ASE-EVAL: Software Evaluation Glossary—Version 1.1

David Budgen

November 12, 2007

General

The terminology used for empirical software engineering in general is drawn from a variety of other disciplines—and hence may not always even be used consistently between different authors and documents. Hence the aim of this document is to try to provide a short glossary of terms, and wherever possible, to cite an authority for any definitions used.

It is something of a 'work in progress' and hence may still be somewhat incomplete. My thanks to Prof. Barbara Kitchenham for reviewing this document and making a number of extremely useful corrections and observations.

A

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*.) [N.B. This term is sometimes used rather differently to refer to a *ratio* scale which incorporates a natural and unequivocal definition of the measurement unit!]

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. (See also *precision*).

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 level*.

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.

B

between-subject (Aka *parallel experiment*) Refers to one of the possible designs of a laboratory experiment. In this form, participants are assigned to different treatment (intervention) groups on the basis of one or more criteria and each participant only receives one treatment. (See also *within-subject*.)

biased Unfairly favouring one treatment over another in an experiment.

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 triple-blind experiments, as well as the participants and researchers, the statisticians analysing the results from an experiment are also blinded by being given treatment identifiers and not being told which identifier refers to which treatment. 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.

box plot A box plot is a means of showing the way that data values are distributed, in particular, of demonstrating any *skew* in the distribution in a visual manner. The core 'box' has a line for the *median* value, and the upper and lower bounds of the box are the *upper quartile* and *lower quartile* values respectively, where these are the data values with index position half way between the median and the upper and lower values of the dataset. The 'whiskers' on the 'tails' are determined by multiplying the box length by 1.5 and adding/subtracting the values of the upper/lower fourths respectively (and then truncating to the nearest actual value in the dataset). Any values lying outside of these are plotted as separate 'outliers'.

C

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 the book (Yin 2003) and more concisely in (Oates 2006).

causality The link between a stimulus and a response, in that one *causes* the other to occur. The notion of some form of causality usually underpins *hypotheses*.

cause-effect (See also *causality*.)

central tendency This is where the majority of data values are to be found. The three different 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 restricts the respondent by giving them a list of all permissible answers. Such a list may optionally include "other" or "don't know" options. See also *open question*.

conclusion validity (See *validity*.)

confidence level To generalise from an experimental sample to the wider population of interest we need to know the size of that population and also determine how sure we are that the values for our sample represent the population (the *confidence level*). See also *accuracy range* and *sample size*.

confounding factor This is an undesirable element in an empirical study that produces an effect that is indistinguishable from that of one of the treatments. (A common example for software engineering is any prior relevant experience that participants may have.)

construct validity (See *validity*.)

content validity (As used in a questionnaire.) 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 treatment and 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* and *quasi-experiment*).

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

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

D

delphi This is a form of survey used with an expert group, with the aim of identifying those issues where the group can reach a consensus, and those where individuals can express disagreement with the group. The survey is administered iteratively to the group, and for each iteration after the first, each participant receives a summary of their own responses and the mean response of the group on the previous iteration, allowing them to change their response to move closer to the norm or to choose to differ.

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 *effect*. The outcomes of a study are based upon measurement of the dependent variable.

descriptive (survey) (See *survey*.)

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 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).

E

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 department's *ethics committee* to ensure that it does not disadvantage any participants in any way.

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

evaluation To judge the quality of; appraise. (Collins English Dictionary)

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. Originating in clinical medicine (EBM) it has been adopted for a number of other domains, including software engineering (Kitchenham 2004, Dybå, Kitchenham & Jørgensen 2005). The normal mechanism for identifying and aggregating research evidence is the *systematic literature review*.

exclusion criteria (In systematic literature reviews.) After performing a search for papers (primary studies) when performing a 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).

exploratory (survey) (See *survey*.)

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

F

field experiment (See *quasi-experiment*.)

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

G

gaussian distribution (Aka *normal distribution*.) A bell-shaped distribution which is the basis for many statistical tests.

H

hypothesis Forms 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. It is then a statistical test to determine the probability that the null hypothesis can or cannot be rejected.

I

inclusion criteria (As used in systematic literature reviews.) After performing a search for papers (primary studies) when performing a 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.

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 process or project attributes.

interpretivism In IS 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.

intrusive (data collection) This involves activities such as filling out forms, attending interviews, answering questions, performing 'think-aloud' etc. and is intrusive as it involves the participant.

L

laboratory experiment Sometimes referred to as a *controlled laboratory experiment* involves the identification of precise relationships between experimental variables by means of a study that takes place in a controlled environment (the 'laboratory') involved 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.

lower quartile The values lying between the 25 percentile and the median.

M

mapping study (Also termed a *scoping review*.) 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 literature review.

maturation This is the process whereby the participants change behaviour between tests, perhaps because of growth (as with children) or because the study has given practice in the skills involved.

mean Often referred to as the *average* and one of the three 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 (an example of this is 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

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 (A.k.a. 50 percentile.) One of the three measures of the *central tendency*. This is the value that separates the upper half of a set of values from the lower half, and we compute it by listing 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 (Petticrew & Roberts 2006). Commonly used within secondary studies such as Systematic Literature Reviews to explore similarities and differences between the outcomes of primary studies, as well as to provide increased statistical power.

mode One of the three measures of the *central tendency*. This is the value that occurs most frequently in a data-set.

N

nominal measurement scale A nominal scale is essentially one that 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*.) Nominal values are often used in regression studies by converting each element in the scale to a dummy binary value. If the nominal scale has n elements, it will be represented by $n - 1$ dummy variables; the nth condition occurs when each of the $n - 1$ dummy variables takes the value zero.

non-intrusive (data collection) Non-intrusive forms are those that do not require any actions on the part of the participant(s). For example, the recording of keystrokes in an experimental interface. (N.B. use of these forms requires that participants are *aware* that they are being employed, in order to avoid ethical problems.)

null hypothesis (See *hypothesis*.)

O

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

observation Non-intrusive forms of data collection (watching, listening,...). This can be 'overt', where those being observed are aware of the presence of the observer, or 'covert' where they are unaware. There is a good discussion of observation in (Oates 2006).

observational scale An observational scale seeks simply to record the actions and outcomes of the study, and there is no attempt to use this to confirm or refute any form of hypothesis. Observational scales are often used to explore an issue and to determine whether more rigourous 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 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.)

outlier A data point that lies beyond the range of values that might be expected for a variable. The expectation may arise from a theoretical consideration (a predicted curve) or from a clustering of empirical values. Note that an outlier may still be a valid data point. Outliers should not be removed from the data set without a good reason. (See also *box plot*.)

P

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.

percentile A value that splits a data set into a stated percentage of values below it. I.e. the 25 percentile or lower quartile is the value such that 25% of data values are less than that value. (See *median*, *lower quartile* and *upper quartile*.)

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 this in an objective manner. It therefore forms the basis for hypothesis-driven research. For a fuller discussion, see (Oates 2006).

power A statistical concept, that determines the probability that a statistical test will correctly reject the null hypothesis, and hence the likelihood that a study can find significant effects. For a discussion of this in the software engineering context, see (Dybå, Kampenes & Sjøberg 2006).

precision This refers to the resolution of a measuring instrument (the smallest mark on its scale). In software terms it is usually the basic counting element (e.g. line of code).

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 broadly similar to a *hypothesis* (and like a hypothesis can be derived from theory). The proposition forms the basis of the case study.

protocol This word 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.
- In the context of *protocol analysis* as a qualitative data analysis technique based upon the use of *think-aloud*, the protocol is a categorisation of possible utterances that is used to analyse the particular sequence produced by a participant while performing a task as well as to strip out irrelevant material.

protocol analysis Widely used in experimental psychology for analysing data related to expert knowledge and for probing into such issues as patterns of information use and strategies employed while performing a task. Requires participants to use *think-aloud*, verbalising their thoughts and ideas while performing a task.

Q

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). Involves participants in answering a series of questions (which may be 'open' or 'closed').

quota A form of population sampling in which the experimenter tries for a balance between different groups within the overall population of interest, for example, male and female participants, Java and C++ programmers, etc.

R

randomised controlled trial (RCT) A form of large-scale controlled experiment using *double blinding* and a random sample from the population of interest. 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 treatement or alternative condition by a random process such as a coin toss or a table of random numbers (Shadish et al. 2002).

ratio scale This is a scale where we have well-defined intervals and also an absolute zero to the scale. Operations are equality, greater than / less than, and ratio (such as 'twice the size'). (See also *measurement scales*.)

reactivity This refers to change in the participant behaviour arising from being tested as part of the study, or from trying the 'help' the experimenter (hypothesis guessing). May also arise because of the influence of the experimenter (such as any bias).

regression The process of fitting data points to a model such as a curve.

research question The research question provides the rationale behind any empirical study, and states in broad terms the issue that the study is intended to investigate (for example, "do structured abstracts make it easier to obtain information about a study simply by reading the abstract?". 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*.

S

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 sample needed to achieve a particular *confidence level* (with a 95% confidence level 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 means by which we select a sample from a sample frame and takes two main forms:

> **probabilistic sampling** An approach whereby we aim to obtain a sample that is a representative cross-section of the sampling frame. Major forms include random, systematic, stratified and cluster sampling.
>
> **non-probabilistic sampling** Such forms may be employed where it is impractical or unnecessary to have a representative sample. Forms include purposive, snowball, self-selection and convenience sampling.
>
> These are discussed in more detail in (Oates 2006).

scatter graph (A.k.a. *Scatter Plot*.) These are often used to demonstrate the relationship between two variables, and the more closely the points cluster around a line, then the closer the relationship between the variables. Conversely, if no pattern can be seen then there is no relationship.

scoping review (See *mapping study*.)

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.

standard deviation The square root of the variance of a set of values. (See *variance*.)

statistical power The ability of a statistical test to reveal a true pattern in the data (Wohlin, Runeson, Host, Ohlsson, Regnell & Wesslen 2000). 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 et al. 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 ("this is more significant than that"). 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:

> **descriptive** The aim is to find out 'what' the attributes of a particular phenomenon are;
>
> **explanatory** The purpose here is to explain 'why' a particular phenomenon occurs;
>
> **exploratory** Here the survey is used to 'scope' a problem to ensure that a later (fuller) study does not omit any relevant issues.

synthesis The process of systematically combining different sources of data (evidence) to answer a research question.

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 and aggregating the data from these. The use of systematic literature reviews in software engineering are discussed in (Kitchenham & Charters 2007) and a more general approach (based upon the social sciences) is provided in (Petticrew & Roberts 2006).

T

tertiary study This from of study effectively performs 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.

think-aloud Used with *protocol analysis*. Participants in a study are trained to 'speak out' their thoughts while performing a task. These verbalisations (often termed *utterances* are recorded and then analysed to identify particular patterns of behaviour.

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 describe 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 studies that reinforce one another in terms of providing evidence, where no one study would be adequately convincing.

U

unbiased See definition of *biased*.

upper quartile The data values lying between the median and the 75 percentile.

utterance (See *think-aloud*.)

V

validity This is concerned with the degree to which we can 'trust' the outcomes of an empirical study, and is usually assessed in terms of four commonly-encountered forms of *threat to validity*. The following definitions are based upon those used 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.

statistical conclusion Relating inferences about the relationship between treatment and outcome variables.

variance A measure of the spread of a set of values. Formally, the mean squared deviation from the mean. (See also *mean* and *standard deviation*.)

W

within-subject (A.k.a. *sequential experiment*.) Refers to one of the possible designs of a laboratory experiment. In this form, participants receive a number of different treatments, with the order in which these are received being randomised. The basic design (two treatments) is an 'A/B–B/A crossover' form whereby some participants receive treatment A and then treatment B, while others receive them in reverse order.

References

Dybå, T., Kampenes, V. B. & Sjøberg, D. (2006), 'A systematic review of statistical power in software engineering experiments', *Information & Software Technology* **48**, 745–755.

Dybå, T., Kitchenham, B. A. & Jørgensen, M. (2005), 'Evidence-Based Software Engineering for Practitioners', *IEEE Software* **22**, 58–65.

Fenton, N. E. & Pfleeger, S. L. (1997), *Software Metrics: A Rigorous & Practical Approach*, PWS Publishing.

Kitchenham, B. (2004), Procedures for undertaking systematic reviews, Technical Report TR/SE-0401, Department of Computer Science, Keele University and National ICT, Australia Ltd. Joint Technical Report.

Kitchenham, B. & Charters, S. (2007), Guidelines for performing Systematic Literature Reviews in Software Engineering, Technical Report EBSE 2007-001, Keele University and Durham University Joint Report.

Oates, B. J. (2006), *Researching Information Systems and Computing*, SAGE.

Petticrew, M. & Roberts, H. (2006), *Systematic Reviews in the Social Sciences: A Practical Guide*, Blackwell Publishing.

Shadish, W., Cook, T. & Campbell, D. (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin Company.

Wohlin, C., Runeson, P., Host, M., Ohlsson, M. C., Regnell, B. & Wesslen, A. (2000), *Experimentation in Software Engineering: An Introduction*, Kluwer.

Yin, R. K. (2003), *Case Study Research: Design and Methods*, SAGE.