were found, it does not indicate how many irrelevant ones were also found.

research question: The research question provides the rationale behind any primary or secondary empirical study, and states in broad terms the issue that the study is intended to investigate. For experiments this will be the basis of the *hypothesis* used, but the idea is equally valid when applied to a more observational form of study.

response rate: For a survey, the response rate is the proportion of surveys completed and returned, compared to those issued.

response variable: An alternative term for the *dependent variable.*

sample: This is the set (usually) of people who act as participants in a study (for example, a survey or a controlled laboratory experiment). Equally, it can be a sample set of documents or other entities as appropriate. An important aspect of a sample is the extent to which this is representative of the larger population of interest.

sample size: This is the size of the sample needed to achieve a particular *confidence interval* (with a 95% confidence interval as a common goal). As a rule of thumb, if any statistical analysis is to be employed, even at the level of calculating means and averages, a sample size of at least 30 is required.

sampling frame: This is the set of entities that could be included in a survey, for example, people who have been on a particular training course, or who live in a particular place.

sampling technique: This is the strategy used to select a sample from a sampling frame and takes two main forms:

non-probabilistic sampling Employed where it is impractical or unnecessary to have a representative sample. Includes purposive, snowball, self-selection and convenience sampling.

probabilistic sampling An approach that aims to obtain a sample that forms a representative cross-section of the sampling frame. Includes random, systematic, stratified and cluster sampling.

secondary study: A secondary study does not generate any data from direct measurements, instead it analyses a set of *primary studies* and usually seeks to aggregate the results from these in order to provide stronger forms of *evidence* about a particular phenomenon.

statistical power: The ability of a statistical test to reveal a true pattern in the data (Wohlin, Runeson, Höst, Ohlsson, Regnell & Wesslen 2012). If the power is low, then there is a high risk of drawing an erroneous conclusion. For a detailed discussion of statistical power in software engineering studies, see (Dybå, Kampenes & Sjøberg 2006).

stimulus variable: (See *independent variable.*)

subjective: Subjective measures are those that depend upon a value judgement made by the observer, such as a ranking ('A is more significant than B). May be expressed as a qualitative value ('better') or in a quantitative form by using an ordinal scale.

survey: A comprehensive research method for collecting information to describe, compare or explain knowledge, attitudes and behaviour. The purpose of a survey is to collect information from a large group of people in a standard and systematic manner and then to seek *patterns* in the

resulting data that can be generalised to the wider population. Surveys can be broadly classified as being

- *experimental* when used to assess the impact of some intervention
- *descriptive* if used to enable assertions to be made about some phenomenon of interest and the distribution of particular attributes—where the concern is not *why* the distribution exists, but *what* form it has

synthesis: The process of systematically combining different sources of data (evidence) in order to create new knowledge.

systematic (literature) review: This is a particular form of *secondary study* and aims to provide an objective and unbiased approach to finding relevant primary studies, and for extracting, aggregating and synthesising the data from these.

tertiary study: This is a secondary study that uses the outputs of secondary studies as its inputs, perhaps by examining the secondary studies performed in a complete discipline or a part of it.

test–retest: Conventionally, this forms a measure of the *reliability* and *stability* of a survey instrument. Respondents are 'tested' at two well-separated points in time, and the responses are compared for consistency by means of a correlation test, with correlation values of 0.7–0.8 usually being considered satisfactory. Use of test–retest is only appropriate in situations where 'learning' effects are unlikely to occur within the intervening time period. In the context where a single researcher is performing a systematic review, the use of test–retest can be interpreted as being for the researcher to perform such tasks as *selection* and *data extraction* twice, with these being separated by a suitable time interval, and to check for consistency between the two sets of outcomes. Where possible, these tasks should be performed using different orderings of the data items, in order to reduce possible bias.

treatment: This is the 'intervention' element of an experiment (the term is really more appropriate to *randomised controlled trials* where the participants are recipients). In software engineering it may take the form of a task (or tasks) that participants are asked to perform such as writing code, testing code, reading documents.

triangulation: Refers to the use of multiple elements that reinforce one another in terms of providing evidence, where no single source would be adequately convincing. The 'sources' may be different forms of data, or the outcomes from different research methods.

validity: This is concerned with the degree to which we can 'trust' the outcomes of an empirical study, usually assessed in terms of four commonly encountered forms of *threat to validity*. The following definitions are based upon those provided in Shadish et al. (2002).

- *internal:* Relating to inferences that the observed relationship between treatment and outcome reflects a cause–effect relationship.
- *external:* Relating to whether a cause–effect relationship holds over other conditions, including persons, settings, treatment variables and measurement variables.
- *construct:* Relating to the way in which concepts are operationalised as experimental measures.
- *conclusion:* Relating inferences about the relationship between treatment and outcome variables.

within-subject: Refers to one of the possible design forms for a quasi-experiment. In this form, participants receive a number of different treatments, with the order in which these are received being randomised. A commonly encountered design (two treatments) is the *A/B–B/A crossover* whereby some participants receive treatment A and then treatment B, while others receive them in reverse order. A weaker version is a *before–after* design, whereby all participants perform a task, are then given some training (the treatment), and are then asked to undertake another task. (Also known as a *sequential* or *repeated-measures* experiment.)

Part I

Evidence-Based Practices in Software Engineering

This page intentionally left blank

Chapter 1

The Evidence-Based Paradigm

Since this is a book that is about the use of evidence-based research practices, we feel that it is only appropriate to begin it by considering what is meant by *evidence* in the general sense. However, because this is also a book that describes how we acquire evidence about software engineering practices, we then need to consider some of the ways in which ideas about evidence are interpreted within the rather narrower confines of science and technology.

Evidence is often associated with *knowledge*. This is because we would usually like to think that our knowledge about the world around us is based upon some form of evidence, and not simply upon wishful thinking. If we go to catch a train, it might be useful to have evidence in the form of a timetable that shows the intention of the railway company to provide a train at the given time that will take us to our destination. Or, rather differently, if we think that some factor might have caused a 'population drift' away from the place where we live, we might look at past census data to see if such a drift really has occurred, and also whether some groups have been affected more than others. Of course the link between evidence and knowledge is rarely well-defined, as in our second example, where any changes in population we observe might arise from many different factors. Indeed, it is not unusual, in the wider world at least, for the same evidence to be interpreted differently (just think about global warming).

In this chapter we examine what is meant by evidence and knowledge, and the processes by which we interpret the first to add to or create the second. We also consider some limitations of these processes, both those that are intrinsic, such as those that arise from the nature of the things being studied, and of data itself, and also those that arise from the inevitable imperfections of research practice. In doing so, we prepare the ground for Chapter 2, where we look at how the discipline of software engineering interprets these concepts, and review the characteristics of software engineering that influence the nature of our evidence—and hence the nature of our knowledge too.

1.1 What do we mean by evidence?

As noted above, evidence can be considered as being something that underpins *knowledge*, and we usually expect that knowledge will be derived from evidence through some process of *interpretation*. The nature of that interpretation can take many forms. For example, it might draw upon other forms of knowledge, as when the fictional detective Sherlock Holmes draws upon his knowledge about different varieties of tobacco ash, or about the types of earth to be found in different parts of London, in order to turn a clue into evidence. Interpretation might also be based upon mathematical or statistical procedures, such as when a scientist gathers together different forms of experimental and observational data—for example, using past medical records to demonstrate that smoking is a cause of lung cancer. Yet another, less scientific, illustration of the concept is when the jury at a criminal trial has to consider the evidence of a set of witnesses in order to derive reasonable knowledge about what actually happened. Clearly these differ in terms of when they arise, the form of knowledge derived, and the rigour of the process used for its derivation (and hence the *quality* of the resulting knowledge). What they do have in common though, is that our confidence about the knowledge will be increased if there is more than one source (and possibly form) of evidence. For the fictional detective, this may be multiple clues; for the clinical analysis it might involve using records made in many places and on patients who have different medical histories; for the jury, it may be that there are several independent witnesses whose statements corroborate each other. This process of *triangulation* between sources (a term derived from navigation techniques) is also an important means of testing the *validity* of the knowledge acquired.

Science in its many forms makes extensive use of these concepts, although not always expressed using this vocabulary. Over the years, particular scientific disciplines have evolved their own accepted set of empirical practices that are intended to give confidence in the validity and quality of the knowledge created from the forms of evidence considered to be appropriate to that discipline, and also to assess how strong that confidence is. Since this book is extensively concerned with different forms of *empirical* study, this is a good point to note that such studies are ones that are based upon *observation* and *measurement*. Indeed, this is a reminder that, strictly speaking, scientific processes never 'prove' anything (mathematics apart), they only 'demonstrate' that some relationship exists between two or more factors of interest. Even physicists, who are generally in the best position to isolate factors, and to exclude the effect of the observation process, are confronted with this issue. The charge on an electron, or the universal gravitational constant, may well be known to a very high level of precision, and with high confidence, but even so, some residual uncertainty always remains. For disciplines where it can be harder to separate out the key experimental characteristics and where (hor-

rors), humans are involved in roles other than as observers, so the element of variability will inevitably increase. This is of course the situation that occurs for many software engineering research studies, and we will look at some of the consequences in the next chapter.

When faced with evidence for which the values and quality may vary, the approach generally adopted is to use repeated observations, as indicated above, and even better, to gather observations made by different people in different locations. By pooling these, it becomes easier to identify where we can recognise repeated occurrences of patterns in the evidence that can be used to provide knowledge. This repetition also helps to give us confidence that we are not just seeing something that has happened by chance.

The assumption that it is meaningful to aggregate the observations from different studies and to seek patterns in these is termed a *positivist* philosophy. Positivism is the philosophy that underpins the 'scientific method' in general, as well as almost all of the different forms of empirical study that are described in this book.

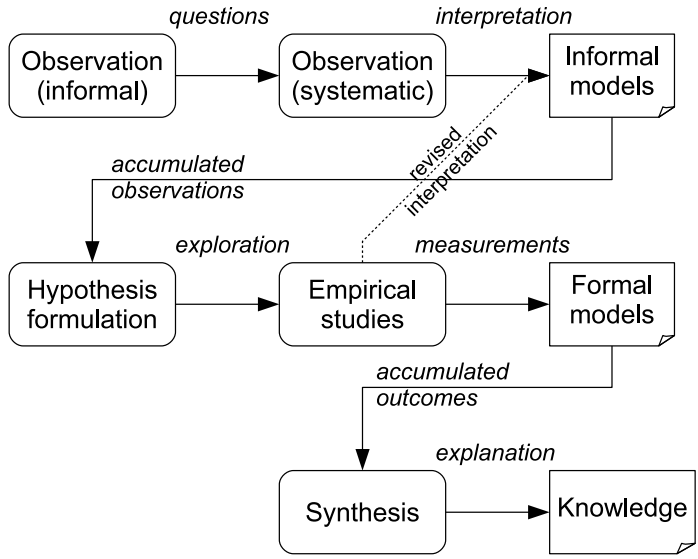

FIGURE 1.1: A simple model of knowledge acquisition.

Figure 1.1 shows a simple model that describes how these concepts relate to one another in a rather general sense. The top row represents how, having noticed the possible presence of some effect, we might begin gathering observations to create a rather informal model to describe some phenomenon. This model might well identify more than one possible cause. If this looks promising, then we might formulate a hypothesis (along the lines that "factor X causes outcome Y to occur") and perform some more systematically

organised studies to explore and test this model, during which process, we may discard or revise our ideas about some of the possible causes. Finally, to confirm that our experimental findings are reliable, we encourage others to repeat them, so that our knowledge is now accumulated from many sources and gathered together by a process that we refer to as *synthesis*, so that the risk of bias is reduced. Many well-known scientific discoveries have followed this path in some way, such as the discovery of X-rays and that of penicillin.

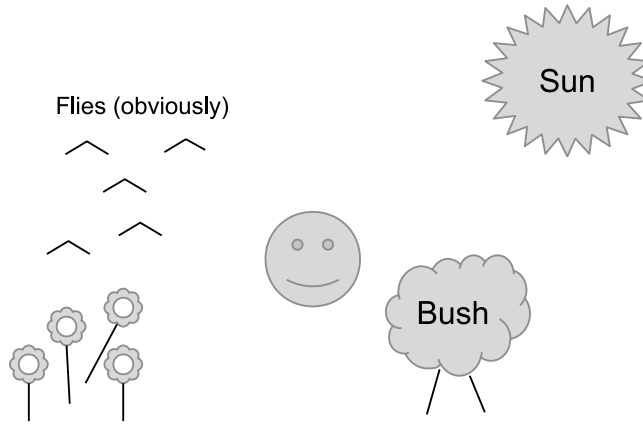

FIGURE 1.2: Does the bush keep the flies off?

Since this is rather abstract, let's consider a simple (slightly contrived but not unrealistic) example. This is illustrated (very crudely) in Figure 1.2. If we imagine that, while sitting out in a garden one day in order to enjoy the summer sunshine, we notice that we are far less bothered by flies when sitting near a particular bush, then this provides an example of informal observation. If we get enough good weather (we did say this example was contrived), we might try repeating the observation, perhaps by sitting near other bushes of that variety. If we continue to notice the effect, then this now constitutes an informal model. Encouraged by visions of the royalties that could arise from discovering a natural insecticide, we might then go on to pursue this rather more systematically, and of course, in so doing we will probably find all sorts of other possible explanations, or indeed, that it is not really an effect at all. But of course, we might also just end up with some systematically gathered knowledge about the insect-repellent nature of this plant (or perhaps, of this plant in conjunction with other factors).

This book is mainly concerned with the bottom two layers of the model shown in Figure 1.1. In Part I and Part III we are concerned with how knowledge from different sources can be 'pooled', while in Part II we provide a subject-specific interpretation of what is meant by the activities in the middle

layer. In particular, we will be looking at ways of gathering evidence that go beyond just the use of formal experiments.

In the next section we examine how the concepts of *evidence-based* knowledge and of *evidence-informed* decision-making, have been interpreted in the 20th and 21st centuries. In particular, we will discuss the procedures that have been adopted to produce evidence that is of the best possible quality.

1.2 Emergence of the evidence-based movement

It is difficult to discuss the idea of evidence-based thinking without first providing a description of how it emerged in clinical medicine. And in turn, it is difficult to categorise this as other than a movement that has influenced the practice and teaching of medicine (and beyond). At the heart of this lies the *Cochrane Collaboration*[1], named after one of the major figures in its development. This is a not-for-profit body that provides both independent guardianship of evidence-based practices for clinical medicine, and also custodianship of the resulting knowledge.

So, who was Cochrane? Well, Archie Cochrane was a leading clinician, who became increasingly concerned throughout his career about how to know what was the best treatment for his patients. His resulting challenge to the medical profession was to find the most effective and fairest way to evaluate available medical evidence, and he was particularly keen to put value upon evidence that was obtained from randomised controlled trials (RCTs). Cochrane's highly influential 1971 monograph "Effectiveness and Efficiency: Random Reflections on Health Services" (Cochrane 1971) particularly championed the extensive use of randomisation in RCTs, in order to minimise the influence of different sources of potential bias (such as trial design, experimenter conduct, allocation of subjects to groups, etc.). Indeed, he is quoted as saying that "you should randomise until it hurts", in order to emphasise the critical importance of conducting fair and unbiased trials.

Cochrane also realised that even when performed well, individual RCTs could not be relied upon to provide unequivocal results, and indeed, that where RCTs on a given topic were conducted by different groups and in different places, they might well produce apparently conflicting outcomes. From this, he concluded in 1979 that "it is surely a great criticism of our profession that we have not organised a critical summary by speciality or subspeciality, adapted periodically, of all relevant randomised controlled trials".

Conceptually, this statement was at complete variance with accepted scientific practice (not just that in clinical medicine). In particular, the role of the *review paper* has long been well established across much of academia, with

[1] www.cochrane.org

specialist journals dedicated to publishing reviews, and with an invitation to write a review on a given topic often being regarded as a prestigious acknowledgement of the author's academic standing. However, a problem with this practice was (and still is) that two people who are both experts on a given topic might well write reviews that draw contrasting conclusions—and with each of them selecting a quite different set of sources in support of their conclusion.

While this does not mean that an expert review is necessarily of little value, it does raise the question of how far the reviewer's own opinions may have influenced the conclusions. In particular, where the subject-matter of the review requires interpretation of empirical data, then how this is selected is obviously a critical parameter. A widely-quoted example of this is the review by Linus Pauling in his 1970 publication on the benefits of Vitamin C for combatting the common cold. His 'cherry-picking' of those studies that supported his theory, and dismissal of those that did not as being flawed, produced what is now regarded as an invalid conclusion. (This is discussed in rather more depth in Ben Goldacre's book, *Bad Science* (2009), although Goldacre does observe that in fairness, cherry-picking of studies was the norm for such reviews at the time when Pauling was writing—and he also observes that this remains the approach that is still apt to be favoured by the purveyors of 'alternative' therapies.)

Finding the most relevant sources of data is, however, only one element in producing reviews that are objective and unbiased. The process by which the outcomes (findings) from those studies are *synthesised* is also a key parameter to be considered. Ideas about synthesis have quite deep roots—in their book on literature reviews, Booth, Papaioannou and Sutton (2012) trace many of the ideas back to the work of the surgeon James Lind and his studies of how to treat scurvy on ships—including his recognition of the need to discard 'weaker evidence', and to do so by using an objective procedure. However, the widespread synthesis of data from RCTs only really became commonplace in the 1970s, when the term *meta-analysis* also came into common use[2].

Meta-analysis is a statistical procedure used to pool the results from a number of studies, usually RCTs or controlled experiments (we discuss this later in Chapters 9–11). By identifying where individual studies show consistent outcomes, a meta-analysis can provide much greater statistical authority for its outcomes than is possible for individual studies.

Meta-analysis provided one of the key elements in persuading the medical profession to pay attention. In particular, what Goldacre describes as a "landmark meta-analysis" looking at the effectiveness of an intervention given to mothers-to-be who risked premature birth, attracted serious attention. Seven

[2]One of us (DB) can claim to have had relatively early experience of the benefits of synthesis, when analysing scattering data in the field of elementary particle physics (Budgen 1971). Some experiments had suggested the possible presence of a very short-lived Σ particle, but this was conclusively rejected by the analysis based upon the composite dataset from multiple experiments.