Research Methods in Computer Science

Matúš Sulír

Table of Contents

Preface

This is a textbook for the Research Methods in Computer Science and Cybersecurity course at the Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice.

1 Overview

In the first chapter, we will try to distinguish what is (and what is not) research. We will continue with a bird's-eye view of the whole research process.

1.1 Research and Science

Let us start with a few definitions of the word *research*:

> "a systematic investigation to find answers to a problem"
>
> – Burns (2000)

> "creative and systematic work undertaken in order to increase the stock of knowledge"
>
> – OECD (2015)

> "a detailed study of a subject, especially in order to discover (new) information or reach a (new) understanding"
>
> – *Cambridge Dictionary*

Therefore, research is an activity that:

- studies something worth studying (relevance),
- uses a systematic approach (soundness),
- and adds new information to the existing body of knowledge (novelty).

A related term, *science*, is usually defined more narrowly:

> "any system of knowledge that is concerned with the physical world and its phenomena and that entails unbiased observations and systematic experimentation"
>
> – *Encyclopedia Britannica*

> "the systematic study of the structure and behavior of the physical and natural world through observation, experimentation, and the testing of theories against the evidence obtained"
>
> – *Oxford Dictionary*

We can see that science concerns a natural, externally observable world. Parts of this world are systematically observed or manipulated, and, based on the results, new knowledge is produced. Such an approach is called *empirical*, from Greek empeiría (experience), where evidence is formed through senses, using measurement and experimentation.

According to the aforementioned definitions, mathematics is not a science since mathematicians do not observe the external world, and they manipulate only abstract entities they devised. Humanities, such as history, literature, or art, are usually not considered sciences either.

1.2 Computer Science

A question now arises: Is computer science a science? According to Denning (2005), yes – as it meets every criterion. We can devise testable hypotheses, observe the external behavior of computers or the people using them, and construct theories based on these observations.

Abelson (1986) thinks the opposite, as he noted in the opening lecture of the Structure and Interpretation of Computer Programs course:

> "It might be engineering or it might be art, but we'll actually see that computer so-called science actually has a lot in common with magic, and we'll see that in this course. So it's not a science. It's also not really very much about computers. And it's not about computers in the same sense that physics is not really about particle accelerators, and biology is not really about microscopes and Petri dishes."

Note, however, that this view originates in the era when empirical studies, such as experiments with human participants or thorough evaluations using computer benchmarks, were rarer than today. Many research papers of that time described a new idea that the researcher got, along with some implementation details, neglecting evaluation. As one of the voices for a change, a paper advocating evidence-based software engineering as an analogy to evidence-based medicine (B. A. Kitchenham et al., 2004) was published at the beginning of this century.

1.3 Computer Science Subfields

It is difficult to talk about research methods in computer science in general. Computer science is a diverse field, encompassing a large variety of subfields (e.g., embedded systems, cybersecurity, high-performance computing, operating systems, databases, or virtual reality). There have been multiple attempts to divide computer science research into categories, including the Association for Computing Machinery (ACM) Special Interest Groups (SIGs), the ACM Computing Classification System (CCS), and the arXiv category taxonomy (expand the "Computer Science" item). Although the areas certainly overlap to some degree, a majority

of researchers specialize in a few narrowly defined topics pertaining to one computer science subfield.

Each subfield has its own typical strategies of advancing knowledge. On one side, there is theoretical computer science. Here, the researcher often starts by defining a theorem, which is split into smaller lemmas. Each lemma is mathematically proved, which finally results in the proof of the main theorem. The research process is thus similar to mathematics as the researchers operate only with abstract notations. However, the proved properties of algorithms, e.g., time complexity, are sometimes empirically verified by executing the program on a computer and measuring the necessary properties.

Human-computer interaction, on the other hand, is people-centric. Traditionally, human participants are observed interacting with a specific hardware or software user interface, and selected quantities are measured.

Some subfields themselves use plenty of different research methods. For example, software engineering papers often describe a new approach designed by the authors, which is subsequently evaluated empirically using simulation, case studies, surveys and interviews, experiments, etc.

Other areas lean toward a commonly used research method. In machine learning, for example, standardized measures such as accuracy or F_1 score are often computed on a popular benchmark for the given problem. This facilitates comparison between multiple approaches but can easily divert attention from problems arising in real-world usage contexts.

The main research area of the author of this text is software engineering, occasionally intersected with human-computer interaction. Despite trying to be impartial, this text may sometimes be a bit biased toward these two fields.

1.4 Interdisciplinarity

Just as the border between computer science subfields is blurry, so is often the border of computer science as a whole. Many research projects span multiple branches of science. The knowledge from multiple fields is combined, and new knowledge, related to one or more of these fields, is produced. This is called an *interdisciplinary* approach.

Interdisciplinary research can bring new insights to real-world problems that would be otherwise difficult to solve. On the other hand, the problem of research shallowness often particularly affects interdisciplinary research. Consider an application of an off-the-shelf machine learning algorithm to weather prediction. Without an expert in meteorology being a part of the research team, there is a risk of misinterpretation of the results and the neglection of methodological or practical issues. Furthermore, while it may be a research contribution to the field of meteorology, we could hardly consider it a contribution to the field of computer science.

1.5 Research Process Overview

Research is an unpredictable activity to a high degree. Despite this, in Figure 1.1, there is an approximation of the journey a computer science researcher can typically take to finish one research project. Be aware that we could add an arrow from almost any activity to any previous activity.

Figure 1.1: An overview of the research process

At the beginning, we must have an idea of what the project will aim to accomplish and why this is relevant for other researchers or the general public. Then we start gathering and studying relevant literature. Throughout the rest of the research process, we will get a better understanding of the topic, so we should return to this step regularly.

Although some researchers leave the writing of the paper as one of the last steps, it is beneficial to start writing as soon as we have a general idea and a few relevant research papers. Putting your ideas into writing helps you clarify your thinking, find inconsistencies soon, and prevent forgetting. Writing should ideally occur after each of the next steps, until sending the paper for review.

The following steps are variable and differ between projects. It is possible to create a software or hardware artifact based on a new approach we designed and described in the paper, e.g., an implementation of an algorithm, a new editor plugin, or a novel pointing device. We suppose the designed approach (and thus the artifact that implements it) has some properties (it is more efficient, more usable, etc.), based on which we pose research questions or hypotheses. Alternatively, we can skip the artifact creation and do purely empirical research instead, by asking research questions or posing hypotheses about currently existing artifacts, people, or processes.

After carefully planning the research methods corresponding to the research questions, we collect and process the data and answer these questions based on the data.

It is possible that the purpose of selected research questions was to find shortcomings of existing approaches or our newly-designed artifacts. If this is the case, we can continue by designing a better artifact and repeating a part of the process.

After the paper is finished, we send it for peer review to an appropriate journal or conference, where about three experts will assess its suitability for publication. If the paper is not rejected, we incorporate the reviewers' comments into it, and, depending on the type of the publication, send it again for review and repeat until it is accepted.

The paper is then published in the given journal or conference proceedings. This is usually considered the final step, but we should remember to spread the knowledge further, e.g., summarize it in a blog post or archive it on our personal website if it is not freely accessible from the publisher.

In the following chapters, we will discuss the individual steps of the research process, starting with the formulation of ideas and searching for relevant literature.

Exercises

1. Suppose you invented a new algorithm to find the shortest path between two vertices in a graph. Is this research? Is it science? Why or why not?
2. Consider this extreme example: You created a program appending the letter "a" to every 15th occurrence of the letter "z". The letters and the occurrence number are not configurable, and your algorithm uses hacks relying on these constants. Based on your measurements, the program is faster, more memory efficient, and more user-friendly than any other program doing it. Should this be called research?

3. According to you, which of the following documents describe research and why?
    a. Fatal Abstraction (Steimann, 2018)
    b. Business Calendar
    c. Fluorescence SpectraViewer
    d. Bringing Rich Experiences to Memory-Constrained TV Devices
    e. DDD — A Free Graphical Front-End for UNIX Debuggers (Zeller & Lütkehaus, 1996)
    f. On the Dichotomy of Debugging Behavior Among Programmers (Beller et al., 2018)
4. Compare a few papers published at the leading software engineering conference, ICSE, in 1982 (Boehm et al., 1982; Futatsugi & Okada, 1982; Nakamoto et al., 1982) and 2024 (Carvalho et al., 2024; Choudhuri et al., 2024; Huang et al., 2024). How do they differ?
5. What computer science subfield(s) do you consider your specialty? For inspiration, you can look at the illustrated Map of Computer Science.
6. Find an example of an interdisciplinary research paper with the experts in two or more domains as co-authors and providing, in your opinion, a valuable contribution to some research field(s). Find another paper where you think the expertise from one field was lacking.
7. Which parts of the research process do you consider the most difficult and why?

2 Forming Ideas and Finding Literature

When starting a research project, we need to define what we would like to accomplish: from a general idea to a more specific research goal. This is tightly interconnected with searching and reading works related to our topic. Recall from Figure 1.1 that there is a loop between the idea formulation and studying relevant literature. None of these two steps is definitely the first one, as they influence each other.

2.1 Coming Up with an Idea

How do researchers come up with a new research idea? The process is usually a chaotic mixture of events, but there are some general guidelines.

2.1.1 Gaps in Literature

The most obvious way is searching the literature for gaps in knowledge. This seems simple, but the reality is more complicated. First, to determine which piece of knowledge is nonexistent but worth studying, one should have a pretty good understanding of the topic obtained by reading tens or hundreds of papers on it.

Suppose we are already knowledgeable on the topic. How specifically do we derive a new idea from existing papers?

Almost all research articles contain a section called Future Work at the end, or at least a paragraph describing what future ideas the authors envision. Sadly, these ideas may be too general, unrealistically difficult, too incremental, or already done by someone in the meantime. If the paper was published relatively recently, the original authors may as well be working on some of those ideas right now.

A particularly interesting idea is to tabulate certain properties of existing approaches and find missing cells. Consider a hypothetical example about automated source code documentation generators. After reviewing existing documentation approaches, we would find they can be characterized by two *dimensions*: input (with *attributes* “method” and “class”) and output (with attributes “sentence” and “paragraph”), as shown in Table 2.1.

Table 2.1: An example of the dimensions of documentation approaches

		Input	
		method	**class**
Output	**sentence**	Foo et al.	Bar et al.
	paragraph	Baz et al.	–

We can see that other researchers have already devised documentation generators that take source code methods or classes as input and produce sentences. Documentation generators producing paragraphs, however, take only methods as input. We could thus try to devise a new documentation approach generating whole paragraphs from classes, provided it was useful and feasible.

Instead of finding a nonexistent combination of attributes, we can even devise completely new attributes or dimensions. In our example, we could explore the possibility of using images as the output of a documentation generator. Other options include generalizing an existing approach, making it more specific, or combining multiple approaches in a nontrivial manner.

If we aim to perform an empirical study instead of designing a new approach, we can apply a ready-to-use framework with a list of dimensions. One of these frameworks, PICOC (B. Kitchenham & Charters, 2007), stands for Population (for example, testers), Intervention (e.g., multi-font syntax highlighting), Comparison (traditional syntax highlighting), Outcome (the number of bugs), and Context (large-scale projects). In this example, we could possibly change the "population" dimension from testers to frontend developers, i.e., determine whether frontend developers produce fewer bugs using multi-font syntax highlighting in a large-scale project.

2.1.2 Personal Experience

Many ideas come from personal experiences, frustrations, and anecdotal evidence of researchers. For instance, when trying to compile ten different open-source Java projects, five of them might fail. We might start wondering if we were just unlucky or whether this is a general phenomenon. To confirm or disprove this, we may design and execute an empirical study with thousands of such projects.

As researchers often also teach, many papers touch on pedagogical topics. Care should be taken to make such studies really systematic, and to critically admit whether the authors are experts on pedagogy before attempting to publish such a paper.

Many times, researchers in computer science represent potential users of approaches they design or study: they develop software, write documentation in markup languages, use human-computer interaction devices, or use video editing software. As such, they can think critically about the current state of the given domain. For instance, when filing a bug in an issue tracking

system, could the title be generated automatically using language processing methods? Note, however, that research tends to be ahead of practice in many fields. If popular systems do not implement a given feature, it definitely does not mean there is no research in the area.

Researchers often extend their own work. A useful pattern is the iteration between empirical studies and artifact creation. We start with an empirical study which finds problems in the current state of the art. Based on these ideas, we design a new approach. It is then empirically evaluated to confirm if the previously present problems were eliminated. During this evaluation, we usually find new problems. With this knowledge, we design a new, improved version of the approach and continue in a loop.

2.1.3 Other People

Talking to other people is another source of inspiration for research ideas. For students, one of the most obvious choices is a thesis advisor. In bachelor's and master's studies, the advisors usually already have an idea prepared for the students, ready to be fine-tuned by communication. On a PhD level, the topic provided by an advisor tends to be much more general, and it is expected that the student will actively suggest a more specific idea.

Talking to colleagues, classmates, and professionals from industry about the current problems they face often brings a multitude of great ideas. It is, however, important not to solve only the symptoms of the immediate, obvious issues. Instead, researchers should strive for conceptual improvements, changing the way things work in general. Design thinking is a set of procedures that can help achieve this. One of its components, the "five whys" technique, tries to find the root cause of a problem by asking "why" iteratively. For example, instead of fixing a bug in a code with an automatically generated patch, we might ask: Why was this bug present in the code? After finding it occurred due to a null pointer exception, we ask again: Why did the exception occur? After a few more whys, we might finally find that programmers need a succinct, runtime-safe way to express objects with default behavior in the source code.

Attending academic conferences is another interesting way to get inspiration by seeing a quick overview of what the other researchers in the area are currently working on and informally discussing with peers from various parts of the world.

2.2 Exploring the Field

The techniques for finding relevant literature depend on whether we would like to get an overview of a broad area or find papers related to a specific idea.

For aspiring researchers, such as master's students thinking about their potential thesis topics or first-year PhD students, it is beneficial to see the big picture of their field. This allows them to get an intuition of what was already done years or decades ago, what the current hot topics are, and how the papers in their area typically look.

Research in computer science is typically published as a paper either in a scientific journal or conference proceedings. Therefore, the first step is to prepare a list of the best journals and conferences in the given field. For journals, there exist catalogs filtrable by sub-categories, such as Journal Citation Reports (accessible only from the university network), Scimago Journal Rank, and Research.com journal rankings. For conferences, there is CORE ranking and Research.com conference ranking. The list of top publication venues at Google Scholar mixes journals and conferences together.

For each journal we are interested in, we should find its official homepage, find a few latest issues, scroll through the list of paper titles, and download a few of them to get a feeling of what constitutes top-notch research in the given area.

For conferences, we should also open their official homepages. Most conferences are annual, so we focus on the last two or three years. Top-tier conferences tend to be large and contain multiple tracks, which consist of sessions. It is advantageous to open the main research track's page and write down the names of its sessions (excluding lunches and coffee breaks, of course). Suppose we would like to explore the field of software engineering. At the ICSE 2024 website, we click Program / ICSE Program, expand the "Filter Program" section, and select Tracks / ICSE Research Track. Then we can see the list of days, each containing multiple sessions, in our case: AI & Security 1, Evolution & AI, Testing 1, Analysis 1, Human and Social 1, Generative AI studies, and many others. Constructing such a list helps us understand the current trends in the given field. For the sessions we are interested in, we can also inspect the list of the corresponding papers.

2.3 Getting an Overview of the Topic

If we narrowed our interest to a more specific topic, such as "time-traveling debugging" or "plant identification using computer vision", we may benefit from reading secondary studies, i.e., articles summarizing existing papers on a given topic. The easiest way to find them is to append the keywords "review", "survey", "systematic literature review", and "systematic mapping study" to the name of our topic and enter them into Google Scholar. Trying synonyms of words in the topic increases the chances of finding the relevant studies. We can also make the topic more specific or general as necessary.

For instance, if we are interested in declarative debugging, we may try:

- "declarative debugging" review
- declarative debugging survey
- algorithmic debugging "systematic mapping study"

Note that some terms are not generally synonyms: *algorithmic* and *declarative* are different terms, but algorithmic and declarative debugging is approximately the same topic. It is thus important to get acquainted with the concepts used by other researchers.

2.4 Searching for Related Papers

Suppose we already have a relatively specific research idea, e.g., "time-traveling debugging using a graphical processor" or "an experimental comparison of cybersecurity education using a traditional and flipped classroom". How do we determine if someone has already published a similar paper and find related papers that we can use for inspiration or for comparison in our Related Work section?

2.4.1 Academic Search Engines

We may use traditional web search engines such as Google or Bing, which return also research papers among other results. However, specialized academic search engines aim to limit the results only to academic literature and offer useful tools such as forward reference lists, which speeds up the process. We will mention some of the more popular ones.

Google Scholar has high document coverage, which is its main benefit and drawback at the same time. Practically all academic documents are indexed – not only journal and conference papers, but also theses, books, and technical reports. It does not exclude low-quality publications, and occasionally includes also files such as manuals and company whitepapers. No subscription is required.

Semantic Scholar is a tool similar to Google Scholar, with a slightly lower document coverage. The usage is free.

Scopus, on the other hand, is a paid service, so it is accessible only from the subscribed universities' networks or after registration with the university e-mail. Coverage is limited to papers passing certain quality criteria. It offers two features indispensable for systematic literature reviews: advanced search based on tens of properties combinable using a custom query language; and export of all search results into a structured file. Sorting the results by relevance is not on par with the other engines, though.

Web of Science is another paid service. It boasts of strict quality criteria for inclusion. Despite its popularity in many fields, its usefulness for computer science research is severely limited. Waiting times for the indexing of new journals and conference proceedings are as long as a few years. Many conferences are not indexed at all.

OpenAlex is a free alternative website, offering export of the results and an API.

Searching for related papers using AI chatbots usefully supplements academic search engines. However, these tools often do not have full-text access to paywalled articles, so using them as a sole source does not suffice. Care must be also taken to verify if the given results are actually relevant to our prompt.

2.4.2 Searching Process

To find as many related works as possible, it is advised to use a high-coverage academic search engine, such as Google Scholar, with a systematic approach. The general procedure is as follows:

1. Enter a search phrase describing your idea succinctly yet precisely.
2. Read the titles of the documents on the displayed results page.
3. For each document that seems relevant based on title, read its abstract. If the abstract seems relevant, inspect the full text quickly: skim the Introduction, Conclusion, figures, and tables. Search the full text for clues – e.g., if finding an experiment with humans, search the terms such as "human", "subject", "participant", or "enrolled". If the paper is still relevant, download it and save it to your reference manager.
4. Go to the next pages of results until you keep seeing relevant papers, repeating steps 2–3.
5. For each saved paper, look at the backward references. These are the works listed in its References section and thus older than the given article. If any of them seems relevant, perform step 3 on it.
6. For each saved paper, list its forward references, i.e., newer papers that cite it. Google Scholar names this feature "Cited by …". Find and save relevant papers in this list.
7. Repeat steps 5 and 6 on all articles saved in the reference manager but not yet processed. Stop when you seem to be approaching infinite recursion, i.e., the documents you see start repeating too much.
8. Read the full text of the saved papers that seem the most relevant.
9. Repeat steps 1–8 with alternative search terms: synonyms or other aspects of the problem. When forming the search terms, use the knowledge you acquired so far.

Obviously, searching for related works is a long-term process. After performing a five-minute search, we should not claim a paper describing ideas similar to our own does not yet exist.

Tip: We can save time by configuring search keywords for academic search engines in a browser (Firefox, Chrome). For instance, after assigning the keyword "s" to Google Scholar, we can simply type "s motion tracking" into the address bar.

2.5 Accessing Papers behind a Paywall

Some papers are inaccessible for general public without paying. As we will explain later in this book, paying for papers as individuals is not reasonable. Instead, we should apply the following tips.

First, on the publisher's websites, there are always the final versions available, so it is a preferred way if we have access to the given paper. Many universities have pre-paid access to digital libraries such as IEEE Xplore, the ACM Digital Library (this one will be free from 2026), Elsevier ScienceDirect, Springer Link, and Wiley Online Library.

If we are outside the university network, we can try a VPN to remotely access the given resource. An alternative is to register on the given site with a university e-mail, or apply for the CVTI remote access available in Slovakia.

If the publisher's version is inaccessible, we should search for the authors' archived version. For example, on Google Scholar, we can click the "[PDF]" or "All versions" buttons. There are also browser plugins that automate this, most notably Unpaywall.

As a last resort, we can try e-mailing the corresponding author of the paper.

2.6 Reference Management

A reference manager is a program designed to add, edit and view a database of bibliographic citations to research-related documents, along with associated metadata, such as notes, tags, or full text links.

2.6.1 Advantages

Using a reference manager makes a huge difference.

First, we build a database of all papers relevant to us for some reason. Most reference manager have some tagging or categorization capability, so papers related to various projects or topics can be distinguished easily.

Second, papers already read can be marked as such. This may seem useless at first, but after reading at least tens of papers, forgetting is easy.

Third, we can, and absolutely should, make notes about papers. This is related to the previous point – having a paper marked as "already read" is fine, but having a short comment summarizing the main points of the paper is even more useful. Instead of objectively summarizing the facts, we should try to focus on exactly how the paper is relevant to us: How could we use the results? What hinders building upon this paper? Is there any research method used that we could use as inspiration?

Finally, with a reference manager, we can cite effortlessly. Reference managers with BibTeX integration offer the possibility of having one central `.bib` file on a computer, which can be included in each paper via `\bibliography{path/to/file}` in LaTeX. Citing is then as simple as pressing a keyboard shortcut and pasting `\cite{citationKey}` into the source code. Unfortunately, during collaboration this does not work as easily since everyone has a separate BibTeX file.

2.6.2 Software

There are many reference managers available. We will mention a few popular ones.

JabRef is a Java-based desktop application. BibTeX is its native storage format, so all information we edit is actually saved in this structured text file. It offers rich customization options. For example, we can set up automated citation key generation for `bib` file entries, based on a pattern. Adding custom fields with user-defined names to entries is also possible. An older version with a classical look and feel can be downloaded too.

Zotero offers, among other features, PDF file annotation and simple citation import from websites via web browser plugins. Although it uses an opaque data storage format, the database can be synchronized with a `bib` file via the Better BibTeX plugin.

KBibTeX is an alternative suitable mainly for Linux distributions with KDE. For macOS, there is also BibDesk.

As a minimalistic alternative, using a spreadsheet or a structured text file is better than nothing.

2.6.3 Citation Sources

Citation entries for the same article from various websites vary in quality. Although it is possible to download a BibTeX entry from Google Scholar by clicking "Cite" and then "BibTeX" under the given document, such a citation often misses necessary fields, such as the issue number of a journal or a Digital Object Identifier (DOI). Downloading citation information directly from the publisher's website usually provides all available information. Alternatively, we can use sites such as the ACM Guide to Computing Literature, which contain citation records from multiple publishers.

If the downloaded BibTeX record lacks some necessary information, it is sometimes necessary to correct it manually.

2.7 Reading Papers

Now, let us consider an important motivational point: Can someone be a good film director without watching a lot of movies first? How would such a director even know what is a good or bad movie? Similarly, can we be a good researcher without reading a lot of papers first?

However, since research papers are usually long and dense pieces, we should not read them like news articles. Instead, we should adopt the three-pass approach (Keshav, 2007):

The first reading pass should take at most 5-10 minutes:

1. Read the title, abstract, and introduction quickly.
2. Skim headings, tables, figures.
3. Briefly read the conclusion.
4. Mark important references.

If the paper is relevant, we can continue with the second pass. We try to read the rest of the paper from start to end in about one hour (of course, this depends on the length), ignoring not immediately important content and information requiring too much time to comprehend: complicated equations not essential to the topic, mathematical proofs, and low-level details including the description of a tool implementation, hardware specification in benchmarks, etc.

If the paper is really important, e.g., we plan to extend it, we consider applying the described methodology, or we deem it one of the three most relevant papers for our upcoming research projects, the third pass is recommended. During it, we can imagine being the author writing this paper. We try to comprehend all details and resolve ambiguities: How exactly would I perform this (unclear) step in the method? How was this result computed? What are the shortcomings? How could I do it better? This pass takes multiple hours.

Sometimes, the second and third passes blend together, particularly if we know beforehand that the third pass will be necessary.

2.8 Systematic Literature Studies

A systematic literature study (often called simply a systematic review) is a research study that aims to collect all works relevant to given questions using a rigorous protocol, analyze them, and synthesize results characterizing them. We recognize the two most common types of systematic literature studies:

- *Systematic literature reviews* (SLRs) in their narrow sense analyze the collected articles deeply, often producing quantitative results by aggregating the results of the original (primary) works. They answer specific research questions, such as: What portion of bugs in Android applications can be resolved using automated tools? Which parsing algorithm is the most efficient for context-free languages?
- *Systematic mapping studies* (SMSs) identify sub-topics, categories, trends, or research gaps. They answer broader, less focused research questions, e.g.: How can automated bug resolution techniques be categorized? Which parsing algorithms exist for context-free languages and what are their shortcomings?

Multiple guidelines exist for conducting systematic reviews, including the ones by Carrera-Rivera et al. (2022), B. Kitchenham & Charters (2007), and Wohlin et al. (2024). The best way to get an intuition about how a systematic review is performed is to read specific examples of high-quality papers of this kind, e.g., by Hall et al. (2012), Shahin et al. (2014), Inayat et al.

(2015), or Arvanitou et al. (2021) in the area of software engineering or many other papers in the respective fields.

To conduct a systematic review, it is first necessary to pose *research questions* that we would like to answer.

Next, we specify a *search strategy*, which is a process of the retrieval of all works potentially relevant to our research questions. There are three major search methods: manual search, automated search, and snowballing.

Manual search is performed on the lists of all papers published in relevant journals and conference proceedings. Relevant journals and conferences are selected by personal knowledge of the research field or from catalogs. We then manually scan the lists of all published papers in the given year range of these journals and conferences, while selecting only papers relevant for the study based on titles, abstracts, or full texts if necessary.

Automated search is done by entering keywords into the search boxes of digital libraries and/or academic search engines. First, we carefully select keywords, considering all aspects of our research questions and including synonyms, e.g., for tree and graph-based software visualization it may look similar to `(software OR program) AND visualization AND (tree OR graph)`. A combination of multiple digital libraries or an academic search engine with sufficient coverage is usually used, in a way that maximizes the union of papers covered and minimizes their intersection. It is beneficial if the used websites offer the export of results, otherwise it would be difficult to obtain a deterministic list of papers that can be inspected non-sequentially or by multiple researchers. Similarly to manual search, we select relevant papers based on titles, abstracts, and eventually full texts.

Snowballing means each relevant paper is scanned for backward and forward references. This extends the set of relevant papers. The added papers can again be searched for references, similarly to like a snowball grows.

These search methods are appropriately combined into a search strategy. For instance, we can perform manual search in five journals and seven conferences, then automated search using a database of academic papers, and finally perform backward and forward snowballing with a depth of two on the union of the results from the manual and automated search. During this process, it is necessary to remove duplicates, so that we do not assess one article twice.

To decide which papers are relevant for our study, we use *inclusion and exclusion criteria*. Each paper included in the study has to fulfill all inclusion criteria and must not fulfill any exclusion criterion. A simple example of such a set of criteria is:

- I1: It is a journal or conference paper published between 2015 and 2024.
- I2: The paper includes an empirical evaluation of a tree or graph-based software visualization.
- E1: Editorials, essays, secondary studies, and tool papers are excluded.
- E2: UML-based visualizations are excluded.

The criteria have to be as precise and unambiguous as possible. Ideally, two researchers should be able to independently include/exclude a set of papers with a perfect overlap.

After constructing a list of papers relevant to our study, *data extraction* is performed. We construct a table where we assign each paper numerical, categorical, or other properties (e.g., year: 2020, type: "graph", evaluation: "controlled experiment"). The set of potential categories for each categorical property is either defined beforehand or refined gradually during the process of data extraction.

Finally, we perform *data synthesis*, where we statistically derive conclusions from the extracted data, such as percentages of papers pertaining to each category. The details of data extraction and synthesis depend on the research questions that we posed in the beginning.

Exercises

1. Select a published paper in your area of interest. Try to guess how the authors came up with a research idea described in this paper. You can use any sources and hints: excerpts from the Introduction section, academic search engines, the authors' websites or blogs, or even generative chatbots, given the explanation is logically consistent and all facts are supported by sources. For example, try to find if this is an extension of the author's own previous work or of another paper by different authors.
2. Choose one of the leading conferences in your favorite subfield of computer science. Open the websites of its last two or three years and find the main research tracks. Compile a list of 10–12 session names per year. If the conference does not have named sessions, list selected paper titles instead.
3. Select a specific enough idea for which you would like to find related papers. For instance, it can be related to your thesis or any paper that you liked. Describe the idea in one or two sentences. Perform the searching process from Section 2.4.2, ignoring step 4, severely limiting or skipping the recursion in step 7, and ignoring step 8. Stop after saving about 10–12 actually relevant papers. Which part of the process was the most efficient in finding relevant papers? Are you satisfied with the results? Do you have any suggestions to improve this process?
4. Use an AI chatbot to find the works related to the idea from the previous exercise. Are the results more or less relevant?
5. Which of the mentioned advantages of reference managers is the most important to you? Can you find any other advantages?
6. Which reference manager do you use and why?
7. Find three papers relevant for you that you have never read at all. Read them with the first pass only. After reading each paper, close it and write down all the information that you remembered. How would you characterize the remembered information? For instance, is it somehow important for your research?

8. (This exercise does not require home preparation.) During your lesson, the teacher will show you a screenshot. Imagine you have an idea that could be explained by a screenshot you see. Your task is to find if a similar idea was already described in research paper(s) and find such a paper. Warning: Do not search for exact texts from the screenshot. Imagine it represents only an abstract idea in your head.

3 Types of Research

When designing a research study, it is necessary to select appropriate research methods based on multiple criteria. In this chapter, we classify types of research according to various characteristics, which will be useful for subsequent discussion throughout this book.

3.1 Quantitative, Qualitative, and Mixed

According to Creswell & Creswell (2018), we distinguish three main research approaches: qualitative, quantitative, and mixed methods research.

Quantitative research typically uses pre-determined procedures and precise numeric measurements. It often includes hypotheses describing relationships between variables, which are objectively tested with the help of statistics. The resulting report tends to have a fixed structure, with standard section names such as "Hypotheses", "Variables", and "Procedure". Typical quantitative research methods include controlled experiments and surveys with closed-ended questions.

In *qualitative* research, the procedure is sometimes not entirely fixed and the collected data are free-form, such as text, images, or videos. Researchers can make subjective interpretations of the collected data and the report does not have a standardized structure. Specific examples of qualitative methods are case studies and unstructured interviews.

Mixed methods research integrates both the quantitative and qualitative approaches. This can be done in multiple ways, but the most common are the convergent, explanatory sequential, and exploratory sequential designs.

In the *convergent mixed methods design*, both quantitative and qualitative data are collected relatively independently, possibly at the same time. Conclusions are drawn by combining the knowledge from both approaches.

Explanatory sequential mixed methods design starts with a quantitative phase, where numerical results are produced or a hypothesis is confirmed or rejected. Subsequently, the researcher tries to explain why such results occurred by conducting qualitative studies. For instance, after finding that a new unit testing technique improves the developers' productivity by 30% in a controlled experiment, we can interview the developers to determine the specific workflows and features that the developers considered useful during their work.

On the other hand, *exploratory sequential mixed methods design* starts with a qualitative phase. Here we try to explore the given topic without a strict, fixed procedure. Based on the collected data, we can find the hypotheses or research questions worth studying, design the instruments (e.g., questions in a questionnaire), or develop an intervention (e.g., a new interaction method with a user interface). In the subsequent quantitative phase, a study is performed based on this newly obtained knowledge.

3.2 Inductive and Deductive

Based on the direction of the relationship between specificity and generality, research can be either inductive or deductive.

Inductive research uses a bottom-up approach as it starts with small, specific observations. By combining and generalizing these pieces of information, a more general theory is developed gradually. As an example, we can mention the grounded theory of self-organizing agile teams (Hoda et al., 2013 and related papers).

Deductive research is top-down: it starts with a general, high-level theory. We formulate more specific hypotheses that should be true if the whole theory is true and test these hypotheses.

3.3 Philosophical Worldviews

Although this is rarely mentioned explicitly in papers, the authors have their philosophical stance toward how our world functions and how research can uncover its laws. Five common philosophical worldviews are positivism, postpositivism, constructivism, transformativism, and pragmatism.

According to *positivists*, there exists an absolute, objective truth. They aim to establish cause-and-effect relationships between variables. The purpose of research is to find such relationships solely through empirical studies and the confirmation of hypotheses.

Postpositivism is similar to positivism, but it accounts for imperfections and bias in the observations and measurements. Therefore, according to postpositivists, hypotheses and theories are never actually confirmed with absolute certainty. The more we test them without rejection, the more confident we are about their truthfulness. Both positivists and postpositivists prefer quantitative research methods and deductive thinking.

According to *constructivists*, the reality is subjective. Therefore, researchers should take the perspectives of different participants into account. Qualitative research methods should be applied, so we can better understand the richness of various views, and construct new theories based on the collected data. Constructivists thus prefer inductive thinking.

The *transformative* worldview comes from the social sciences, where the aim of the researchers is to focus on minorities that are often overlooked during classical quantitative research and suggest transformations so that their situation improves. The transformative worldview is relatively rare in computer science research, compared to the other stances.

Pragmatists combine qualitative, quantitative, and mixed methods research approaches as necessary. They do not see the world as a set of universally applicable laws. Instead, they try to find practical solutions fitting particular specific contexts.

3.4 Basic and Applied

We can also categorize research as basic or applied. Note that this categorization is rather artificial, and the distinction is often not clear.

Basic research (also called pure or fundamental) aims to obtain new knowledge about the given phenomenon, without any particular practical application in mind. It is mainly driven by the researchers' curiosity. This does not mean it is useless – many pieces of knowledge obtained by basic research were successfully applied in practice.

Applied research tries to solve a particular practical problem. The results are immediately actionable and useful to the general public or a selected population.

There exists another related term, *experimental development*, which utilizes the results of basic and applied research to create a novel product with the goal of subsequent commercialization. Together with research, they form the concept of research and development, known as R&D.

3.5 Types of Contribution

According to Wobbrock & Kientz (2016), there are seven possible types of contribution of research papers: theoretical, empirical, methodological, dataset, artifact, survey, and opinion. Although their taxonomy is tailored for the field of human-computer interaction (HCI), it is applicable for many other computer science subfields.

Theoretical contribution represents the creation of a theory, i.e., a proposition explaining certain general phenomena. It is often based on the generalization of a large quantity of evidence in the area.

Empirical contribution consists of a specific empirical study that observes the real world. Each empirical study can partially contribute to the development of a theory.

Methodological contributions improve ways we measure, analyze, and design things. Examples include a new type of questionnaire for measuring user experience or a research method to evaluate human interface devices for visually impaired people.

A *dataset* benefits the research community by offering pre-processed data ready to be analyzed by other researchers and standardized benchmarks to compare multiple approaches.

Artifacts are newly designed and innovative prototypes of software systems, hardware devices, or techniques.

Surveys analyze, categorize, and summarize a large number of existing research works on a specific topic. This should not be confused with questionnaire surveys that pertain to empirical contributions.

Finally, *opinions*, sometimes called essays, aim to lay out the author's thoughts and convince the reader using strong arguments based on scientific evidence and logical justifications.

3.6 Actors, Behavior, and Context

Stol & Fitzgerald (2018) devised a software engineering research methods classification scheme, which is, however, applicable to many other computer science subfields beyond software engineering. It is based on two key dimensions: obtrusiveness, i.e., how much the researchers modify the conditions during the measurement, and generalizability – how much the findings are applicable in other settings.

Research methods are then characterized based on whether they are maximally generalizable over *actors* (A), maximally precisely measure the actors' *behavior* (B), or preserve the maximally realistic *context* (C), thus the name "the ABC of software engineering research". By actors, the authors mean people such as developers or managers, software systems, artifacts (e.g., issues), and techniques. Behavior means system behavior (e.g., performance, storage space) and the persons' productivity, motivation, and other characteristics. Context includes industrial vs. academic settings, specific software projects, or development teams.

3.7 Primary and Secondary

Our final categorization is based on the sources of information used. *Primary* research includes the collection of new data. *Secondary* research analyzes only existing literature or existing data sources. For instance, systematic literature reviews are considered secondary studies. *Tertiary* studies review multiple secondary studies.

Exercises

1. Suppose you would like to determine which of the three database insertion techniques is the most energy efficient. Which research approach would you use – quantitative, qualitative, or mixed methods?

2. Find a qualitative research study that you consider interesting.
3. Find a mixed methods research study in your area of interest. Can convergent, explanatory sequential, or exploratory sequential mixed methods design be clearly recognized in the paper?
4. Which of the mentioned philosophical worldviews do you hold and why?
5. Which philosophical worldview do Shreeve et al. (2023) hold in their paper?
6. Find any paper in your research area whose main contribution type is:

    a. theoretical,
    b. empirical,
    c. dataset,
    d. artifact.

7. Inspect Figure 1 in the ABC framework paper (Stol & Fitzgerald, 2018). Select one of the eight research strategies shown in the circle (experimental simulations, field experiments, …). Find a paper in your area using the selected research strategy.
8. Find any secondary study in computer science that synthesizes quantitative results from a large number (at least 20) of existing papers.

4 Choosing Research Questions and Methods

After we select a specific research topic and idea, it is necessary to specify the research question(s) and/or hypotheses and choose a research method to answer them.

4.1 Research Questions

A *research question* (RQ) is a clearly formulated question that we would like to answer in our study using research methods. In general, a good research question should be:

- Focused: If we have multiple research questions per paper, they should all relate to the main goal.
- Researchable and feasible: We must be able to answer it scientifically and using reasonable resources.
- Specific: A research question should not be vague.
- Complex enough and relevant: We should not study questions that will be useless for sure.

To provide more specific examples, Begel & Zimmermann (2014) and Huijgens et al. (2020) provide lists of questions that industrial software engineers find relevant. Similar lists may exist for other subfields.

Sometimes, a research question (or a hypothesis) is only implicit in a paper. This is the case mainly for papers whose main contribution is an artifact – the research question then might be "Is it feasible to design X?" or "How to design an approach X so that it has some property Y?". However, particularly for empirical studies, we should always state a research question explicitly.

4.2 Hypotheses

A *hypothesis* is a declarative sentence whose truthfulness is yet unknown. It often specifies a relationship between variables, e.g., "A high-contrast mouse cursor improves the clicking precision on a target rectangle on a screen." A hypothesis must be falsifiable, which means we can reject it using an empirical research method.

Any hypothesis can be written as a research question, but the reverse does not always hold. For instance, the question "How do developers unit test database modules in procedural languages?" cannot be written as a hypothesis. If a statement is statistically testable, the hypothesis is a preferred form.

A hypothesis is a statement about a more specific phenomenon, such as "When debugging, having dynamic information from Senseo (Rothlisberger et al., 2012) available reduces the time for solving maintenance tasks." In contrast, a *theory* is a proposition explaining a certain general phenomenon. It is usually based on the generalization of a large quantity of evidence in the area. For example, an information foraging theory for debugging (Lawrance et al., 2013), explains how the developers navigate the program, similarly to a predator following prey, using concepts such as scent, proximal cues, or topology.

4.3 Operationalization

We know that a research question or a hypothesis should be specific and precisely defined. However, even well-stated hypotheses and RQs contain terms with many possible interpretations. Consider the hypothesis:

> Using a time-traveling debugger improves developers' efficiency in fixing bugs.

There exist multiple time-traveling debuggers with different features. Many people consider themselves developers: from high-school students coding for fun to senior developers in corporations. What actually constitutes efficiency is a notoriously debated problem.

We thus have to provide *operational definitions* of important terms. An operational definition expresses how the given term will be interpreted in this specific study, i.e., how it will be measured, observed, or applied.

For example, we will use a specific time-traveling debugger TimeDebugX, implementing certain features. By "developers", we will understand professional programmers with at least one-year industrial Java experience. We will measure "efficiency" as the number of successfully solved bug-fixing tasks per hour.

4.4 Choosing a Research Method

Many beginning researchers make a mistake of selecting a research method without having any specific hypothesis or RQ in mind. For instance, they design a new approach, which they would like to empirically evaluate. They start thinking about a questionnaire survey as a relatively easy way to collect data from multiple respondents. They start adding many questions to the survey, ranging from demographic details of every kind to many unclear and suggestive questions about how the respondents liked various aspects of the newly designed system.

This is definitely a wrong practice. When doing research, we should always *define RQs/hypotheses first* and only then proceed to choosing research methods. Otherwise, we risk that the method will answer only questions that are irrelevant or will not reliably answer any research question at all.

There are multiple guides and overviews suggesting the most suitable research methods for given types of research questions, e.g., by Wohlin & Aurum (2015) or Easterbrook et al. (2008). In the next chapters, we will describe some of the frequently used research methods, including controlled experiments, surveys, interviews, case studies, and benchmarking studies among others. For each method, we will mention examples of RQs it is suitable for answering. However, in practice, the methods and their principles are sometimes not clearly separated, so we may use, e.g., the concept of sampling known from surveys in a benchmarking study.

During a research project, it is often beneficial to use multiple methods or data sources in a practice called *triangulation*. For example, we can obtain requirements for our new approach using a survey, assess its machine-based efficiency by a benchmark, find its advantages and disadvantages in a case study, and finally assess its human-based efficiency using a controlled experiment with humans. This increases the validity of the study and provides a more comprehensive understanding of the subject.

Exercises

1. Which of the following are good research questions and why?
    a. Why do students fail in the operating systems bachelor's courses?
    b. How to determine if a program written in C terminates without running it?
    c. What proportion of C++ Gists on GitHub is compilable without downloading third-party libraries?
2. Which of the following are good hypotheses? Why?
    a. Java is a fast programming language.
    b. An LL parser is faster than an LR parser for converting JSON into an in-memory tree structure.
3. Find a paper in your field that explicitly lists operational definitions of terms used in a hypothesis or a research question.

5 Controlled Experiments

In this chapter, we explain one of the fundamental research methods, controlled experiments. A controlled experiment is typically used to answer research questions in the form "Does X cause Y?" or "Is A more efficient/precise than B?".

5.1 Correlation vs. Causation

Suppose we would like to know what is the relation between playing platform games and kids' typing speed on a computer keyboard. We start asking a very large random sample of kids whether they are playing platform games. During the study, we also measure the typing speed of each participant using a standardized test. As a result, data similar to Table 5.1 is produced, with hundreds of rows.

Table 5.1: A sample of the collected data on the relation between playing platform games and typing speed

Participant ID	**Plays platform games**	**Typing speed (words/min)**
1	false	32
2	false	21
3	false	25
4	true	65
5	true	53
6	true	67
…	…	…

For all participants not playing platform games, the mean is 27 words per minute, while for the playing ones, a mean of 62 was computed. Platform game players thus seem to be markedly faster at typing. Suppose we report our findings in a research paper that gets published. After some time, a famous tabloid comes with a flashy headline:

> "Playing platform games improves typing speed" say researchers.

Based on this fantastic news, many parents sign their kids up for a platform gaming club. Every week, they play Super Mario, Crash Bandicoot, Seiklus, and other popular platformers. Sadly, after a year of playing, the parents find out there is absolutely no difference in their kids' typing speed.

Why is this the case? Clearly, there is a difference between correlation and causation. The study showed there is a *correlation*: "Kids who play platform games tend to type faster". However, it did not show *causation* at all: "Playing platform games causes kids to type faster."

The question arises about what is the true causation in our example. Is it reversed, i.e., typing faster causes kids to play computer games? In general, reverse causality is possible, but in our specific example, this does not seem plausible. Instead, we should look at other hidden variables that were not measured in our study. During interviews with the participants, we might find out that practically all kids that play platform games also chat using various instant messaging applications. In Table 5.2, there is a new column "chats", representing a *confounding factor*, which is a variable related both to the incorrectly presumed cause (playing platform games) and the outcome (typing speed).

Table 5.2: Internet chatting as a confounding factor

Participant ID	**Plays platform games**	**Chats**	**Typing speed (words/min)**
1	false	**false**	32
2	false	**false**	21
3	false	**false**	25
4	true	**true**	65
5	true	**true**	53
6	true	**true**	67
...	...	...	...

Upon closer investigation, we might find that the reality is even more complicated, as the root cause of both game-playing and chatting is that a given kid has a dedicated computer at home. This situation is displayed in Figure 5.1.

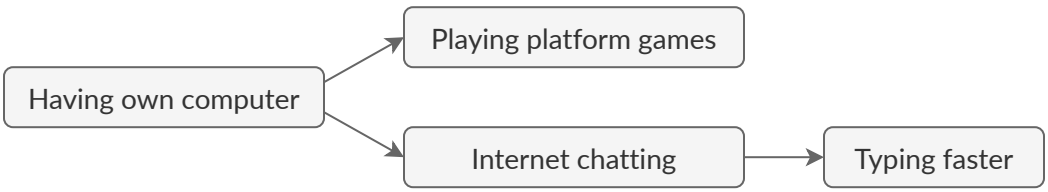


Figure 5.1: Probable true causation; arrows represent causality

In general, there may be many confounding factors. We could carefully search for such confounding factors and take them into account during statistical calculations. This process is

called *controlling for* variables. However, if practically possible, a much more valid approach should be applied instead: a controlled experiment.

A *controlled experiment* is based on the idea that we keep all conditions fixed, while manipulating only the presumed cause. In a very common type of a controlled experiment, a *randomized controlled trial*, we randomly divide a set of subjects (e.g., people, programs) into two or more groups. Each group receives a different treatment; if a group receives no treatment at all (or a placebo), it is called a *control group*. At a specified time(s), we measure the outcome for each subject and compare the results between the groups. This approach works, particularly with a large enough number of subjects, because randomization diffuses the subjects having various characteristics into all groups.

In our example, we could divide the kids (preferably the ones not playing platform games regularly) randomly into two groups. The first one will play platform games for two hours a week, while the second one will not play platform games at all. After a few months, we compare the typing speed of the groups.

5.2 Variables

Before performing a controlled experiment, we need to define variables. A *variable* in research is a measurable, observable, or manipulable characteristic that can vary. An *independent variable* is a condition which we manipulate in the controlled experiment. A *dependent* variable is an outcome which we measure or observe. Its name is derived from the fact that we hypothesize it depends on the value of the independent variable.

Each variable has a *scale*. There are four possible scales:

- *Nominal*: one of n possible values called *levels*, without any particular order, such as Linux/Windows. A nominal variable with two levels is called *dichotomous* or *binary*.
- *Ordinal*: one of n values that can be ordered but not subtracted, e.g., primary, secondary, or tertiary education.
- *Interval*: it has equal differences between subsequent values, e.g., a calendar date.
- *Ratio:* a ratio of two values is computable, with a meaningful zero (e.g., time spent on development in hours). It can be further characterized by its statistical distribution, usually normal/non-normal.

The list is ordered from the most general scale to the most specific one, so if an appropriate statistical test exists only for a more general scale, it can be applied instead.

In our example experiment, an independent variable of a nominal scale is playing/not playing platform games. A dependent variable of a ratio scale is the typing speed in words per minute.

5.3 Hypotheses

Next, we define a null and alternative hypothesis. A *null hypothesis*, denoted H_0, assumes there will be no difference between the groups in the experiment, which means changing the value of the independent variable does not cause a change of the dependent variable. An *alternative hypothesis* (H_1 or H_a) stipulates a difference between the groups will be present, so there is a causal relationship between the independent and dependent variable.

In our example, the null and alternative hypotheses could be defined as follows:

- H_0: Playing platform games makes no difference in the typing speed of kids.
- H_1: Playing platform games makes a difference in the typing speed of kids.

The alternative hypothesis in our example is a *two-tailed hypothesis* since it considers a change in both directions; the difference could be either positive or negative. A *one-tailed hypothesis* would consider only the specified direction and deny the possibility of the opposite one, e.g., "Playing platform games improves the typing speed of kids." We should generally use a two-tailed hypothesis unless we have good reasons not to do so, such as when the opposite change is impossible or irrelevant.

5.4 Experimental Design

An *experimental design* formally defines how the experiment will be executed. It tells us how the assignment of the subjects to groups will be performed. It also specifies how many times and in what order the treatment will be administered and the outcome measured.

Based on the randomness of the assignment of the treatments to the subjects, we distinguish:

- *quasi-experiments*, where assignment is performed using some convenient criterion, e.g., the groups will consist of employees working in given teams or students attending given classes
- and *true experiments*, where the assignment is random.

From the perspective of the sequence of treatment administration and measurement events, there exist multiple experimental designs. The most common ones are:

- the *pretest–posttest design*, in which the value of the dependent variable is measured both before and after the administration of the treatment
- and the *posttest-only design*, where we measure the outcome only after the treatment.

Another categorization specifies the number of treatments each subject receives. In a *between-subject design*, each subject is assigned to one specific group during the whole experiment, receiving only one treatment. Table 5.3 shows an example of such a design.

Table 5.3: Assignment of treatments to subjects in a between-subject design

Table 5.4: Group 1

Subject ID	Treatment
1	1
3	1
5	1
...	...

Table 5.5: Group 2

Subject ID	Treatment
2	2
4	2
6	2
...	...

When using a *within-subject design*, each subject receives a treatment, and then the outcome is measured. Then the same subject receives another treatment, and the outcome is measured again. If there are more treatments, these steps are repeated. Table 5.6 displays an example of the assignment of the subjects to treatments in a within-subject design with two treatments. Note that we used *counterbalancing*, i.e., subject #1 was administered treatment 1 first and then treatment 2, while subject 2 received them in the reversed order. This is done to prevent systematic bias, where one of the treatments would be given an advantage.

Table 5.6: Assignment of treatments to subjects in a within-subject design

Subject ID	First treatment	Second treatment
1	1	2
2	2	1
3	1	2
4	2	1
...	...	...

A between-subject design is usable in almost all situations, but it requires more subjects than the within-subject design. On the other hand, a within-subject design requires fewer subjects, but the learning effect and fatigue are a problem when the subjects are human.

An experiment can have multiple independent variables. The most straightforward way is to have a group for every possible combination of independent variables. For instance, with two dichotomous variables, we would have four groups. This is called a *factorial design*.

5.5 Effect Size

Suppose we execute our example experiment and collect the data. Let us say the mean of the not-playing group in the controlled experiment is 45.1 words/minute, and the mean of

the playing group is 45.8. Therefore, we can say that on average, the playing group was 0.7 words/minute – or 1.6% – faster. These are one of the simplest ways to express the *effect size*. Although they are valid and easily comprehensible, there exist standardized metrics of effect sizes, such as Cohen's d, which take variability (standard deviation) into account and simplify the meta-analysis of multiple papers. Kampenes et al. (2007) provide a systematic review of effect sizes commonly used in software engineering, including an overview of the possibilities.

5.6 Statistical Testing

One question remains: Are these two mean values (45.1 and 45.8) sufficient to reject the null hypothesis (no difference) and accept the alternative hypothesis? The answer is no because the difference in means could be observed purely by chance. We need to analyze the whole dataset to find whether the result is statistically significant.

For this, we compute a *p-value*, which is the probability of observing the data at least as extreme as we observed if H_0 was in fact true. The largest acceptable p-value is called a *significance level*, denoted α, and must be stated before the execution of the experiment. The most commonly used value of alpha is 0.05 (i.e., 5%). If $p < \alpha$, we reject the null hypothesis, otherwise we fail to reject it.

The p-value is computed by an appropriate statistical test. To select a test, we can use, e.g., a table by UCLA or by Philipp Probst. Each test should be applied only in situations when its assumptions are met. For instance, the independent-samples t-test has multiple assumptions: the data should be approximately normally distributed, there should be no significant outliers, etc. Many statistical tests are implemented in the libraries of programming languages, such as R or Python.

Let us say that in our example, the computed p-value would be 0.67. This means we could not reject the null hypothesis, so we would not show there is any effect of playing platform games on typing speed. If the p-value was, for instance, 0.04, we would reject the null hypothesis and show an effect.

In Table 5.7, we plotted the computed result of the experiment (rejection of H_0 or a failure to do so) against the reality, which is unknown. Correctly rejecting or not rejecting H_0 is called a true positive and a true negative, respectively. Accepting the alternative hypothesis, i.e., showing an effect, while there is in fact no effect, is called a Type I error. The maximum acceptable probability of a Type I error is α, as we already mentioned.

Table 5.7: Possible outcomes of an experiment compared to the reality

		In reality, H_0 is	
		true (no effect)	**false (effect)**
We reject H_0	**no**	true negative	Type II error

	In reality, H_0 is	
	true (no effect)	**false (effect)**
yes	Type I error	true positive

There is also another type of error in Table 5.7: a Type II error. It means we failed to show an effect when there actually is one. The probability of a Type II error is denoted β (beta), with a common value of 0.2. To mitigate a Type II error, we should perform the experiment with a large enough number of subjects and choose a suitable statistical test, so the experiment will have enough *power* $(1 - \beta)$. It is possible to estimate the number of subjects required for a given statistical test to have the given power using a process called power analysis.

5.7 Threats to Validity

Every research project has issues that could negatively affect its validity, which we should mention in the report in a section called "Threats to Validity". In controlled experiments, these are some common types of validity which can be threatened:

- *Construct validity*: Did we choose the right variables and measures?
- *Internal validity*: Are the results affected by other factors beyond causality?
- *External validity*: Are the results generalizable to other situations?

In addition to listing the validity threats, we should also mention how we tried to mitigate them or why it was not possible.

5.8 Complete Example

To show the whole picture, now we will describe a complete example of an imaginary controlled experiment, along with the data analysis using the R statistical programming language. Suppose we have the following *research question (RQ)*: Does the syntax highlighting type (none, mild, strong) affect the developers' speed when locating a code fragment?

5.8.1 Hypotheses

We formulate the null and alternative hypothesis:

- H_0: There is no difference in the developers' mean code location speed when using none, mild, or strong syntax highlighting.
- H_1: There is a difference in the developers' mean code location speed when using none, mild, or strong syntax highlighting.

We will test with $\alpha = 0.05$.

5.8.2 Variables

There is one *independent* variable: syntax highlighting type. It is nominal (categorical) with 3 levels. The *dependent* variable is of the ratio type (code location time in seconds).

5.8.3 Design

One developer cannot search for the same code fragment multiple times using different highlighting styles because of a learning effect. We could use different code locations and source files for each syntax highlighting type, but different locations/files can be of varying difficulties. We also had a large pool of participants available. Therefore, we used the between-subject posttest-only design.

5.8.4 Procedure

We recruited 33 final-year computer science Master's students. We divided them randomly into 3 groups (none, mild, strong).

They were given 10 source code files. For each source code file, there were two tasks specified, such as: "Find the first ternary operator in the function sum()" or "Find the exception handling code in a function which receives a JSON object".

After reading each task, the subject pressed a shortcut to display the code file. At this moment, the timer started. After finding the fragment, the participants placed a cursor on it and pressed another shortcut – if the location was correct, the timer stopped. The total time was recorded for each subject.

5.8.5 Results

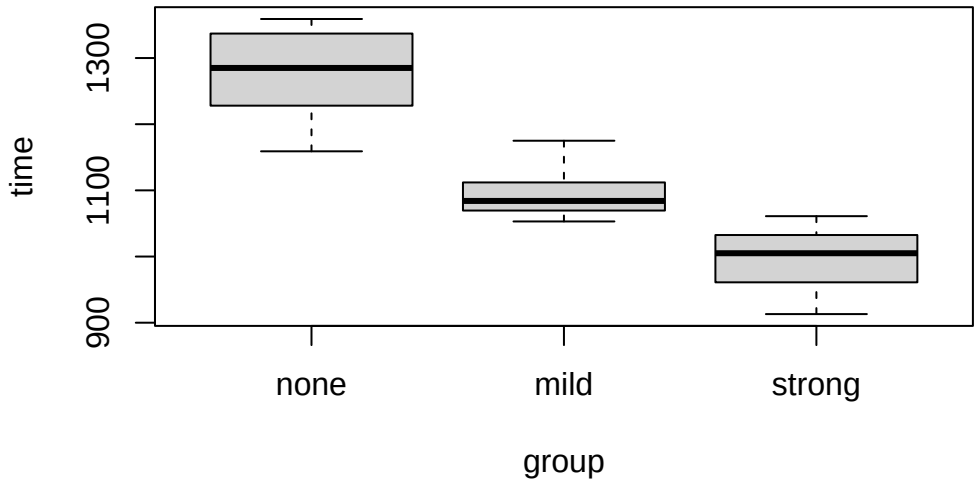
The measured values are located in a CSV file, of which we show only the first few rows for brevity.

```
results <- read.csv("data/experiment.csv", header = TRUE)
head(results)
```

```
   group time
1   none 1241
2   none 1215
3   mild 1060
4   mild 1084
5 strong 1005
6 strong 1005
```

A graphical overview of the differences between groups in the form of a box plot follows.

```
results$group <- factor(results$group, c("none", "mild", "strong"))
boxplot(time ~ group, results)
```



We compute the means of each group:

```
aggregate(time ~ group, results, mean)
```

```
   group      time
1   none 1275.0000
2   mild 1097.0909
3 strong  993.5455
```

How much faster were the developers (i) with mild formatting compared to none and (ii) with strong formatting compared to none?

```
difference <- function(col1, col2) {
    diff = (mean(results[results$group == col1, "time"])
            / mean(results[results$group == col2, "time"]))
    paste(round(100 * (1 - diff), 3), "%")
}

print(difference("mild", "none"))
```

```
[1] "13.954 %"
```

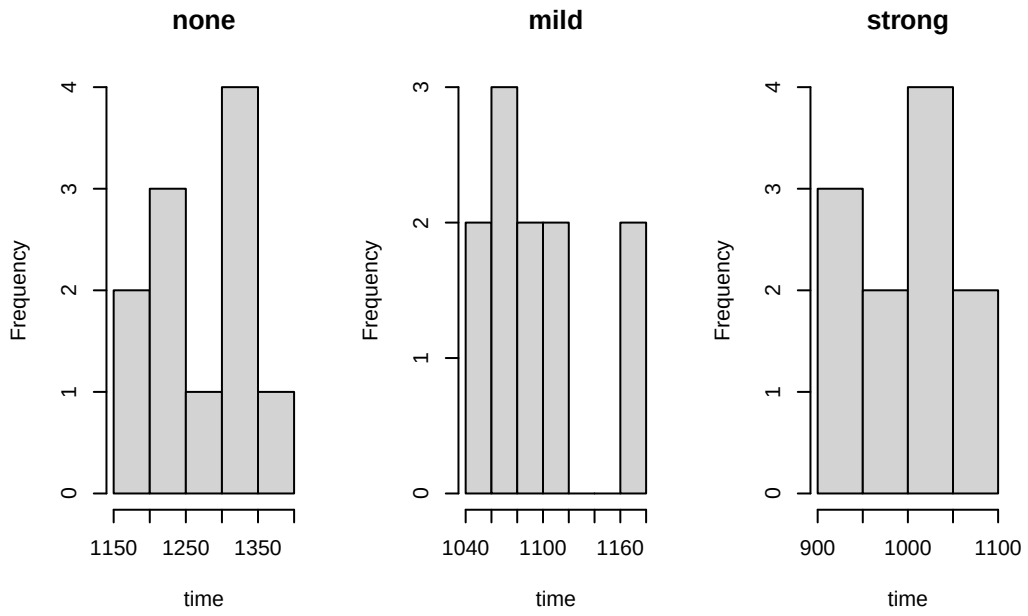
```
print(difference("strong", "none"))
```

```
[1] "22.075 %"
```

To find an appropriate statistical test, we first determine whether the data are normally distributed.

```
options(repr.plot.width = 6, repr.plot.height = 3)
par(mfrow = c(1, 3))

for (group in levels(results$group)) {
    hist(results[results$group == group, "time"], main = group, xlab = "time")
}
```



Since the data are not normally distributed, we will use the Kruskal-Wallis test.

```
kruskal.test(time ~ group, results)
```

```
    Kruskal-Wallis rank sum test

data:  time by group
Kruskal-Wallis chi-squared = 27.312, df = 2, p-value = 1.173e-06
```

Since the p-value is less than 0.05, we reject the null hypothesis. There is a statistically significant difference between the groups.

5.8.6 Threats to Validity

- *Construct validity*: Timing may not be perfect (the subject can press the shortcut too late).
- *Internal validity*: Some developers might be far more experienced. However, this threat should be mitigated by randomization to a large extent.
- *External validity*: The source code and tasks may not represent the typical source code and location tasks used in practice. It is also questionable to what extent the Master's students represent developers in general.

5.8.7 Conclusion

Developers performed the best using the strong highlighting type (22% faster than with none), followed by the mild highlighting (14% faster than with none). The results are statistically significant.

Exercises

1. Describe a specific situation when performing a controlled experiment in computer science to show causation would be:

   a. unethical,
   b. too time-consuming or expensive,
   c. physically impossible.

   What would you do instead?
2. Which are independent and dependent variables in the following hypotheses and what are their scales?

   a. Computer science teachers with more years of experience spend more time weekly doing research.
   b. Using a terminal instead of a graphical user interface to launch the debugged application lowers the number of launches.
   c. On a scale from saddest to happiest, using IntelliJ IDEA makes Java developers happier than Visual Studio Code.

3. State a two-tailed hypothesis and both possible one-tailed hypotheses about an experiment comparing a touchpad and a trackball with respect to the precision of clicking on a target.
4. Name a hypothesis for which selecting a quasi-experiment would be most suitable.
5. Suppose we are designing a posttest-only controlled experiment. Describe a situation (other than the examples in this chapter) in which:

   a. a within-subject design would be perfectly usable,
   b. a between-subject design has to be used.

   Why?
6. Which statistical test would you use:

   a. for comparing two groups in a between-subject controlled experiment with a non-normal ratio dependent variable,
   b. for comparing three groups in a within-subject controlled experiment with an ordinal dependent variable?

7. Find a paper describing a controlled experiment. How does it categorize threats to validity? Name one from each category.

8. Describe an imaginary experiment ruined by a critical threat to its internal validity. How would you prevent it?
9. Find a paper, preferably in your research area, reporting on a controlled experiment. In its text, mark the null and alternative hypotheses, variables (along with their scales), type of the experimental design, and the used statistical test. How was the effect size reported? If some of this information is not explicitly stated, deduce it.
10. A experiment can also be performed without human participants – the "subjects" can be programs, files, executions, and so forth. Find an example of such a controlled experiment.

6 Human-Centered Methods

In this chapter, we provide an overview of common research methods that directly involve human subjects. This should not be considered an exhaustive list of strictly defined and clearly separated methods, but rather a selection of typical examples with certain characteristics that can be customized as necessary.

Many of the concepts and principles described here are also usable in purely computer-based studies. For instance, sampling also applies to projects in software repository mining studies, and qualitative coding (assigning tags to parts of text) can be performed on texts from an Internet forum.

6.1 Ethical Considerations

Before performing any empirical study with human participants, it is necessary to consider the ethical standpoint of research. Specific rules that apply depend on the institution and country where the research is performed, but there exist several general documents providing general guidelines.

One of them, the Belmont report specifies three basic ethical principles: respect for persons, beneficence, and justice. They imply a number of practical rules.

First, subjects have to know that they are participating in a research study, and their participation has to be *voluntary*. They should be informed about the general goal of the study. If the knowledge of specific details would affect the validity of research negatively, it can be postponed to the *debriefing* phase after the study. Deception, i.e., providing outright false information about the purpose of the study, should be used only if absolutely necessary and the true purpose must be revealed during debriefing. Apart from voluntariness, a person has to be able to withdraw from the study at any point.

Second, the risks must outweigh the benefits of the study. This applies both to the individual and society level. For instance, at the individual level, the participants in the study can receive small compensation, such as a voucher, or be informed about the results sooner than the general public if they wish so. At the society level, a slightly higher-risk study is justified if a significant breakthrough is expected thanks to the results, but the risks should not be higher than absolutely necessary and never cross a certain line.

Finally, the sample that will be studied should be selected fairly. For instance, if our research is aimed at professional software engineers, we should not constantly use only students in every experiment because they are conveniently available.

Before conducting research with human subjects, an *institutional review board* (IRB) approval is required in many cases. An institutional review board is a group of people at an institution, such as a university, which decides whether a given research study is ethical. An IRB application is institution-specific, but generally it contains information about the investigators, a hypothesis or research question, a succinct but complete description of the planned research method, and a sample informed consent document. An *informed consent* is an agreement of an individual to participate in a study. It is best if it has a written form (electronic or paper-based), since verbal informed consent is difficult to prove. An informed consent should contain:

- the name of the research project and the purpose of the study in simple terms,
- contact information of the researchers,
- the types of tasks that the subjects will perform,
- the data collection procedures and storage conditions,
- risks and benefits,
- and the confirmation that the person participates voluntarily and can withdraw at any time.

6.2 Surveys

A survey, in this sense, is a method to collect and analyze data from humans using a questionnaire with a fixed set of questions. This questionnaire can be either paper-based or, very commonly, administered as a web form. The usual purpose is to find certain self-reported information about a population, e.g., the average number of monitors used by professional 3D modelers, or the most common problems Haskell developers face when using monads.

6.2.1 Sampling

While it is sometimes possible to administer a survey to the whole population, often this is not feasible. This is also the case for our example, i.e., the population of all Haskell developers in the world. In order to make a study practically executable, we need to select a *sample* from the population. There are two principal ways to attain this: probability and non-probability sampling.

In *probability sampling*, every member (or item) of the population has a known, computable probability of being selected for the sample. For this, we ideally need a list of all members of the population. If it is not available, an imperfect *sampling frame* can be used, listing a large portion of the population. Two of the most common probability sampling strategies are:

- *Simple random sampling*: Each member has the same probability of being selected.
- *Stratified random sampling*: We divide the population into groups called strata. A *stratum* contains members having certain characteristics, such as junior or senior developers only. Then we can randomly sample members from each stratum. For instance, if there are 400 junior and 100 senior developers in the population, and we use a proportionate allocation strategy for a sample of 50 developers, then 40 junior and 10 senior developers will be in the sample.

When a list of all population members or a sampling frame is not available, we need to use *non-probability sampling*. Some of the commonly used strategies are:

- *Convenience sampling*: Subjects who are most conveniently accessible to researchers are used. This includes, for example, students in a teacher's class or employees in a company where the researcher used to work.
- *Self-selection* (*volunteer sampling*): Here, instead of the researchers sampling the subjects, the participants select themselves to become a part of the sample. For instance, after publishing a link to a questionnaire on a public forum, people who are interested can participate.
- *Snowball sampling*: After an initial group of people participates in the study, they recruit additional participants, which call more of them, etc.

6.2.2 Questions

Before designing questions in a survey, it is necessary to carefully consider what we would like to find out. Each survey question should contribute to answering one of our research questions or confirming our hypotheses.

Some RQs about simple variables and facts can be directly transformed into an item in a questionnaire. For example, if we want to know what IDEs Haskell programmers use, we can simply ask: "What IDE do you use?"

However, we often want to study abstract, multi-dimensional *constructs*, which cannot be directly measured or observed. For instance, asking "What is your cognitive load when merging Git commits?" does not make much sense. Constructs such as cognitive load need to be operationalized into measurable items. In the case of surveys, one multi-dimensional construct is usually operationalized in a set of multiple related questions, either standardized or designed specifically for the given study by a researcher. In our example, the construct of cognitive load could be operationalized using the standardized NASA Task Load Index (TLX).

The questions asked in a survey can be categorized into two main groups: closed-ended and open-ended questions.

Closed-ended questions include numeric, single-choice (including the Likert scale), multiple-choice, item-sorting (ranking), and similar questions that can be directly analyzed quantitatively.

They are useful to answer many kinds of RQs, mainly about counts, frequencies, proportions, such as: What is the average value of X? What proportion of X is Y? Considering multiple questions at once, we can also ask: Are X and Y correlated?

Open-ended questions require free-form texts, images, sounds or other unstructured data as answers. After qualitative analysis, they can be used to answer many kinds of "how", "why", "what", and other RQs, e.g.: What is the users' attitude to X? Why do people do X? How could X be improved?

6.2.3 Pilot Testing

Sending a survey to hundreds of subjects and then realizing the most important questions was ambiguous is really unpleasant. Before starting the actual response collection, first try the survey on a small group of people. This group does not have to be representative, and it is usually obtained by convenience sampling. Responses from the pilot testing will not be included in the final analysis.

During the pilot testing, we should focus on the clarity of the questions, appropriateness of the response scales, the time required to complete the survey, and technical issues. We can then modify the survey according to the feedback if necessary.

Pilot testing is often applied also to other research methods involving human participants, especially controlled experiments.

6.2.4 Preprocessing

If we used a sampling and recruitment method that involved invitations of particular persons, most likely not all of them actually filled in and submitted the questionnaire. It is thus necessary to mention the response rate in the report.

The questionnaire form should be configured to validate the inputs whenever possible, such as forbidding empty answers for mandatory questions. Nevertheless, the received responses should be checked for signs of invalid responses, such as free-form answers containing nonsensical texts or combinations of answers to two questions that do not make sense. There are three basic options to deal with invalid responses:

- Exclude the whole response of the given participant.
- Exclude only the invalid answer to the given question, but leave the rest of the answers of the participant as is.
- *Impute* missing or invalid values (Miao et al., 2023). This has to be done carefully so as not to lower the validity of the study.

Whatever the choice was, it has to be clearly documented in the report, including the specific counts of invalid responses.

6.2.5 Analysis of Results

The analysis and reporting of the results depend on the nature of the questions. Open-ended questions require specialized analyses, such as qualitative coding, which we will describe in the next section about interviews. Closed-ended questions are usually reported using descriptive or inferential statistics.

When choosing *descriptive statistics*, we can report means, medians, and standard deviations of percentages or absolute frequencies. With non-probability sampling, we should generally not claim representativeness for the whole population. For instance, you cannot say "40% of all C# developers use lambda expressions" if you surveyed only programmers from one company.

A large variety of *inferential statistics* can be computed from survey results. Some of the most common ones include the calculation of confidence intervals, differences between groups, and relationships between variables.

When the participants were selected using probability sampling, we typically report confidence intervals for individual variables. A *confidence interval* is a range of values that likely contains the true value, e.g.: "40% of subjects selected this option (95% confidence interval is +/-4%)". It consists of the definition of a confidence level and a margin of error. The *confidence level*, usually 95% or 99%, tells us that if we repeated the random sampling infinitely, the given percentage of intervals would contain the true value. The given interval is defined by the *margin of error*, e.g., +/-4% in our example.

Instead of post-hoc reporting of confidence intervals only to find that they are too broad because of an insufficient sample size, we can calculate the required sample size before starting the recruitment of participants. This can be accomplished with statistical libraries or an online sample size calculator for simple random sampling.

Differences between groups can be calculated when we have a categorical variable representing groups (e.g., the role: a manager, programmer, or tester) and another concept of interest (such as the number of lines of code written per day) captured in a survey. Then we can compute averages of the groups and determine whether the differences are statistically significant using an appropriate test, which is a procedure similar to the one used in controlled experiments (Section 5.6). Of course, unless we include confounding factors in the survey and control for them, we cannot show causation.

Other *relationships between variables* captured in the survey can be computed too, such as *correlation* between two numeric variables. We will demonstrate this on a made-up example.

6.2.6 Example

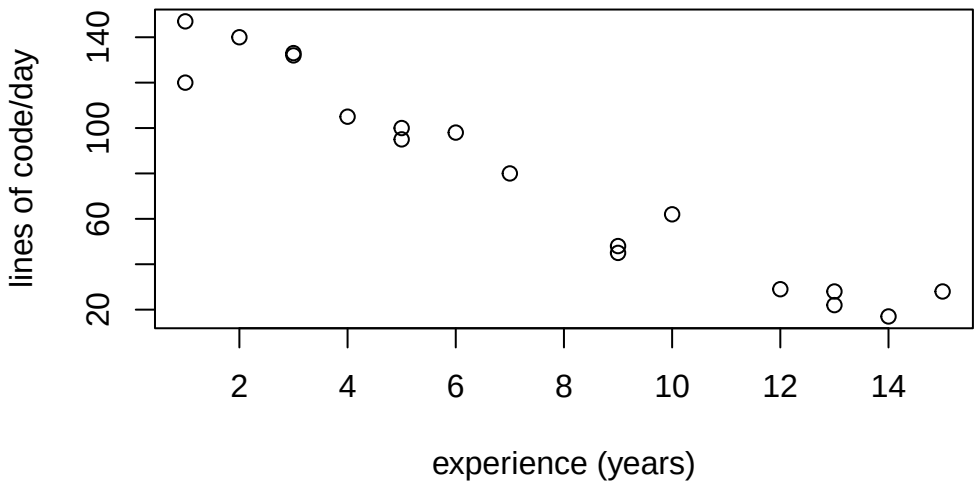
We have a hypothesis that developers with more years of experience write more lines of code per day. Suppose we collected data using a survey.

A web questionnaire was designed, containing only two questions, both mandatory:

1. How many years of industrial experience do you have? (numeric answer)
2. How many lines of code do you write on average per day? (numeric answer)

Pilot testing with three subjects did not reveal any problems. Convenience sampling was used; an invitation to participate in the survey was sent to 30 developers in four software companies in one city. Two weeks were allocated for response collection. Twenty-one developers filled the questionnaire, resulting in a response rate of 70%. Two responses were excluded since they reported more than 100 years of experience.

```
results <- read.csv("data/survey.csv", header = TRUE)
plot(lines ~ experience, results,
     xlab = "experience (years)",
     ylab = "lines of code/day")
```



We calculate Spearman's ρ, which is a correlation coefficient usable for monotonic (including non-linear) relationships:

```
cor(results$lines, results$experience, method = "spearman")
```

```
[1] -0.9632526
```

The results indicate a strong negative correlation (-0.96). The hypothesis was not confirmed; on the contrary, according to our results, more experienced developers tend to write fewer lines of code per day. Our study is limited by an important construct validity threat, which is manual, approximate reporting of the lines of code written per day instead of automated measurement in an IDE. Obtaining responses from companies in one city is an external validity threat.

For more realistic and high-quality examples of survey research studies, we recommend papers by Inal et al. (2020) and Begel & Zimmermann (2014).

Exercises

1. In what research study with human participants in computer science would deception be necessary for validity?
2. Is it ethical if students receive course credit (points) for participating in research studies? Why or why not?
3. Find a paper in your research area describing a survey (questionnaire research). Was the population or the sampling frame explicitly mentioned? Which sampling strategy was used?
4. Find a computer science paper using stratified random sampling – not necessarily for a survey. What were the strata?
5. Some researchers claim that everything, including simple variables such as the age of a person or the number of lines in a file, is a construct that cannot be measured directly and thus needs operationalization. What arguments support their claim?
6. What specific problems can be discovered during the pilot testing of a research study? You can answer using your own experience, an existing paper, or a made-up situation.
7. What are advantages and disadvantages of excluding only invalid answers to specific questions in a response instead of the exclusion of the whole response of a participant?
8. If you had a list of all 12,000 users of the library FooBar, how many participants would you need for a survey to claim that a certain percentage of the users wish FooBar was implemented also for Lua with a 2% margin of error and 95% confidence interval?

Assignments

Assignment 1: List of Papers for a Systematic Review

Imagine you are doing a systematic review (see Section 2.8). Leave out data analysis and synthesis – perform only the initial steps:

- define the topic and research question(s),
- specify the search strategy in detail,
- define inclusion and exclusion criteria,
- perform the search,
- and list the found relevant articles matching the criteria.

The narrower the topic, the less work you will potentially have during filtering of the papers; recall that a systematic review aims to ideally include *all* relevant works. However, there should be at least 20 relevant papers in the resulting list. If inclusion criteria specify a time range, it should span at least 5 years.

Your goal is to achieve as high recall and as high precision as possible with respect to all published research papers relevant to the given topic/question and matching the criteria. Since the set of all relevant published papers is unknown, the reviewer (teacher) can only guess it by assessing the quality of the research method and/or by performing random keyword searches.

You can use the fact that you are working in teams to your advantage: for example, label a subset of papers by multiple persons, compute inter-rater reliability, and argue with this for the quality of your study. Alternatively, label all papers by two people and then resolve disagreements.

Semi-automated tools are allowed and encouraged. However, simply entering your research question into a chatbot and copying all results is definitely not enough. As the internal workings of the tool are opaque, you have to ensure the research method is valid, a large portion of relevant papers are included and irrelevant ones excluded.

The assignment should be submitted as a report in PDF format, containing a clear description of the method (including a diagram is helpful) and the list of relevant papers as bibliographic citations.

Assignment 2: Controlled Experiment Data Processing

Your task is to design and describe an imaginary controlled experiment. It will not be actually executed, so you can make up the data and state this in the report. Although unethical in practice, as an assignment this is a practical approach since executing a controlled experiment is often resource-intensive, and finding existing raw data from a controlled experiment without a paper already describing this experiment in detail is rare.

The report should describe at least:

- the research question, the null and alternative hypotheses,
- variables (including their scales),
- experimental design,
- details of the procedure,
- results (the made-up data, effect size, statistical testing),
- threats to validity,
- and conclusion.

The topic of the experiment should be related to computer science (interdisciplinarity is accepted). Quasi-experiments are allowed, but this has to be clearly stated in the report.

The experiment should be reported using a computational notebook such as Jupyter, marimo, or Observable Notebooks. A ZIP file should be submitted, containing:

- notebook source files (*.ipynb, *.py, etc.),
- data files, if any (e.g., *.csv),
- an *HTML (including images) or PDF export* of the notebook.

References

Abelson, H. (1986). *Lecture 1A: Overview and introduction to lisp [lecture transcript]*. MIT OpenCourseWare 6.001 Structure and Interpretation of Computer Programs. https://ocw.mit.edu/courses/6-001-structure-and-interpretation-of-computer-programs-spring-2005/resources/1a-overview-and-introduction-to-lisp/

Arvanitou, E.-M., Ampatzoglou, A., Chatzigeorgiou, A., & Carver, J. C. (2021). Software engineering practices for scientific software development: A systematic mapping study. *Journal of Systems and Software*, *172*, 915–929. https://doi.org/10.1016/j.jss.2020.110848

Begel, A., & Zimmermann, T. (2014). Analyze this! 145 questions for data scientists in software engineering. *Proceedings of the 36th International Conference on Software Engineering*, 12–23. https://doi.org/10.1145/2568225.2568233

Beller, M., Spruit, N., Spinellis, D., & Zaidman, A. (2018). On the dichotomy of debugging behavior among programmers. *Proceedings of the 40th International Conference on Software Engineering*, 572–583. https://doi.org/10.1145/3180155.3180175

Boehm, B. W., Elwell, J. F., Pyster, A. B., Stuckle, E. D., & Williams, R. D. (1982). The TRW software productivity system. *Proceedings of the 6th International Conference on Software Engineering*, 148–156. https://dl.acm.org/doi/10.5555/800254.807757

Burns, R. B. (2000). *Introduction to research methods* (4th ed.). SAGE Publications.

Carrera-Rivera, A., Ochoa, W., Larrinaga, F., & Lasa, G. (2022). How-to conduct a systematic literature review: A quick guide for computer science research. *MethodsX*, *9*, 101895. https://doi.org/10.1016/j.mex.2022.101895

Carvalho, L., Degiovanni, R., Cordy, M., Aguirre, N., Le Traon, Y., & Papadakis, M. (2024). SpecBCFuzz: Fuzzing LTL solvers with boundary conditions. *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. https://doi.org/10.1145/3597503.3639087

Choudhuri, R., Liu, D., Steinmacher, I., Gerosa, M., & Sarma, A. (2024). How far are we? The triumphs and trials of generative AI in learning software engineering. *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. https://doi.org/10.1145/3597503.3639201

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage.

Denning, P. J. (2005). Is computer science science? *Commun. ACM*, *48*(4), 27–31. https://doi.org/10.1145/1053291.1053309

Easterbrook, S., Singer, J., Storey, M.-A., & Damian, D. (2008). Selecting empirical methods for software engineering research. In F. Shull, J. Singer, & D. I. K. Sjøberg (Eds.), *Guide*

to advanced empirical software engineering (pp. 285–311). Springer London. https://doi.org/10.1007/978-1-84800-044-5_11

Futatsugi, K., & Okada, K. (1982). A hierarchical structuring method for functional software systems. *Proceedings of the 6th International Conference on Software Engineering*, 393–402. https://dl.acm.org/doi/10.5555/800254.807782

Hall, T., Beecham, S., Bowes, D., Gray, D., & Counsell, S. (2012). A systematic literature review on fault prediction performance in software engineering. *IEEE Transactions on Software Engineering*, *38*(6), 1276–1304. https://doi.org/10.1109/TSE.2011.103

Hoda, R., Noble, J., & Marshall, S. (2013). Self-organizing roles on agile software development teams. *IEEE Transactions on Software Engineering*, *39*(3), 422–444. https://doi.org/10.1109/TSE.2012.30

Huang, Y., Wang, J., Liu, Z., Wang, Y., Wang, S., Chen, C., Hu, Y., & Wang, Q. (2024). CrashTranslator: Automatically reproducing mobile application crashes directly from stack trace. *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. https://doi.org/10.1145/3597503.3623298

Huijgens, H., Rastogi, A., Mulders, E., Gousios, G., & Deursen, A. van. (2020). Questions for data scientists in software engineering: A replication. *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 568–579. https://doi.org/10.1145/3368089.3409717

Inal, Y., Clemmensen, T., Rajanen, D., Iivari, N., Rizvanoglu, K., & Sivaji, A. (2020). Positive developments but challenges still ahead: A survey study on UX professionals' work practices. *J. Usability Studies*, *15*(4), 210–246.

Inayat, I., Salim, S. S., Marczak, S., Daneva, M., & Shamshirband, S. (2015). A systematic literature review on agile requirements engineering practices and challenges. *Computers in Human Behavior*, *51*, 915–929. https://doi.org/10.1016/j.chb.2014.10.046

Kampenes, V. B., Dybå, T., Hannay, J. E., & Sjøberg, D. I. K. (2007). A systematic review of effect size in software engineering experiments. *Information and Software Technology*, *49*(11), 1073–1086. https://doi.org/10.1016/j.infsof.2007.02.015

Keshav, S. (2007). How to read a paper. *SIGCOMM Comput. Commun. Rev.*, *37*(3), 83–84. https://doi.org/10.1145/1273445.1273458

Kitchenham, B. A., Dyba, T., & Jorgensen, M. (2004). Evidence-based software engineering. *Software Engineering, 2004. ICSE 2004. Proceedings. 26th International Conference on*, 273–281. https://doi.org/10.1109/ICSE.2004.1317449

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering* (Technical Report EBSE-2007-01). Keele University; Durham University Joint Report. https://legacyfileshare.elsevier.com/promis_misc/525444systematicreviewsguide.pdf

Lawrance, J., Bogart, C., Burnett, M., Bellamy, R., Rector, K., & Fleming, S. D. (2013). How programmers debug, revisited: An information foraging theory perspective. *IEEE Transactions on Software Engineering*, *39*(2), 197–215. https://doi.org/10.1109/TSE.2010.111

Miao, X., Wu, Y., Chen, L., Gao, Y., & Yin, J. (2023). An experimental survey of missing data imputation algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *35*(7),

6630–6650. https://doi.org/10.1109/TKDE.2022.3186498

Nakamoto, Y., Iwamoto, T., Hori, M., Hagihara, K., & Tokura, N. (1982). An editor for documentation in -system to support software development and maintenance. *Proceedings of the 6th International Conference on Software Engineering*, 330–339. https://dl.acm.org/doi/10.5555/800254.807775

OECD. (2015). *Frascati manual 2015: Guidelines for collecting and reporting data on research and experimental development* (p. 398). OECD Publishing. https://doi.org/10.1787/9789264239012-en

Rothlisberger, D., Harry, M., Binder, W., Moret, P., Ansaloni, D., Villazon, A., & Nierstrasz, O. (2012). Exploiting dynamic information in IDEs improves speed and correctness of software maintenance tasks. *IEEE Transactions on Software Engineering*, *38*(3), 579–591. https://doi.org/10.1109/TSE.2011.42

Shahin, M., Liang, P., & Babar, M. A. (2014). A systematic review of software architecture visualization techniques. *Journal of Systems and Software*, *94*(Supplement C), 161–185. https://doi.org/10.1016/j.jss.2014.03.071

Shreeve, B., Gralha, C., Rashid, A., Araújo, J., & Goulão, M. (2023). Making sense of the unknown: How managers make cyber security decisions. *ACM Trans. Softw. Eng. Methodol.*, *32*(4). https://doi.org/10.1145/3548682

Steimann, F. (2018). Fatal abstraction. *Proceedings of the 2018 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, 125–130. https://doi.org/10.1145/3276954.3276966

Stol, K.-J., & Fitzgerald, B. (2018). The ABC of software engineering research. *ACM Trans. Softw. Eng. Methodol.*, *27*(3). https://doi.org/10.1145/3241743

Wobbrock, J. O., & Kientz, J. A. (2016). Research contributions in human-computer interaction. *Interactions*, *23*(3), 38–44. https://doi.org/10.1145/2907069

Wohlin, C., & Aurum, A. (2015). Towards a decision-making structure for selecting a research design in empirical software engineering. *Empirical Softw. Engg.*, *20*(6), 1427–1455. https://doi.org/10.1007/s10664-014-9319-7

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2024). Systematic literature studies. In *Experimentation in software engineering* (2nd ed., pp. 51–63). Springer. https://doi.org/10.1007/978-3-662-69306-3_4

Zeller, A., & Lütkehaus, D. (1996). DDD—a free graphical front-end for UNIX debuggers. *SIGPLAN Not.*, *31*(1), 22–27. https://doi.org/10.1145/249094.249108