taken at the university in computer science may be a poor measure of the subject's experience in a programming language, i.e. has poor construct validity. The number of years of practical use may be a better measure, i.e. has better construct validity.

Threats to external validity concern the ability to generalize experiment results outside the experiment setting. External validity is affected by the experiment design chosen, but also by the objects in the experiment and the subjects chosen. There are three main risks: having wrong participants as subjects, conducting the experiment in the wrong environment and performing it with a timing that affects the results.

A detailed list of threats to the validity is presented in Sect. 8.8. This list can be used as a checklist for an experiment design. In the validity evaluation, each of the items is checked to see if there are any threats. If there are any, they have to be addressed or accepted, since sometimes some threat to validity has to be accepted. It may even be impossible to carry out an experiment without certain threats and hence they have to be accepted and then addressed when interpreting the results. The priority between different types of threats is further discussed in Sect. 8.9.

8.8 Detailed Description of Validity Threats

Below, a list of threats to the validity of experiments is discussed based on Cook and Campbell [37]. All threats are not applicable to all experiments, but this list can be seen as a checklist. The threats are summarized in Table 8.10 and the alternative and limited classification scheme [32] is summarized in Table 8.11.

8.8.1 Conclusion Validity

Threats to the conclusion validity are concerned with issues that affect the ability to draw the correct conclusion about relations between the treatment and the outcome of an experiment.

Low statistical power. The power of a statistical test is the ability of the test to reveal a true pattern in the data. If the power is low, there is a high risk that an erroneous conclusion is drawn, see further Sect. 8.2 or more specifically we are unable to reject an erroneous hypothesis.

Violated assumptions of statistical tests. Certain tests have assumptions on, for example, normally distributed and independent samples. Violating the assumptions may lead to wrong conclusions. Some statistical tests are more robust to violated assumptions than others are, see Chap. 10.

Fishing and the error rate. This threat contains two separate parts. Searching or 'fishing' for a specific result is a threat, since the analyses are no longer independent and the researchers may influence the result by looking for a specific outcome.

Table 8.10 Threats to validity according to Cook and Campbell [37]

Conclusion validity	Internal validity
Low statistical power	History
Violated assumption of statistical tests	Maturation
Fishing and the error rate	Testing
Reliability of measures	Instrumentation
Reliability of treatment implementation	Statistical regression
Random irrelevancies in experimental setting	Selection
Random heterogeneity of subjects	Mortality
	Ambiguity about direction of causal influence
	Interactions with selection
	Diffusion of imitation of treatments
	Compensatory equalization of treatments
	Compensatory rivalry
	Resentful demoralization
Construct validity	External validity
Inadequate preoperational explication of constructs	Interaction of selection and treatment
Mono-operation bias	Interaction of setting and treatment
Mono-method bias	Interaction of history and treatment
Confounding constructs and levels of constructs	
Interaction of different treatments	
Interaction of testing and treatment	
Restricted generalizability across constructs	
Hypothesis guessing	
Evaluation apprehension	
Experimenter expectancies	

Table 8.11 Threats to validity according to Campbell and Stanley [32]

Internal validity	External validity
History	Interaction of selection and treatment
Maturation	Interaction of history and treatment
Testing	Interaction of setting and treatment
Instrumentation	Interaction of different treatments
Statistical regression	
Selection	

The error rate is concerned with the actual significance level. For example, conducting three investigations with a significance level of 0.05 means that the total significance level is $1 - (1 - 0.05)^3$, which equals 0.14. The error rate (i.e. significance level) should thus be adjusted when conducting multiple analyses.

Reliability of measures. The validity of an experiment is highly dependent on the reliability of the measures. This in turn may depend on many different factors, like poor question wording, bad instrumentation or bad instrument layout. The basic principle is that when you measure a phenomenon twice, the outcome shall be the same. For example, lines of code are more reliable than function points since it does not involve human judgement. In other words, objective measures, that can be repeated with the same outcome, are more reliable than subjective measures, see also Chap. 3.

Reliability of treatment implementation. The implementation of the treatment means the application of treatments to subjects. There is a risk that the implementation is not similar between different persons applying the treatment or between different occasions. The implementation should hence be as standard as possible over different subjects and occasions.

Random irrelevancies in experimental setting. Elements outside the experimental setting may disturb the results, such as noise outside the room or a sudden interrupt in the experiment.

Random heterogeneity of subjects. There is always heterogeneity in a study group. If the group is very heterogeneous, there is a risk that the variation due to individual differences is larger than due to the treatment. Choosing more homogeneous groups will on the other hand affect the external validity, see below. For example, an experiment with undergraduate students reduces the heterogeneity, since they have more similar knowledge and background, but also reduces the external validity of the experiment, since the subjects are not selected from a general enough population.

8.8.2 Internal Validity

Threats to internal validity are influences that can affect the independent variable with respect to causality, without the researcher's knowledge. Thus they threat the conclusion about a possible causal relationship between treatment and outcome. The internal validity threats are sometimes sorted into three categories, *single group threats, multiple group threats* and *social threats*.

Single group threats. These threats apply to experiments with single groups. We have no control group to which we do not apply the treatment. Hence, there are problems in determining if the treatment or another factor caused the observed effect.

History. In an experiment, different treatments may be applied to the same object at different times. Then there is a risk that the history affects the experimental results, since the circumstances are not the same on both occasions. For example if one of the experiment occasions is on the first day after a holiday or on a day when a very rare event takes place, and the other occasion is on a normal day.

Maturation. This is the effect of that the subjects react differently as time passes. Examples are when the subjects are affected negatively (tired or bored) during the experiment, or positively (learning) during the course of the experiment.

Testing. If the test is repeated, the subjects may respond differently at different times since they know how the test is conducted. If there is a need for familiarization to the tests, it is important that the results of the test are not fed back to the subject, in order not to support unintended learning.

Instrumentation. This is the effect caused by the artifacts used for experiment execution, such as data collection forms, document to be inspected in an inspection experiment etc. If these are badly designed, the experiment is affected negatively.

Statistical regression. This is a threat when the subjects are classified into experimental groups based on a previous experiment or case study, for example top-ten or bottom-ten. In this case there might be an increase or improvement, even if no treatment is applied at all. For example if the bottom-ten in an experiment are selected as subjects based on a previous experiment, all of them will probably not be among the bottom-ten in the new experiment due to pure random variation. The bottom-ten cannot be worse than remain among the bottom-ten, and hence the only possible change is to the better, relatively the larger population from which they are selected.

Selection. This is the effect of natural variation in human performance. Depending on how the subjects are selected from a larger group, the selection effects can vary. Furthermore, the effect of letting volunteers take part in an experiment may influence the results. Volunteers are generally more motivated and suited for a new task than the whole population. Hence the selected group is not representative for the whole population.

Mortality. This effect is due to the different kinds of persons who drop out from the experiment. It is important to characterize the dropouts in order to check if they are representative of the total sample. If subjects of a specific category drop out, for example, all the senior reviewers in an inspection experiment, the validity of the experiment is highly affected.

Ambiguity about direction of causal influence. This is the question of whether A causes B, B causes A or even X causes A and B. An example is if a correlation between program complexity and error rate is observed. The question is if high program complexity causes high error rate, or vice versa, or if high complexity of the problem to be solved causes both.

Most of the threats to internal validity can be addressed through the experiment design. For example, by introducing a control group many of the internal threats can be controlled. On the other hand, multiple group threats are introduced instead.

Multiple groups threats. In a multiple groups experiment, different groups are studied. The threat to such studies is that the control group and the selected experiment groups may be affected differently by the single group threats as defined above. Thus there are interactions with the selection.

Interactions with selection. The interactions with selection are due to different behavior in different groups. For example, the selection-maturation interaction means that different groups mature at different speed, for example if two groups apply one new method each. If one group learns its new method faster than the other, due to its learning ability, does, the selected groups mature differently. Selection-history means that different groups are affected by history differently, etc.

Social threats to internal validity. These threats are applicable to single group and multiple group experiments. Examples are given below from an inspection experiment where a new method (perspective-based reading) is compared to an old one (checklist-based reading).

Diffusion or imitation of treatments. This effect occurs when a control group learns about the treatment from the group in the experiment study or they try to imitate the behavior of the group in the study. For example, if a control group uses a checklist-based inspection method and the experiment group uses perspective-based methods, the former group may hear about the perspective-based method and perform their inspections influenced by their own perspective. The latter may be the case if the reviewer is an expert in a certain area.

Compensatory equalization of treatments. If a control group is given compensation for being a control group, as a substitute for that they do not get treatments; this may affect the outcome of the experiment. If the control group is taught another new method as a compensation for not being taught the perspective-based method, their performance may be affected by that method.

Compensatory rivalry. A subject receiving less desirable treatments may, as the natural underdog, be motivated to reduce or reverse the expected outcome of the experiment. The group using the traditional method may do their very best to show that the old method is competitive.

Resentful demoralization. This is the opposite of the previous threat. A subject receiving less desirable treatments may give up and not perform as good as it generally does. The group using the traditional method is not motivated to do a good job, while learning something new inspires the group using the new method.

8.8.3 Construct Validity

Construct validity concerns generalizing the result of the experiment to the concept or theory behind the experiment. Some threats relate to the design of the experiment, others to social factors.

Design threats. The design threats to construct validity cover issues that are related to the design of the experiment and its ability to reflect the construct to be studied.

Inadequate preoperational explication of constructs. This threat, despite its extensive title, is rather simple. It means that the constructs are not sufficiently defined, before they are translated into measures or treatments. The theory is not clear enough, and hence the experiment cannot be sufficiently clear. For example, if two inspection methods are compared and it is not clearly enough stated what being 'better' means. Does it mean to find most faults, most faults per hour, or most serious faults?

Mono-operation bias. If the experiment includes a single independent variable, case, subject or treatment, the experiment may under-represent the construct and thus not give the full picture of the theory. For example, if an inspection experiment is conducted with a single document as object, the cause construct is under-represented.

Mono-method bias. Using a single type of measures or observations involves a risk that if this measure or observation gives a measurement bias, then the experiment will be misleading. By involving different types of measures and observations they can be cross-checked against each other. For example, if the number of faults found is measured in an inspection experiment, where fault classification is based on subjective judgement, the relations cannot be sufficiently explained. The experimenter may bias the measures.

Confounding constructs and levels of constructs. In some relations it is not primarily the presence or absence of a construct, but the level of the construct which is of importance to the outcome. The effect of the presence of the construct is confounded with the effect of the level of the construct. For example, the presence or absence of prior knowledge in a programming language may not explain the causes in an experiment, but the difference may depend on if the subjects have 1, 3 or 5 years of experience with the current language.

Interaction of different treatments. If the subject is involved in more than one study, treatments from the different studies may interact. Then you cannot conclude whether the effect is due to either of the treatments or of a combination of treatments.

Interaction of testing and treatment. The testing itself, i.e. the application of treatments, may make the subjects more sensitive or receptive to the treatment. Then the testing is a part of the treatment. For example, if the testing involves measuring the number of errors made in coding, then the subjects will be more aware of their errors made, and thus try to reduce them.

Restricted generalizability across constructs. The treatment may affect the studied construct positively, but unintenionally affect other constructs negatively. This threat makes the result hard to generalize into other potential outcomes. For example, a comparative study concludes that improved productivity is achieved with a new method. On the other hand, it can be observed that it reduces the maintainability, which is an unintended side effect. If the maintainability is not measured or observed, there is a risk that conclusions are drawn based on the productivity attribute, ignoring the maintainability.

Social threats to construct validity. These threats are concerned with issues related to behavior of the subjects and the experimenters. They may, based on the fact that they are part of an experiment, act differently than they do otherwise, which gives false results from the experiment.

Hypothesis guessing. When people take part in an experiment they might try to figure out what the purpose and intended result of the experiment is. Then they are likely to base their behavior on their guesses about the hypotheses, either positively or negatively, depending on their attitude to the anticipated hypothesis.

Evaluation apprehension. Some people are afraid of being evaluated. A form of human tendency is to try to look better when being evaluated which is confounded to the outcome of the experiment. For example, if different estimation models are compared, people may not report their true deviations between estimate and outcome, but some false but 'better' values.

Experimenter expectancies. The experimenters can bias the results of a study both consciously and unconsciously based on what they expect from the experiment. The threat can be reduced by involving different people which have no or different expectations to the experiment. For example, questions can be raised in different ways in order to give the answers you want.

8.8.4 External Validity

Threats to external validity are conditions that limit our ability to generalize the results of our experiment to industrial practice. There are three types of interactions with the treatment: people, place and time:

Interaction of selection and treatment. This is an effect of having a subject population, not representative of the population we want to generalize to, i.e. the wrong people participate in the experiment. An example of this threat is to select only programmers in an inspection experiment when programmers as well as testers and system engineers generally take part in the inspections.

Interaction of setting and treatment. This is the effect of not having the experimental setting or material representative of, for example, industrial practice. An example is using old-fashioned tools in an experiment when up-to-date tools are common in industry. Another example is conducting experiment on toy problems. This means wrong 'place' or environment.

Interaction of history and treatment. This is the effect of that the experiment is conducted on a special time or day which affects the results. If, for example, a questionnaire is conducted on safety-critical systems a few days after a big software-related crash, people tend to answer differently than a few days before, or some weeks or months later.

The threats to external validity are reduced by making the experimental environment as realistic as possible. On the other hand, reality is not homogenous. Most important is to characterize and report the characteristics of the environment, such as staff experience, tools, methods in order to evaluate the applicability in a specific context.

8.9 Priority Among Types of Validity Threats

There is a conflict between some of the types of validity threats. The four types considered are internal validity, external validity, conclusion validity and construct validity. When increasing one type, another type may decrease. Prioritizing among the validity types is hence an optimization problem, given a certain purpose of the experiment.

For example, using undergraduate students in an inspection experiment will probably enable larger study groups, reduce heterogeneity within the group and give reliable treatment implementation. This results in high conclusion validity, while the external validity is reduced, since the selection is not representative if we want to generalize the results to the software industry.

Another example is to have the subjects measure several factors by filling out schemes in order to make sure that the treatments and outcomes really represent the constructs under study. This action will increase the construct validity, but there is a risk that the conclusion validity is reduced since more, tedious measurements have a tendency to reduce the reliability of the measures.

In different experiments, different types of validity can be prioritized differently, depending on the purpose of the experiment. Cook and Campbell [37] propose the following priorities for theory testing and applied research:

Theory testing. In theory testing, it is most important to show that there is a casual relationship (internal validity) and that the variables in the experiment represent the constructs of the theory (construct validity). Adding to the experiment size can generally solve the issues of statistical significance (conclusion validity). Theories are seldom related to specific settings, population or times to which the results should be generalized. Hence there is little need for external validity issues. The priorities for experiments in theory testing are in decreasing order: internal, construct, conclusion and external.

Applied research. In applied research, which is the target area for most of the software engineering experiments, the priorities are different. Again, the relationships under study are of highest priority (internal validity) since the key goal of the experiment is to study relationships between causes and effects. In applied research, the generalization – from the context in which the experiment is conducted to a wider context – is of high priority (external validity). For a researcher, it is not so interesting to show a particular result for company X, but rather that the result is valid for companies of a particular size or application domain. Third, the applied researcher is relatively less interested in which of the components in a complex treatment that really causes the effect (construct validity). For example, in a reading experiment, it is not so interesting to know if it is the increased understanding in general by the reviewer, or it is the specific reading procedure that helps the readers to find more faults. The main interest is in the effect itself. Finally, in practical settings it is hard to get sufficient size of data sets, hence the statistical conclusions may be drawn with less significance (conclusion validity).

The priorities for experiments in applied research are in decreasing order: internal, external, construct and conclusions.

It can be concluded that the threats to validity of experimental results are important to evaluate and balance during planning of an experiment. Depending on the purpose of the experiment, different validity types are given different priority. The threats to an experiment are also closely related to the practical importance of the results. We may, for example, be able to show a statistical significance, but the difference is of no practical importance. This issue is further elaborated in Sect. 10.3.14.

8.10 Example Experiment

This description is a continuation of the example introduced in Sect. 7.2. The input to the planning phase is the goal definition. Some of the issues related to planning have partially been addressed in the way the goal definition is formulated in the example. It is already stated that students will be the subjects and the text also indicates that the experiment will involve more than one requirements document. Planning is a key activity when conducting an experiment. A mistake in the planning step may affect the whole outcome of the experiment. The planning step includes seven activities as shown in Fig. 8.1.

Context selection. The type of context is in many cases at least partially decided by the way the goal definition is formulated. It is implicitly stated that the experiment will be run off-line, although it could potentially be part of a student project, which would have meant on-line although not as part of an industrial development project. The experiment will be run with a mixture of M.Sc. and Ph.D. students.

An off-line experiment with students implies that it may be difficult to have time to inspect a requirements document for a fully-fledged real system. In many cases, experiments of this type have to resort to a requirements document with limited features. In this specific case, two requirements documents from a lab package (material available on-line for replication purposes) will be used. The choice to use two requirements documents has some implications when it comes to the choice of design type, which we will come back to. The requirements documents have some limitations when it comes to features and hence they are to some extent to be considered as ‘toy’ requirements documents.

The experiment can be considered as general in the sense that the objective is to compare two reading techniques in general (from a research perspective), and it is not about comparing an existing reading technique in a company with a new alternative reading technique. The latter would have made the experiment specific for the situation at the company. In both these cases, there are some issues to take into account to ensure a fair comparison.

In the general research case, it is important that the comparison is fair in the sense that the support for the two techniques being investigated is comparable. It is of