Is it a case study? – A critical analysis and guidance

Claes Wohlin[a], Austen Rainer[b]

[a]*Blekinge Institute of Technology, SE-371 79 Karlskrona, Sweden*
[b]*Queen's University Belfast, 18 Malone Road, Computer Science Building, BT9 5BN, Belfast, Northern Ireland, UK*

Abstract

The term "case study" is not used consistently when describing studies and, most importantly, is not used according to the established definitions. Given the misuse of the term "case study", we critically analyse articles that cite case study guidelines and report case studies. We find that only about 50% of the studies labelled "case study" are correctly labelled, and about 40% of studies labelled "case study" are actually better understood as "small–scale evaluations". Based on our experiences conducting the analysis, we formulate support for ensuring and assuring the correct labelling of case studies. We develop a checklist and a self–assessment scheme. The checklist is intended to complement existing definitions and to encourage researchers to use the term "case study" correctly. The self–assessment scheme is intended to help the researcher identify when their empirical study is a "small-scale evaluation" and, again, encourages researchers to label their studies correctly. Finally, we develop and evaluate a smell indicator to automatically suggest when a *reported* case study may not actually be a case study. These three instruments have been developed to help ensure and assure that only those studies that are actually case studies are labelled as "case study".

Keywords: case study, guidelines, citation analysis, small–scale evaluation, checklist, smell indicator

1. Introduction

The term "case study" is still not consistently used in software engineering research. For example, in a recent short communication, Wohlin (2021) classified 100 articles that were reported as case studies and found that close to

half of those articles were not case studies according to the established definitions. Similar findings were reported over 15 years ago. Zannier et al. (2006) conclude, from a study of articles published at the International Conference on Software Engineering, that, "...our sample indicated a large misuse of the term case study." Runeson and Höst (2009) reached a similar conclusion: "...the presented studies range from very ambitious and well organised studies in the field, to small toy examples that claim to be case studies." Despite the problem being pointed out by several authors, and over many years, the problem of the misuse of the term "case study" persists.

Our initial motivation for this article was to extend the preliminary work of Wohlin (2021) as we wanted to more deeply investigate the extent of misuse (however unintentional) of the term "case study". We identified a dataset of 188 articles for analysis. Our analysis identified a number of misclassified studies. We organised these into several categories of misclassification. One category, which we label "small–scale evaluation", dominated the misclassifications. Our findings both corroborate previous investigations that approximately 50% of studies reported as case studies are *not* case studies according to the definitions, but also reveal a type of study – the "small–scale evaluation" – that is often mislabelled as "case study".

Given these findings, we then started to think about practical support that we might provide to researchers in their use of the term "case study". We started by developing a checklist for distinguishing a case study from other types of study. The checklist is not concerned with designing, conducting, or reporting a case study (cf. Runeson and Höst (2009)), but instead concerns determining whether a study is in fact a case study. The checklist is intended to be used as a self–checking instrument for authors or for assessment by, for example, reviewers and other readers. To complement the checklist, we also develop a self–assessment scheme and a smell indicator. The self–assessment scheme is intended to help the researcher identify whether their *empirical study* is a "small-scale evaluation". The smell indicator is intended to help researchers automatically check whether an article being *reported* as a case study is, in fact, *not* a case study.

Whilst we have not formally evaluated or validated the checklist or the scheme, we did want to illustrate the application of both. We therefore identified a small number of additional articles, independent of our original dataset, for those illustrations. Furthermore, we apply the checklist and the smell indicator to six articles identified as exemplars of case study by the Empirical Standards for Software Engineering Research (Ralph, 2021). Formally, the

Empirical Standards were initially developed by the ACM SIGSOFT Paper and Peer Review Quality Task Force, who released the first version of the standard in October 2020 (Ralph, 2021). In the absence of any alternative set of standards, these standards are referred to as "the Empirical Standards for Software Engineering Research" or simply "Empirical Standards."

The remainder of the article is structured as follows. Background, related work and some examples concerning the misuse of the term "case study" are presented in Section 2. Our research approach is described in Section 3. The subsequent five sections, i.e., Section 4 through Section 8, then each present our work relating to one or more of the research contributions. Section 4 focuses on the analysis of the articles citing the case study guidelines by Runeson and Höst (2009) and presents high–level results, including introducing the "small–scale evaluation". Section 5 presents our more detailed analysis of studies that were not actually case studies. Section 6 presents the checklist, intended to ensure that studies are correctly classified as case study. The self–assessment scheme is presented in Section 7. Section 8 then introduces our smell indicator, intended to assure that studies are correctly labelled as case study by suggesting studies that are more likely to *not* be case studies. Section 9 discusses the limitations to our research and, finally, Section 10 presents conclusions.

2. Background and related work

For this section, we begin, in Section 2.1 with a brief review of the history of case studies in software engineering, and of the definitions of case study. Then, in Section 2.2, we focus on the challenges of labelling studies correctly, considering research method terminology and misunderstandings concerning case study research. Having established background and related work, we then present and discuss examples of the misuse of the term "case study", in Section 2.3. We use these examples to illustrate our motivation for this article.

2.1. Background – Case study research

Case study research is common in many disciplines and a range of general guidelines have been published. The most well–known source in software engineering research is probably the book by Yin (2018) with the first edition published in 1984. As a precursor to the book, an article was published by Yin (1981), where some reflections on case study research are presented.

Case studies have become more common in software engineering and are primarily used as a way to study software development in the field. Some early guidelines for case study research were published by Kitchenham et al. (1995). New and more detailed guidelines have subsequently been published in software engineering by, for example, Runeson and Höst (2009), Verner et al. (2009) and Runeson et al. (2012).

Runeson and Höst (2009) do not provide a definition of case study of their own; instead they highlight several other definitions, including those provided by Benbasat et al. (1987), Robson (2002) and Yin (2003). Runeson and Höst (2009) do, however, stress the real world setting: "Case studies are by definition conducted in real world settings, and thus have a high degree of realism, mostly at the expense of the level of control."

We noted earlier that the most well–known source for case study research in software engineering is probably the book by Yin (2018). Yin (2003) provides the following definition for case study, which is used in the case study guidelines by Runeson and Höst (2009):

"A case study is an empirical inquiry that investigates a contemporary phenomenon within its real–life context, especially when the boundaries between phenomenon and context are not clearly evident."

Although not stated explicitly in his definition, one of Yin's principles of case study research is the use of multiple sources of evidence.

Runeson et al. (2012) do provide an explicit definition of case study for software engineering, basing their definition on definitions from other areas including the definitions by Yin (2003), Benbasat et al. (1987) and Robson (2002). The definition by Runeson et al. (2012) is as follows:

"Case study in software engineering is an empirical inquiry that draws on multiple sources of evidence to investigate one instance (or a small number of instances) of a contemporary software engineering phenomenon within its real–life context, especially when the boundary between phenomenon and context cannot be clearly specified."

Based on the definitions cited above, Wohlin (2021) proposed a revised definition of *case study*, intended to help clarify the need for more than a single data collection *method* for case studies, and not just multiple *sources* of evidence. Furthermore, Wohlin included a formulation to separate case

studies from action research. Wohlin's (2021) definition is:

"A case study is an empirical investigation of a case, using multiple data collection methods, to study a contemporary phenomenon in its real–life context, and with the investigator(s) not taking an active role in the case investigated."

The definition by Wohlin (2021) includes the following five essential components for a case study:

- Empirical investigation of a case,
- Multiple data collection methods,
- Contemporary phenomenon,
- Real–life context, and
- Investigator's role.

2.2. Related work

As recognised in Section 1, the term "case study" is often misused in software engineering research. The term "case study" may be misused for a study because no clear alternative term is available. A common example is the study of an instance of a phenomena: the *method* of study does not conform to the definition of, or the guidelines for, a case study, however as there is a single instance being studied, the instance is treated as a case leading to the use of the term "case study" (cf. (Wohlin, 2021)).

Baker et al. (1992) uses the term, "method slurring" to refer to situations where researchers "... blur distinctions between the various qualitative approaches and combine their methodological prescriptions eclectically." They take two research approaches, grounded theory and phenomenology, and contrast them in terms of role, sources of data, sampling, data collection and analysis, and validity. Similarly, Haslam (2016) presents two forms of "concept creep" in psychology, i.e., "horizontal" and "vertical" concept creep, and discusses several instances of these forms. With these two examples, i.e., Baker et al. (1992) and Haslam (2016), it is clear that the concern is not unique for software engineering. Within software engineering, activities often takes places in the context of software projects that seek to deliver

on some concept of requirements, we may understand "method slurring" or "concept creep" as a kind of "research method creep" or "research method label creep", i.e., in a similar way as scope creep in software projects. We define "research method label creep" as when a research method label is used beyond the boundaries originally set by an established definition of the research method label.

The misuse of the label is a different, though related problem, to the misuse of the method itself: a *method* may be used beyond its proper scope, whilst a *label* may be misused to describe a study. Later in this article, we show an example where a high–quality ethnographic study has been subsequently labelled by others (not the original authors) as a "case study". The example illustrates how others change a label in relation to how the study was originally presented. However, our main focus for this article is on the mislabelling of studies by the original authors.

There is also definitional confusion in terminology (Yin, 2003), and a lack of consensus on definitions. For example, on the one hand, Yin (2003) views ethnography and participant–observation as data collection *techniques*. He states that a definitional flaw has resulted in confusing these techniques with the case study research strategy. But, on the other hand, in the Empirical Standards for Software Engineering Research (Ralph, 2021), case study research subsumes ethnography, e.g., that ethnography is a particular type of case study. In their book on case study research in software engineering, Runeson et al. (2012) also subsume ethnography within case study research. Others, such as Sharp et al. (2016) and Easterbrook et al. (2008), argue for ethnography as a research method in its own right, and not as a data collection technique or a type of case study.

The different opinions concerning ethnography are not critical to our discussion, however our preference is to view ethnography separate from case study. Doing so is consistent with our separation of the case study method from action research. The Empirical Standards also separate action research from case study. Furthermore, the Empirical Standards separate grounded theory from case study, recommending the use of grounded theory when research questions are not formulated upfront, and data collection and analysis are interleaved. As a contrast, others (e.g. Allen (2003)) take the view that the main difference between case study and grounded theory is that in case study research, hypotheses are intended to be formulated upfront, while in grounded theory hypotheses are intended to emerge based on the data analysis. The Empirical Standards state that when a study intervenes in the

context then action research ought to be considered. The latter is aligned with the case study definition by Wohlin (2021).

The problem of labelling studies incorrectly does not only occur for case study research. In software engineering, Ayala et al. (2021) observe a similar problem with the term "experiment" in research related to mining software repositories. Another example, described by Stol et al. (2016), relates to the use of grounded theory in software engineering. According to Stol et al. (2016), several articles in software engineering claim to use grounded theory but they do not include all core characteristics of such a study.

In our view, it is essential to label empirical studies correctly to avoid misunderstandings. This requires three steps: 1) the research community should develop appropriate and clear definitions, 2) authors should cite which definition they use, and 3) authors should label their studies in a way that is consistent with the definitions they use. For case studies, the research community has already covered step 1: we have definitions. We therefore focus on steps 2 and 3 in this paper.

2.3. Motivational examples of the misuse of the term "case study"

Two common features relating to the misuse of the term "case study" are the lack of real–life context and the limited use of data collection. We consider these in turn.

The first common misuse of the term "case study" appears to relate to the *context* of the case, i.e., whether the case being studied actually occurs in a real–life context. Unfortunately, the definitions of case study research do not define what constitutes a real–life context. Whether a context is "real–life" depends on the research objective, e.g., a case study of a student programmer tackling a programming problem may be a legitimate real–life context if the objectives of the research are to better understand how students tackle programming problems. But this situation is not a real–life context when the objectives are to infer from the study of students the behaviour of professional programmers tackling a programming problem for a commercial software product.

To study a real–life context in software engineering research often requires the conduct of a study in close collaboration with industry, which implies observing, over time, some authentic professional environment, e.g., being present in a face–to–face office environment, or interviewing a distributed team with team members working remotely. The latter example may include open source projects, where contact with the development community

may be combined with collecting and analysing information generated in a project. In other words, a case study requires studying some ongoing development, which often includes direct contact with people in the project being studied. One example is the study by Llerena et al. (2019), where the authors investigate the use of usability techniques in four open source projects.

Examples of investigations that are not case studies include:

- "Harvesting" *historical* data, such as from one or more open source projects, does not make for a study a *case study*, e.g., because this is no longer the study of a contemporary phenomenon.
- Conducting a study in a laboratory environment is not a case study, even when "real" artifacts are used, or when professional practitioners are asked to participate in the laboratory study, as the study is not conducted in a real–life context.
- *Re-analysing* a previously conducted case study, potentially with some change, is also not a case study as it is no longer a study of a contemporary phenomenon.

The second common misuse relates to data collection. It seems that some researchers are assuming that multiple instances of *one* method of data collection, e.g., multiple interviews, constitutes multiple sources of evidence. The definition by Runeson et al. (2012) states that case studies should draw on evidence from multiple sources. This is also highlighted by Yin (2018). In the book by Yin (2018), the first core principal is that multiple sources of evidence should be used for triangulation purposes. But Yin (2018) refers to six main evidence sources: documentation, archival records, interviews, direct observation, participant-observation, and physical artifacts; other sources may also be relevant, such as films and pictures. These sources require different methods of data collection. Also, given that all research methods have limitations and, therefore, the triangulation of data from different sources is important, multiple *methods* of data collection are needed for case study research.

Overall, our examples indicate the importance of accessing people at the time of the real–life context and also of collecting information from multiple sources. Further examples of the misuse of the term "case study" are presented in Section 5 based on our analysis of articles citing the guidelines by Runeson and Höst (2009).

In summary, researchers need to become more aware of the appropriate labels to use for the research studies being conducted and reported, and use those labels appropriately. Here, our focus is on research method label creep of case study research within software engineering.

3. Research approach

In this section, we first explain how we identified a dataset of articles being reported as case studies and, following that, how we identified a sample of additional articles for illustration purposes. Then, we explain how we classified our dataset of articles. This is followed by a discussion concerning how we developed and illustrated our checklist, and how we subsequently assessed the Empirical Standards against that checklist. We then discuss how we formulated and illustrated the self–assessment scheme for "small-scale evaluations". Finally, we discuss how we developed and applied the case study smell indicator.

3.1. Identifying a dataset of articles reported as case studies

We chose to focus on articles citing the guidelines by Runeson and Höst (2009) instead of conducting a systematic literature review. The objective was not to identify all articles misusing the "case study" label. Our focus has been to highlight the problem of "research method *label* creep", or "method slurring" as expressed by Baker et al. (1992) and Stol et al. (2016). The primary motivation to focus on citations to the guidelines by Runeson and Höst (2009) (henceforth also referred to as R&H guidelines) is that we believe that studies citing the guidelines are more likely, or at least ought to be more likely, to adhere to established definitions of case study research.

To identify an appropriate dataset of articles to analyse, we focused on journal articles that reported as case studies *and that also* cited the case study guidelines of Runeson and Höst (2009). Journal articles usually have a more in–depth and iterative review process, compared to most conference papers, and therefore ought to be of higher quality, or at least provide more information on their research methodology. Also, by citing Runeson and Höst (2009), these studies should be aware of the definition and guidelines for case study research in software engineering.

We used the following procedure to search and select articles:

1. We used Scopus to conduct the initial searches. The choice of Scopus is motivated by the fact that Scopus has a good coverage of different

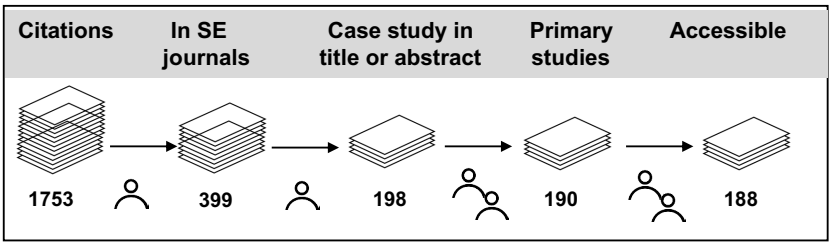

Figure 1: Search and selection process.

outlets (Dieste et al., 2009), including publications from many different publishers. Similar conclusions concerning Scopus have been reached by others, for example, by Garousi and Mäntylä (2016) and Mourão et al. (2020).

2. All searches were conducted by the first author.
3. We investigated all citations to the R&H guidelines in Scopus.
4. The inclusion criteria for full article analysis are: publication in software engineering journals, with the term "case study" present in the title or the abstract of the article.
5. The search for articles citing the case study guidelines was conducted on December 10, 2020.
6. The total number of citing articles retrieved was 1753. Including only articles published in software engineering *journals*, and excluding all others, retained 399 articles. Retaining only articles where the term "case study" was mentioned in the title or the abstract left 198 articles. A further 10 articles were subsequently removed during the analysis. Eight of these articles were methodological articles or SLRs. The other two articles were not accessible, i.e., behind paywalls. We therefore had 188 articles to analyse.

We note here that *none* of the six studies identified by the Empirical Standards (Ralph, 2021) as exemplars of case studies we retrieved in our search.

The process of search and selection is illustrated in Figure 1. A complete listing and classification of the 188 articles with links to the articles being assessed can be found through the link provided in Appendix A.

3.2. Identifying a sample of articles for use as illustrations

We did not formally evaluate or validate the checklist we developed, or the self–assessment scheme of "small-scale evaluations". But we did want to *illustrate* the checklist and the self–assessment scheme, and we wanted this illustration to be on a set of articles independent of the 188 articles we initially classified. We therefore used the following procedure to search for and select additional articles for the illustrations:

1. A search for articles published in 2021 and citing the guidelines by Runeson and Höst (2009) was conducted in Scopus for the Computer Science subject area, unfortunately only six articles were identified using the inclusion and exclusion criteria below. This was judged to be too few, hence we decided to do a search in Google Scholar instead. It is worth noting that articles published in 2021 were not included in the original search, since the original search was conducted in December 2020, hence there is no overlap with the 188 articles. Also, we searched both journal articles and conference papers.
2. We used the following inclusion criteria – the articles should:
   - relate to software engineering,
   - focus on research and not education,
   - be published in venues with peer–review, e.g., journals and conferences,
   - be published in English, and
   - include "case study" in the title.

   We also used the following exclusion criterion – the articles should *not* include:
   - more than one case study, since we wanted articles focusing on as single case study.

Our procedure identified 12 articles from 347 articles citing the guidelines by Runeson and Höst (2009), when the search was conducted on October 19, 2021. We ranked the 12 articles in the order they were found and then selected the six odd–ranked articles in the list to illustrate the application of the checklist, and used the six even–ranked articles in the list to illustrate the "small-scale evaluations". The decision to use odd and even numbered articles for two different purposes is equivalent to a random assignment,

since the order in which the articles appear is not related to the illustration objectives. A listing of the 12 articles and their classification can be found through the link provided in Appendix A.

3.3. Classifying the dataset of articles reported as case studies

We used the following procedure to analyse and classify the 188 articles:

1. Although we search for articles that cite the R&H guidelines, we use the Yin (2003) definition of "case study" and his principle concerning multiple data sources. We do this because Runeson and Höst (2009) did not provide an explicit definition in their guidelines, but instead cite the definition from the *third* edition of Yin's (2003) book.
2. We organised our analysis of articles into batches of 10 articles each.
3. We selected an initial sample of ten articles to pilot our classification. Essentially, we sorted the articles by journal and then took the first article in each journal. We independently classified all ten articles, then discussed all ten articles to agree on a classification. We concluded that a second pilot was warranted.
4. We selected a second sample of ten articles (second article from each journal) and again independently classified those, discussing all ten articles.
5. We then decided to proceed in samples of ten articles each, i.e., one author would classify a batch and, concurrently, the second author would classify the next batch in the sequence. Each author would then identify any articles that required independent review from their respective batches. Then we discussed the two batches (twenty articles in each round of analysis), resolving differences.
6. Each reviewer assessed sufficient parts of the full articles to feel confident in his judgement. If, after discussion, we considered that the study being reported was on the borderline of being or not being a case study, we accepted the study as a case study. An example of this situation is the study by Demirsoy and Petersen (2018). They refer to the conduct of a validation of their interviews, however they refer to this validation only in the discussion section of their paper, and not earlier in their paper, e.g., in the data collection or analysis sections. We accepted that some kind of validation occurred, and we view this as a second method of data collection hence more than one method of

data collection was used in their study. Thus, we accepted their self–classification of "case study". One implication of our heuristic (i.e., for borderline papers, to err on the side of "generousity") is that we may be over–reporting the number of actual case studies.

Section 4 presents the results of our analysis, and Section 5 discusses examples of studies that are reported as case studies but that are not actually case studies, according to our application of the definition of "case study".

3.4. Developing and illustrating the checklist

During our analysis of the first 60% of the articles (strictly speaking, the first 113 articles, i.e., our two pilot batches plus the next 10 matches in the dataset, less seven rejected articles), we each independently developed our own list of items for checking that a study is indeed a case study. Upon completion of the first 60% of the articles, we then discussed our respective lists, aggregating these items into one checklist.

Our aggregated checklist has not been formally evaluated, or validated, using the remaining 40% of articles (75 articles). We did not formally evaluate or validate the checklist because we wanted to continue with the same process to assess all 188 articles. The checklist is, however, consistent with the components identified in the case study definitions, as discussed in Section 2.1, and the checklist was *sense–checked* as we continued with the remaining articles in our dataset. The checklist is presented in Section 6.1.

As noted in Section 3.2, we identified six articles, independent of the 188 classified articles, to illustrate the application of the case study checklist, as described in Section 6.2.

3.5. Assessing the Empirical Standards' exemplars with the checklist

In addition to the articles that we use to illustrate the checklist, we also use the checklist to assess the exemplar articles for case studies, identified in the Empirical Standards (Ralph, 2021). We discuss the assessment of the Empirical Standard's exemplars in Section 6.3. A listing of the six articles and their classification can be found through the link provided in Appendix A. The six exemplar articles are also listed in the Empirical Standards (Ralph, 2021).

3.6. Formulating and illustrating the assessment of small–scale evaluations

Again, as noted in the preamble to this section, we noticed, as we assessed the 188 articles, that a majority of the misclassified studies were of a common type, a type we refer to here as "small–scale evaluations" (Robson, 2017). Thus, we saw a need to separately illustrate how these small–scale evaluations may be identified. Given this observation, we discussed aspects separating small–scale evaluations from case studies. Three core aspects were identified and provided the basis for formulating the self–assessment scheme for small–scale evaluations. The distinguishing aspects of a small–scale evaluation are discussed in Section 7.1.

To illustrate the assessment scheme, as explained above, we took the second six articles from a set of 12 additional articles, and used these to illustrate the identification of small–scale evaluations. Our intentions here are to highlight that small–scale evaluations, as a type of *empirical study*, can be misclassified as case studies, and to highlight why, or how, these misclassifications might occur. The assessment scheme is illustrated in Section 7.2.

3.7. Formulating a case study smell indicator

We also wondered, as the assessment of articles progressed, whether it might be possible to formulate an indicator for studies that may *not* be case studies. Our objective was to develop something similar in concept to a *code smell*. We emphasis here that the indicator is *only* intended to be used for studies being reported as case studies. Here, the indicator has been validated and evaluated against studies being classified as case studies and which also cite the R&H guidelines.

During our assessments of the R&H articles, we each independently developed a candidate *smell indicator*. Then, having completed the assessment of 113 articles, we compared the performance of the two indicators. We found that one smell indicator (described below) was both more powerful in its predictions and simpler in its design; but also "contentious", for reasons we explain in due course. We selected this indicator and subsequently evaluated it using the remaining 75 articles of our dataset. The smell indicator is also applied to the six exemplar case study articles listed in the Empirical Standards (Ralph, 2021).

4. Analysis of articles citing the Runeson and Höst guidelines

In this section we first present a summary of our analysis of the 188 articles being reported as case studies, and then summarise the types of misclassification. We discuss the misclassifications in more detail in the next section, Section 5, where we present examples of those misclassifications.

4.1. Summary of the correct classification of articles

Table 1 presents the high–level results from our analysis of the articles in the listed journals citing the guidelines by Runeson and Höst (2009). The table is organised and ranked by journal. For each journal, the table shows the number and percentage of articles presenting true case studies (see columns entitled *True case study*), according to the definition of case study by Yin (2003). For the four journals publishing the most case study articles, the percentage of true case study articles varies between about 40-70%. The percentages of true case studies published in the other journals varies substantially. One explanation for this variation is that the samples are small.

Table 1: Classification of studies citing the R&H guidelines and being reported as case studies.

Journal	**Total**	**True case study**	
		Count	**Percentage**
Information and Software Technology	57	27	47%
Journal of Systems and Software	45	18	40%
Empirical Software Engineering	21	12	57%
Journal of Software Evolution and Process	19	13	68%
Requirements Engineering	9	4	44%
Software Quality	9	4	44%
Software and Systems Modelling	6	2	33%
Software Practice and Experience	4	2	50%
Transactions on Software Engineering	4	1	25%
Automated Software Engineering	3	1	33%
e-Informatica Software Engineering	3	3	100%
IET Software	3	1	33%
Software Engineering and Knowledge Engineering	2	0	0%
Open Source Software and Processes	1	0	0%
Software Engineering and its Applications	1	1	100%
Transactions on Software Engineering and Methodology	1	1	100%
Total	188	90	

Overall, 90 (48%) articles are correctly classified as case studies though, as recognised earlier, this number might be over–reporting. In other words, less than 50% of the articles being classified as case studies are actually case studies and this percentage might be *even lower*. Our results are close

to those found previously by Wohlin (2021), who used a different dataset, where the number of correctly classified *case study articles* was 53%. The main difference between our dataset and Wohlin's (2021) is that we restrict our search only to journal articles that cite the guidelines by Runeson and Höst (2009).

4.2. A high–level breakdown of the misclassification of articles

Given that over 50% of the articles were *mis*classified, we looked more closely at how these non–case study articles should be classified. Table 2 breaks down the different types of study. The definition of these types is given below:

Table 2: Our classification of articles being reported as "case study".

Type of study	**R&H**	
	Total	**Percentage**
Case study	90	48%
Small–scale evaluation	72	38%
Interview study	14	7%
Archival analysis	9	5%
Action research	2	1%
Discourse analysis	1	0.5%
Survey	1	0.5%
Workshop or focus group	0	0%

- **Small–scale evaluation** is the most common type of study that has been misclassified as case study. According to Robson (2017), a small-scale evaluation comprises a single researcher or a small team, run over a short period, with limited resources, and most often occurring at a single site. Robson's (2017) description fits well with studies where researchers would like to illustrate, demonstrate, prove a concept, conduct a feasibility study or, in some other way, show the value of a research endeavour. Small–scale evaluations are closely related to engineering research where something novel is constructed and its (prospective) value needs to be assessed. These evaluations are essential for software engineering and software engineering research however such evaluations are not case studies.
  The term "small–scale evaluation" is focused on the objective of the study, which then may be conducted using different approaches, for example, simulation or benchmarking. This is well–aligned with case study research, which is defined in terms of the objective of conducting

a study of a contemporary phenomena in a real–life context, and then different data collection methods may be used, for example, interviews and document analysis.
A potential drawback with the term "small–scale evaluation" is that it may not be perceived as sufficiently "impressive". However, it is essential that we, as researchers, label our studies correctly and do not mislabel them to make them sound more substantive than they are. It also implies that the research community ought to see the real value of small–scale evaluation, e.g., as a natural and valuable stepping stone to propose new research solutions.

- **Other methods**

  - Interview study – A research method for obtaining data from individuals through direct interaction in person, either face–to–face or via some other media such as a phone call.
  - Archival analysis – A research method that obtains data through extraction from archives, such as email records or, potentially, code repositories.
  - Action research – A research method which, whilst sharing many similarities with the case study research method, focuses on active participation for change, i.e., the researcher is a change agent.
  - Discourse analysis – A research method for analysing the written or spoken language within its social context.
  - Survey – A research method for collecting data from a sample of items (often individuals) from (ideally) a predefined population to solicit opinions or experiences.
  - Workshop or focus group – A research method for obtaining data from people as a group of people, e.g., interview several people collectively.

Table 2 shows that the main challenge is to separate different types of small–scale evaluation from case studies. This is discussed in more detail in the next section.

5. Exemplification of misuse of the term case study

As indicated in the preceding section, we realised that there was a substantial subset of the articles that may be better understood as some kind of "research evaluation", rather than case study. These articles were not case studies. They were empirical studies of an *instance* but these instances did not occur in natural settings. In this section we review some of these variations. To illustrate how studies are incorrectly presented as case studies, some examples are taken from our analysis. When citing and discussing examples, our intention is not to criticise the study; our objective is only to give examples of the types of issues that we identified.

The variations we discuss map to Table 2 and are presented in the following sections. Section 5.1 discusses different forms of small–scale evaluation. Other types of studies not being case studies include: studies conducting archival analysis, as discussed in Section 5.2; using a single data collection method, most often only interviews, as discussed in Section 5.3; mining and analysing conversations from a forum or repository, which is further discussed in Section 5.4; studies being better described as action research, as discussed in Section 5.5; and one survey conducted within a company, as discussed in Section 5.6.

5.1. Small–scale evaluations

We found a number of articles that conducted some kind of empirical investigation but *not* in a real–life context. Instead the researchers use some kind of constructed environment or a simulated environment, or reuse an existing case study for new objectives.

Researchers often evaluate a technology (e.g., research method, technique, process, or tool) or some other research outcome, after having first built or adapted a suitable piece of software. For example, Arcaini et al. (2020) present a proof–of–concept by modelling a solution using open source software. In other cases, authors are explicit about addressing an industrial challenge, but then they build software for evaluating a potential solution for the problem identified in a laboratory setting, for example, as presented by Aleti (2015). Furthermore, authors conduct studies in a laboratory setting, but then discuss the study with practitioners. However, consulting or asking practitioners concerning a study is insufficient to make it into a case study. For example, Yang et al. (2017) asked software architects to use a framework developed through the research for a real and non–trivial project

from their work. However, the non-trivial project was not undertaken within the participants' everyday work environment; it was done as an extra activity to demonstrate and evaluate the framework. Thus, the study is not conducted in a real–life context, since it is an extra activity albeit with practitioners participating.

In other studies, authors use an existing case study and adapt it for their evaluation objective, for example, as presented by Pohlmann and Hüwe (2019). It is originally a case study, but it is adapted and hence the study is neither conducted in a real–life context, nor is it any longer the study of a contemporary phenomenon. In an article by Rainer (2010), the author reused a case study (previously conducted by himself) for a different purpose, but without making any changes to the case study. This article is classified here as a case study article, although the focus of the article is not to present the case study as such.

We found articles that report evaluation studies based on using data from available online resources. For example, Herbold et al. (2017) use three datasets from different sources to compare and evaluate global vs. local models for cost estimation. The first two datasets came from the tera-PROMISE repository, containing 35 and 12 projects respectively. The third dataset was also publicly available online, and contained five projects. If one treats each project as a case, then there is data on 52 cases. Furthermore, these cases have been removed from their contexts. The projects are not being investigated as contemporary cases in their contexts; they are used for evaluation purposes.

We also found articles that report studies using data from open source projects. Such studies may be acceptable as case studies, e.g., where the researchers conduct an investigation of an on–going open source project, as reported by Llerena et al. (2019). By contrast, Amanatidis et al. (2020) use fifty open source projects with two programming languages (Java and JavaScript) to illustrate and evaluate their proposed framework; again, these projects are not studied as cases in their contexts.

In a formal case study, researchers would typically collect information from, or about, people who naturally participate in the case being studied, e.g., software engineers, project managers, or potentially also students if education is the focus of the case study.

We found articles where the researchers used students or other kinds of *non–representative participants* in their studies. For example, students are not representative participants if the objective is to generalise to an industrial

context. However, they may be representative if it is a case study related to studying the educational environment as such. As one example, Razavian et al. (2016) use students to investigate design thinking, but formulate conclusions concerning software designers, which give the impression that the conclusions are valid for other kinds of participants than students.

5.2. Archival analysis: Non–contemporary phenomenon

We discussed the use of repositories earlier in relation to studies conducted for evaluation purposes. Another way in which repositories might be used is to analyse historical phenomenon. For example, Lehtinen et al. (2017) analyse 37 sprint retrospectives in a large, distributed and agile software development organisation. Thus, the study is done in an industrial setting. However, the study is not of a contemporary phenomenon, since the analysis is done on historical data. We recognise that this example, and this issue (of non–contemporary phenomenon) is something of an edge–case because one can report historical case studies. In software engineering, archival analysis is often labelled data mining or mining repositories.

5.3. Using only a single data collection method

As noted earlier, case study research is expected to use more than a single data collection method. In some studies, only a single data collection method is presented in the article, and hence it is not a case study. For example, Bjarnason et al. (2014) only describe interviews.

In other studies, researchers use a primary data collection method, for example, interviews. The primary method is complemented with one or more other methods, although the complement is very limited. For example, and as briefly discussed earlier, Demirsoy and Petersen (2018) use interviews and validation workshops with the main contact person at the company. The latter is not primarily for data collection, but it may be assumed that additional information is conveyed in the workshops. It should also be noted that interviews are mentioned in the title of the article, but in the article it is presented as a case study. Thus, the article is borderline to not being a case study. For the latter type of study, we have accepted it as a case study. However, it is essential that authors clearly describe the research methods used and how they are combined to form a case study.

In the article by Akarsu and Yilmaz (2020), the authors use a survey hence it is also an example of using a single data collection method. This article is further described below.

5.4. Discourse analysis: Archival and a single method

In one article (Wang and Redmiles, 2016), the authors use what is best described as discourse analysis. They analysed the development logs including both quantitative and qualitative data from two open source projects. They had a particular focus on non–work related conversations to understand how these helped cooperation and created trust in software development teams.

The study by Wang and Redmiles (2016) use a single data collection method and the data is collected from archives. Thus, the study is not done in a real–life context and only a single data collection method is used.

5.5. Action research: Investigator's role

The distinction between case study research and action research is not clear cut; the two methodologies share many aspects. The main difference relates to the role of the researchers. Action research focuses on setting research into action, and hence not primarily to create propositional knowledge to be used in other settings. Thus, action research focuses on the researchers being involved in a change process, i.e., also having some responsibility for the change. Further information concerning action research in a software engineering context are described by Staron (2020), and Wohlin and Runeson (2021). As an example concerning action research, Nashaat et al. (2019) describe how the researchers were treated as part of a stakeholder group. It gave the research team a possibility to integrate theory with practice and then continuously validate their hypotheses.

5.6. Survey: What is the case?

We identified one article, by Akarsu and Yilmaz (2020), where the authors conducted a survey within a company. Here, we use the word "survey" in relation to people, i.e., to solicit the opinion or knowledge of individuals. In this example, it is difficult to identify what the *case* actually is. The survey is conducted at one site within a company, but neither the whole company nor the site is being studied. The overall aim of the study is to explore the personality traits of software development teams. The 132 software practitioners surveyed are drawn from 20 teams within one site of one company. Thus, the research objective appears to concern teams, rather than individuals, sites or companies; but overall, the case is hard to identify and no specific case, within its context, appears to be studied.

The Akarsu and Yilmaz (2020) article also uses a second data collection method. The authors validated the survey through interviews with experts at

the company site. Despite the authors using more than a single data collection method, it does not make the study into a case study, since the case itself is unclear. The study might be better described as a survey complemented with validation interviews.

5.7. Summary

The examples presented in this section illustrate how the term "case study" is (unintentionally) misused in many different ways. Studies are labelled "case study" without taking the definitions of case study research sufficiently into account. Unfortunately, *case study* has become a collective term for studying an instance in some way. Or, as Wohlin (2021) phrased it, "It is a Case, and it is a Study, but is it a Case Study?" In subsequent sections of this article, we introduce novel contributions that are intended to support researchers in determining whether or not they are conducting a case study (according to the definitions).

6. Checklist for case study research

In this section, we first present a simple checklist to help researchers self–check whether they are designing a case study, or for reviewers, editors or readers to assess whether a study is a case study. We then apply the checklist to six articles to illustrate the checklist. Our selection of the six articles was explained in Section 3.4. Finally, we apply the checklist to all six papers identified by the Empirical Standards (Ralph, 2021) as exemplars of case studies.

6.1. Checklist

Given the identified issue with the usage of the term "case study", we propose a simple checklist to help researchers self–check whether they are designing a case study according to the definitions cited in Runeson and Höst (2009) and suggested by Wohlin (2021). The most appropriate point at which to consider the checklist is obviously during the design of the study itself. Thus, the objective is to consider each *study* being designed, rather than, for example, consider the design of a research article. Whilst the checklist is primarily intended to help researchers design their studies, the checklist may also help others to evaluate studies, e.g., when reviewing a manuscript submitted for publication.

The checklist consists of five steps, intended to be used as a flow chart. For conciseness, the checklist is *briefly* summarised in Table 3 however the full checklist – the one we encourage researchers to actually use – is presented through a link provided in Appendix A.

The five steps are formulated as closed questions with each question supported by a brief explanation. The order of the steps align with the five components of Wohlin's (2021) definition – the five components are identified in Section 2.1 – and are intended to be easy to answer. It is required that the answers to the questions in Steps 1–4 are "Yes", and "No" in Step 5 to qualify for being a case study. A different answer to a question in the sequence strongly indicates that the researcher will *not* be conducting a formal case study, or alternatively, that the researcher needs to redesign the study before it is conducted. Once a different answer, than those needed for a case study, has been encountered, the researcher should not need to proceed with answering the remaining questions.

The formulation concerning the study being an empirical investigation does not distinguish case studies from several other research methods. Thus, the focus is on the formulation of "the case" in the study. There is a need to be able to define the case, which is the first assessment criterion and it is formulated as a question.

In summary, the first four questions in the five steps should all be answered with "Yes", and the final question with "No" otherwise, the study is most likely not a case study. To qualify as a case study, an empirical study must clearly satisfy the five components. If researchers are not able to clearly satisfy all components, they are unlikely to be doing a case study, according to the established definitions of case study, and hence they ought to *not* label their research as case study research.

6.2. Illustration of the checklist

The six articles for illustration of the case study checklist were independently assessed by both authors. The first author used the checklist, whilst the second author assessed the six articles using the approach we used for the 188 articles. After our independent assessments, we were in agreement for four of the articles. For the remaining two articles (articles #5 and #7), we agreed after a short discussion. For these two articles, we decided to classify the articles based on the guidance of the checklist. These articles were borderline articles and, as per our approach in Section 3.1, we accepted the articles as case studies.

Table 3: Brief version – Case study checklist.

Step 1: Will/Did the empirical study investigate an identifiable case or case(s)?
☐ No or uncertain: the study is most likely not a case study. ☐ Yes: proceed to the next step. *Explanation*: For a study to be a case study it must be possible to define the case, for we need to be able to identify what it is we are studying. Describing the case can be difficult, because the boundary between the case and the context may not be clearly evident (Yin, 2018).
Step 2: Will/Was the study (be) conducted in a real–life context?
☐ No or uncertain: the study is most likely not a case study. ☐ Yes: proceed to the next step. *Explanation*: For a study to be a case study, it must be conducted in a real–life context. What constitutes "real–life" will depend on the "target population" to which the research findings are intended to apply.
Step 3: Will/Did the study use more than a single method of data collection?
☐ No or uncertain: the study is most likely not a case study. ☐ Yes: proceed to the next step. *Explanation*: For a study to be a case study it should use multiple methods of data collection so as to collect multiple sources of evidence and support triangulation.
Step 4 Will/Did the study investigate a contemporary phenomenon, i.e., a phenomenon occurring at the time of the study?
☐ No or uncertain: the study is probably a small–scale evaluation of an instance or archival analysis, depending on the research question. ☐ Yes: proceed to the next question.
Explanation: In general, contemporary refers to events occurring in the present. In a software development context, it is more realistic to understand "contemporary" to refer to a period of reasonably recent events for which it is possible to collect *new* information (e.g., through interviews) that can complement any other collected information, e.g., repositories of code or emails.
Step 5 Will/Did the researchers take active part in establish a long–term change beyond the duration of the research?
☐ Yes: the study is more likely to be action research, i.e., if at least one researcher is part of the decision–making team for change. ☐ No: the study is more likely to be a case study.
Explanation: In action research, one or more researchers are actively involved in implementing change based on the research, i.e., the researcher acts as a change agent.

Of the six articles, we found that three articles satisfy the components of case studies, as described in Section 2.1. The three articles are: Ljungberg et al. (2021) (article #5), Tsai (2021) (article #7) and Hussein and Zein (2021) (article #11). This is illustrated in Table 4 with a "Yes" on the four first case study components, and "No" on the fifth and final component. Furthermore, when an article fails on one component, the remaining components are not assessed. For the three articles that failed our checklist, we have the following reasons for not classifying the articles as case studies according to the definitions:

- Article #1 by Brataas et al. (2021): the study only uses a single data collection method, interviews, and hence it is not a case study. The

study fails on the third case study component, i.e., do the study use more than a single method of data collection? The study is best labelled as an interview study.

- Article #3 by Laverdière et al. (2021): the authors investigate the security evolution of two open source projects for web applications. Given that the authors investigate two applications, it becomes unclear what the case is, the article refers to one case study. Thus, it fails on the first case study component, does the empirical study investigate an identifiable case or case(s)? If the article would have referred to two cases, it would not fail on the first component. However, it is worth noting that even if studying one of the applications, the study fails on other case study components. For example, the study is not conducted in a real–life context. The study is best labelled as a small–scale evaluation.

- Article #9 by Sheikh Bahaei et al. (2021): the authors develop a framework for assessing risk of AR-equipped socio-technical systems. Through their contacts with a company they gain access to a system developed in industry. The authors use the system from the company to demonstrate or illustrate the modelling and analysis capabilities of their proposed framework. Hence, the researchers are not investigating a contemporary phenomenon in its real-world setting, but using that phenomenon to demonstrate their framework. This study is also best labelled as being a small–scale evaluation.

Table 4: Case study classification using our case study checklist.

	Case study components					
Article no.	**Case?**	**Real–life?**	**Data collection?**	**Contemporary?**	**Responsibility?**	**Case study?**
1	Yes	Yes	No			No
3	No					No
5	Yes	Yes	Yes	Yes	No	Yes
7	Yes	Yes	Yes	Yes	No	Yes
9	Yes	No				No
11	Yes	Yes	Yes	Yes	No	Yes

The main outcome of using the checklist is to decide whether or not a study qualifies as being a case study. The checklist is not intended to help with deciding which *other* type of study it may be.

6.3. Empirical Standards and the checklist

Concerning the Empirical Standards (Ralph, 2021), it starts from the definition of case study by Yin (2018). The standard highlights six essential attributes. The essential attributes are (Ralph, 2021):

- "justifies the selection of the case(s) or site(s) that was(were) studied,
- describes the site(s) in rich detail,
- reports the type of case study,
- describes data sources (e.g. participants' demographics and work roles),
- defines unit(s) of analysis or observation,
- presents a clear chain of evidence from observations to findings."

None of these are controversial as such, although the three latter attributes are valid for all empirical studies, and not unique for case study research. However, the attributes do not cover several essential elements of the definition of a case study. The real–world context and the contemporary nature of case study research are not covered. Furthermore, it does not highlight the need for more than a single data collection method, which is one of the four principles highlighted by Yin (2018). Use of more than a single data source and triangulation is listed as a desirable attribute, and not an essential attribute in the Empirical Standards. To further compare the standards' view on case study research with our assessment of case study research, we assess the exemplar articles listed in the standard with the five components of case study research listed in Section 2.2.

In the Empirical Standards (Ralph, 2021), six articles are listed as exemplars for case study and ethnographic research. It should be noted that case study subsumes ethnography according to the standard. Based on the case study checklist presented in Section 6.1, we assessed the six exemplars. An overview of the outcome is presented in Table 5. When we considered articles to be borderline of being or not being a case study, we accepted the study as a case study, as done in the previous classification, see Section 3.1.

The articles by Felderer and Ramler (2016) and Stol and Fitzgerald (2014) are both clearly case studies according to the definitions of case study research. Furthermore, both studies refer to the books by Yin (2003) and Runeson et al. (2012). Mockus et al. (2002) present studies of two open source

Table 5: Case study classification using our case study checklist.

Article	Case study components					Case study?
	Case?	**Real–life?**	**Data collection?**	**Contemporary?**	**Responsibility?**	
Alami and Wasowski	Yes	Yes	Yes	Yes	Yes	No
Felderer and Ramler	Yes	Yes	Yes	Yes	No	Yes
Mockus et al.	Yes	Yes	Yes	Yes	No	Yes
Sharp and Robinson	Yes	Yes	Yes	Yes	No	Yes
Spinellis and Avgeriou	Yes	Yes	Yes	No		No
Stol and Fitzgerald	Yes	Yes	Yes	Yes	No	Yes

projects, namely Apache and Mozilla. The study is primarily using archival data, but the authors also produce software development process descriptions which are reviewed by core team members of the two project. Thus, the results are triangulated, and the studies qualify to be labelled as case studies. The article does not refer to the case study literature, which is partially explained by its publication year.

The article by Sharp and Robinson (2004) is somewhat special. It is presented as an ethnographic study and there is no mentioning of case study in the article. However, if accepting that case study subsumes ethnographic studies then the article fulfils the definition of case study. However, it is interesting to note that Sharp et al. (2016) disagree with the view that ethnography is a special form of case study research, and then the ethnographic study by Sharp and Robinson (2004) is used as an exemplar article for case study research. Easterbrook et al. (2008) also view ethnography as a main research method on the same level as, for example, case study research and action research.

Finally, we have two (even) more problematic articles:

- The article by Alami and Wasowski (2019) is presented as being action research since both authors are active participants in the open–source project studied. They actively take part in improving the quality and quality assurance in the project. The authors do not mention case study research in the article and they do not refer to case study research literature. The article is better described as action research than case study research.

- Spinellis and Avgeriou (2019) present their work as being case study research. The article refers to the case study book by Runeson et al. (2012). Data is collected from analysis of source code, reference documentation, and related publications. The case is the Unix open source project. It is done in real–life, i.e., open source development. However, it is not contemporary given that the data collection is done from archival records. The authors write: "The case study as an empirical method is used for investigating a phenomenon in its real life context." However, this is only a partial view of the definition of a case study. It is not mentioned that it should be a study of a contemporary phenomenon. The study meets all components for case study research except for being contemporary. The study is best described as being archival analysis.

In summary, four out of six of the exemplar articles for case study research come out as being case studies using the checklist for case study research, including one article that is presented as being an ethnographic study. Unfortunately, two of the exemplar articles are better described as action research respectively archival analysis. The findings are disappointing given that it is an empirical standard, particularly in relation to the need to avoid research method label creep.

7. Small–scale evaluations

Given that small–scale evaluations are in many cases labelled as "case study" (see Table 2), we consider how small–scale evaluations can be differentiated from *field studies* more generally. Differentiating small–scale evaluations from field studies more generally also, by implication, differentiates the small–scale study from the case study and from other types of field study, such as the interview and action research. We illustrate this differentiation with six examples. We also show how small–scale evaluations relate to the level of control in relation to the following three aspects:

- Context,
- Case in terms of artefacts or data,
- Participants.

In Section 7.1 we first consider control. We then illustrate how small-scale evaluations differ from field studies in Section 7.2

7.1. Control in small–scale evaluations

In terms of levels of control, the level of control in a field study, for example a case study, is limited because the study investigates a contemporary phenomenon in a real–world context. When the researchers are able to exercise a higher level of control, this suggests the study is more likely to be an evaluation, rather than a field study. For example, in a small–scale evaluation the context is often within a more controlled context than a real–life context.

In terms of artefacts or data, where an artefact is moved from its natural context, this suggests the study is a small–scale evaluation. Similarly, where data is not collected in a real–life context, or is in some way removed from its natural context, this again suggests the study is a small–scale evaluation. For example, using historical data from a repository means that the study is not conducted in a contemporary, natural setting and studies that use that data are therefore more likely to be small–scale evaluations rather than case studies, or any other type of field study.

In terms of the participants, the study of participants who are doing their work in a natural context is more likely to be a field study, for example, a case study. Placing professional software engineers in an artificial or in some form of controlled setting, however realistic that setting may be, suggests the study being conducted is a small–scale evaluation.

To ease identification of small–scale evaluations, we propose a self–assessment scheme focusing on the three aspects listed above. The purpose of the self–assessment scheme is to help researchers decide whether they are conducting a small–scale evaluation or some other type of study. However, the assessment scheme does not specify which type of study the researchers will be conducting, for example, case study or interview.

7.2. Illustrations identification of small–scale evaluation

We assessed six articles to investigate whether the studies presented ought to be classified as small–scale evaluations rather than case studies. The authors assessed the articles independently, and were in agreement concerning all six articles without a need to come to a consensus. It is sufficient to have one "No" response in a row in Table 6 to most likely be a small–scale evaluation. The three studies by Bredahl Rasmussen et al. (2021) (article #4), Muñoz and Rodríguez (2021) (article #6) and Heeager and Nielsen (2020) (article #12) are not small–scale evaluations. The nature of this assessment means that we cannot determine what type of study they should

actually be, only that they are or are not likely to be small–scale evaluations. These three studies might be case studies or interview studies or some other form of field study.

Table 6: Small–scale evaluation assessment.

Article no.	Real–life?	Artefacts/data from studied context?	Regular work for participants or contributors?	Small–scale evaluation?
2	No	No	No	Yes
4	Yes	Yes	Yes	No
6	Yes	Yes	Yes	No
8	Partially	No	No and Yes	Yes
10	No	No	No	Yes
12	Yes	Yes	Yes	No

For the other three articles, we conclude that they are small–scale evaluations, and therefore not case studies, for the following reasons:

- Article #2 by Rahad et al. (2021): The study uses code from several projects on GitHub in the research. Thus, the study is not done in a real–life context, since it uses existing code. Furthermore, the study does not investigate any specific context. Finally, there are no real participants in the study. The analysis is conducted by the researchers using the code from GitHub projects.

- Article #8 by Márquez et al. (2021): The researchers have developed and implemented a framework, which is then evaluated by stakeholders from industry. Although the evaluation involves practitioners, the practitioners are not evaluating the framework through their normal work in their natural setting. Thus, the evaluation is related to a real–life context, but not conducted in a real–life context.

- Article #10 by Hernández et al. (2021): The researchers evaluate and compare two tools using open source software. Thus, it is not conducted in a real–life context. Furthermore, it is not conducted in a specific context, and it is not the regular work of the researchers. From a study point of view, the article has much in common with article #2, by Rahad et al. (2021).

In summary, to evaluate these three aspects helps researchers to identify studies as more likely being small–scale evaluations and not, incorrectly, label them as case studies.

8. A smell indicator for incorrect classifications of case studies

In this section, we propose and evaluate a simple indicator for suggesting when an article has been *incorrectly* classified as a case study. The indicator is therefore a kind of complement to the checklist. The checklist primarily helps researchers to prospectively check whether they are correctly designing and reporting a case study (though the checklist may also be applied retrospectively, as we have done in Section 6.2), whilst the indicator helps researchers to *retrospectively* check whether an article has been *incorrectly* classified. We evaluate the indicator's performance with a development dataset, an evaluation dataset and the Empirical Standards (Ralph, 2021). Because the indicator is counter–intuitive, we also discuss how the indicator, and its performance, should be interpreted. We discussed the development of the indicator in Section 3.7. Here we discuss the application of the indicator.

8.1. Explanation of the smell indicator

The smell indicator is intended to indicate articles that *incorrectly* classify a study as a case study. These misclassifications may be understood as "false case studies". It is essential to highlight that the smell indicator should only be used on studies that are already classified as case studies.

The indicator we selected is essentially based on a simple rule:

```
If the main body of the article reporting the ‘‘case study’’
has less than three occurrences of the word ‘‘interview’’
then the article is NOT reporting a case study.
```

We limit our rule to the main body of the article as the Reference section of an article may contain references that include the word "interview" in their titles.

The rule was implemented in a simple Python script which was then applied to the first 113 articles (the development dataset), and subsequently applied to the remaining 75 articles (the evaluation dataset).

For all included articles (i.e., the 113 articles and later the 75 articles) the following algorithm was used to classify articles using the smell indicator:

1. Download the PDF of the article.
2. Extract the text from the PDF of the article.
3. Remove all text following the Reference heading of the article.

4. Count the number of literal occurrences of the word "interview" in the remaining text.
5. Apply the indicator rule to the count, and label the article accordingly.

8.2. Performance of the indicator with the development and evaluation datasets

Having developed an indicator, we wanted to assess the performance of the indicator. Our objective for the indicator is for it to identify articles where the authors *incorrectly* label their studies as case studies.

To assess the indicator, we compared its performance against our own classification of each article. We did this for the development dataset (i.e., the first 113 articles we analysed, from which we first developed the indicator) and then for the evaluation dataset (i.e., the subsequent 75 articles we assessed).

Table 7 presents the confusion matrix for the performance of the indicator against the development dataset and Table 8 presents the confusion matrix for the performance of the indicator against the evaluation dataset. Because of our objective for the indicator, precision and recall are not suitable measures of performance for the evaluation. We therefore use accuracy instead.

Overall the accuracy measure suggests a reasonably effective indicator.

Table 7: Confusion matrix for the development dataset, with performance measure.

Confusion matrix:		
	Smell indicator suggests the study is	
We think the study is	**Case study**	**Not case study**
Case study	46	**10**
Not case study	16	41

Performance measure:	
Accuracy	$(46 + 41)/(46 + 41 + 16 + 10) = 0.77$

Table 8: Confusion matrix for the evaluation dataset, with the performance measure.

Confusion matrix:		
	Smell indicator suggests the study is	
We think the study is	**Case study**	**Not case study**
Case study	19	**14**
Not case study	10	32

Performance measure:	
Accuracy	$(19 + 32)/(19 + 32 + 10 + 14) = 0.68$

8.3. Applying the indicator on the exemplar studies in the Empirical Standards

We applied the rule for the smell indicator to the six exemplar studies presented in the Empirical Standards (Ralph, 2021). The assessment of the articles using the case study checklist is presented in Section 6.3. A listing of the six articles and their classification can be found through the link provided in Appendix A. The outcome of the application of the smell indicator is as follows:

- Strictly speaking, the smell indicator should not be applied to the articles by Sharp and Robinson (2004) and Alami and Wasowski (2019) since these articles do not present their studies as case study, however the Empirical Standards have classified these two articles as case studies. Sharp and Robinson (2004) present their study as an ethnographic study, and Alami and Wasowski (2019) view their study as action research.

- Two studies are classified as being case studies by the indicator. These are the studies by Felderer and Ramler (2016) and Stol and Fitzgerald (2014). These two articles are clearly classified as case studies when applying the case study checklist in Section 6.1.

- Finally, two articles are not case studies according to the smell indicator. The two studies are those presented by Mockus et al. (2002) and Spinellis and Avgeriou (2019). It should be noted that the study by Mockus et al. (2002) primarily uses archival analysis, but given contacts with some core team members concerning development process descriptions, the article qualifies as being a case study. However, interviews are not used, hence the smell indicator indicates that it may not be a case study. The article by Spinellis and Avgeriou (2019) is indicated as not being a case study, which is consistent with the assessment when using the case study checklist.

Overall, the smell indicator is not applicable to two studies given that they do not present themselves as case studies. The indicator identifies the two studies most clearly being case studies. Finally, it indicates correctly that one article is not a case study, and then fails to identify the article by Mockus et al. (2002) as a case study since they did not use interviews.

8.4. Reflections on the indicator

We are surprised at the effectiveness of the rule, especially given its simplicity. The rule is so simple as to appear simplistic. However, Yin (2018) highlights the interview as one of the most essential sources of evidence in a case study. The interview demonstrates the importance of gathering information from people as they participate in the real–life context. We did evaluate an alternative indicator which is more aligned with the definition of case study, albeit more complex. However, the alternative indicator did not perform as well as the simpler indicator that is based only on the occurrence of the word "interview".

One obvious weakness with the rule is that, when interview studies are *presented* as case studies, the rule equates interview studies and case studies. Most case studies include the conduct of interviews as a method of *data collection*, however interview *studies* – in which interviews are, usually, the single method of data collection – by definition exclude other methods of data collection. Given this weakness to our indicator, the results from applying the indicator need to then be complemented by a subsequent analysis to distinguish actual case studies from actual interview studies. A second weakness of the indicator is that the indicator does not recognise case studies that do not use interviews.

But, on the other hand, the rule exploits what appears to be a natural relationship between case studies and interview studies, i.e., that case studies and interview studies appear to be more similar to each other than they are to, for example, small–scale evaluations. Phrased another way, and informally, interview studies and case studies cluster together and cluster separately from small–scale evaluations.

8.5. Summary

Overall, our two evaluations suggest that the indicator is reasonably accurate, (i.e., with measures of accuracy of 0.68 and 0.77) at identifying articles that incorrectly report their studies as case studies. We have recognised two *weaknesses* in the indicator, i.e., that it equates interview studies, labelled as case studies, with case studies, and that it fails to identify case studies that do not include interviews. These *weaknesses* exploit the relative similarity of case studies and interview studies compared to small–scale evaluations. This similarity between case studies and interviews requires that a separate analysis of the data collection methods used in a study needs to be conducted to separate interview studies from case studies. Also, our development and

evaluation datasets for the indicator comprise journal articles that explicitly cite the Runeson and Höst (2009) guidelines and that are reported as being case studies, and the exemplars in the Empirical Standards (Ralph, 2021). Our indicator may not perform as effectively for other kinds of dataset.

9. Limitations

As with all research, our research comes with some limitations. First, we only used one database for generating our main dataset, i.e., Scopus. We could have used other databases too. However, our objective was to highlight a problem; hence, using multiple databases would most likely not change the outcome. There is no reason why articles in other databases should show a substantially different pattern. Furthermore, Scopus has good coverage of articles published by various publishers, and several authors highlight Scopus as being a good alternative (Dieste et al., 2009), (Garousi and Mäntylä, 2016) and (Mourão et al., 2020).

Second, it is essential to highlight that we did not conduct any quality assessment of the journal articles analysed. Our focus was on articles being reported as case study research. On the one hand, higher–quality case studies may be more likely to follow case study definitions. On the other hand, studies not being case studies may be of high quality, although misusing the case study label.

Third, we chose to primarily base our assessment on consensus discussions based on individual assessment. However, the individual assessments were foremost seen as a basis for our discussions; hence we have not calculated any agreement index for the assessment.

Finally, the research includes providing methodological guidelines, which is difficult to evaluate empirically, in particular by the researchers formulating the support. The support provided has been used, although its usefulness can only be argued based on it being based on one of the definitions of case study research.

10. Conclusions

Several authors, including Zannier et al. (2006), Runeson and Höst (2009) and Wohlin (2021) recognise that the term "case study" is misused in software engineering research. In the current article, we have further analysed the misuse of the term by investigating journal articles citing the case study

guidelines by Runeson and Höst (2009). We chose to investigate articles citing the guidelines because those articles ought to be more likely to follow the definitions for case study research than an arbitrary article presenting a "case study".

In our analysis of articles citing the case study guidelines by Runeson and Höst (2009), we find that more than 50% of the articles claiming to present case studies did not actually report a case study, according to the definitions. The percentage is consistent with the findings by Wohlin (2021), who collected a different dataset. Thus, we conclude that approximately half of the studies presented in software engineering research as case studies are not case studies according to the established definitions. Given the way we analysed the dataset, this statistic may be *over–reporting* the number of true case studies.

We also provide support for authors, reviewers, editors and readers to decide whether a specific study is a case study or not. We developed a case study checklist to help ensure or assure that a study actually is a case study. The checklist targets the five components identified in the definition of case study by Wohlin (2021), and helps the user of the checklist to decide whether or not a study is a case study. We intend for this checklist to be used primarily by researchers during the design of the study to ensure the study is a case study. Furthermore, we suggest an assessment scheme, primarily for authors, to identify if a study is a small–scale evaluation. The objective is to ensure that studies being small–scale evaluations are not mislabelled as, for example, case study.

Finally, we also developed a case study smell indicator, which can be used as a first indication of whether or not a reported study is in fact a case study. The development and the evaluation of the implemented indicator is based on the dataset analysed and presented in Section 4.

Further research includes, preferably, independent evaluation of the checklist, the assessment scheme and the smell indicator using other datasets. Based on these evaluations, we expect there will be the need to consider further improvements of the three instruments we present in this article. Furthermore, we also intend to address the limitations of the smell indicator, in particular the indicator's current inability to distinguish between true case studies and interview studies reported as being case studies, and true case studies not using interviews.

In summary, we corroborate claims about the misuse of the term "case study", and provide support for researchers to ensure and assure the proper

use of the term "case study". Unfortunately, our study illustrates that there is either a lack of awareness concerning the use of research method labels, or a lack of appreciation of the need to adhere to the definitions of research methods in software engineering.

Acknowledgement

We thank the anonymous reviewers for their time and effort. We appreciate their constructive feedback which helped us improve the article.

Appendix A. Supplementary material

The following is the supplementary material is related to this article.

S1: A listing of the articles classified (**Link to be provided by the publisher.**)

S2: The case study checklist (**Link to be provided by the publisher.**)

References

Akarsu, Z., Yilmaz, M., 2020. Managing the social aspects of software development ecosystems: An industrial case study on personality. Journal of Software: Evolution and Process 32, e2277.

Alami, A., Wasowski, A., 2019. Affiliated participation in open source communities, in: International Symposium on Empirical Software Engineering and Measurement, pp. 1–11.

Aleti, A., 2015. Designing automotive embedded systems with adaptive genetic algorithms. Automated Software Engineering 22, 199–240.

Allen, G., 2003. A critique of using grounded theory as a research method. Electronic Journal of Business Research Methods 2, 1–9.

Amanatidis, T., Mittas, N., Moschou, A., Chatzigeorgiou, A., Ampatzoglou, A., Angelis, L., 2020. Evaluating the agreement among technical debt measurement tools: building an empirical benchmark of technical debt liabilities. Empirical Software Engineering 25, 4161–4204.

Arcaini, P., Mirandola, R., Riccobene, E., Scandurra, P., 2020. Msl: A pattern language for engineering self-adaptive systems. Journal of Systems and Software 164, 110558.

Ayala, C., Turhan, B., Franch, X., Juristo, N., 2021. Use and misuse of the term experiment in mining software repositories research. IEEE Transactions on Software Engineering doi: http://doi.org/10.1109/TSE.2021.3113558.

Baker, C., Wuest, J., Stern, P.N., 1992. Method slurring: the grounded theory/phenomenology example. Journal of Advanced Nursing 17, 1355–1360.

Benbasat, I., Goldstein, D.K., Mead, M., 1987. The case research strategy in studies of information systems. MIS Quarterly 11, 369–386.

Bjarnason, E., Runeson, P., Borg, M., Unterkalmsteiner, M., Engström, E., Regnell, B., Sabaliauskaite, G., Loconsole, A., Gorschek, T., Feldt, R., 2014. Challenges and practices in aligning requirements with verification and validation: a case study of six companies. Empirical Software Engineering 19, 1809–1855.

Brataas, G., Martini, A., Hanssen, G.K., Ræder, G., 2021. Agile elicitation of scalability requirements for open systems: A case study. Journal of Systems and Software 182, 111064.

Bredahl Rasmussen, J., Haug, A., Shafiee, S., Hvam, L., Henrik Mortensen, N., Myrodia, A., 2021. The costs and benefits of multistage configuration: A framework and case study. Computers & Industrial Engineering 153, 107095.

Demirsoy, A., Petersen, K., 2018. Semantic knowledge management system to support software engineers: Implementation and static evaluation through interviews at Ericsson. e-Informatica Software Engineering Journal 12, 237–263.

Dieste, O., Grimán, A., Juristo, N., 2009. Developing search strategies for detecting relevant experiments. Empirical Software Engineering 14, 513–539.

Easterbrook, S., Singer, J., Storey, M.A., Damian, D., 2008. Selecting empirical methods for software engineering research, in: Shull, F., Singer, J., Sjøberg, D.I. (Eds.), Guide to advanced empirical software engineering, pp. 285–311.

Felderer, M., Ramler, R., 2016. Risk orientation in software testing processes of small and medium enterprises: an exploratory and comparative study. Software Quality Journal 24, 519–548.

Garousi, V., Mäntylä, M.V., 2016. Citations, research topics and active countries in software engineering: A bibliometrics study. Computer Science Review 19, 56–77.

Haslam, N., 2016. Concept creep: Psychology's expanding concepts of harm and pathology. Psychological Inquiry 27, 1–17.

Heeager, L.T., Nielsen, P.A., 2020. Meshing agile and plan-driven development in safety-critical software: a case study. Empirical Software Engineering 25, 1035–1062.

Herbold, S., Trautsch, A., Grabowski, J., 2017. Global vs. local models for cross-project defect prediction. Empirical Software Engineering 22, 1866–1902.

Hernández, C.M., Martínez, A., Quesada-López, C., Jenkins, M., 2021. Comparison of end-to-end testing tools for microservices: A case study, in: Rocha, Á., Ferrás, C., López-López, P.C., Guarda, T. (Eds.), Information Technology and Systems, pp. 407–416.

Hussein, M.A., Zein, S., 2021. Quadruple factors interference and its effects on quality of outsourcing testing: A case study. International Journal of Advanced Trends in Computer Science and Engineering 10, 228–236.

Kitchenham, B., Pickard, L., Pfleeger, S., 1995. Case studies for method and tool evaluation. IEEE Software 12, 52–62.

Laverdière, M.A., Julien, K., Merlo, E., 2021. RBAC protection-impacting changes identification: A case study of the security evolution of two PHP applications. Information and Software Technology 139, 106630.

Lehtinen, T.O.A., Itkonen, J., Lassenius, C., 2017. Recurring opinions or productive improvements—what agile teams actually discuss in retrospectives. Empirical Software Engineering 22, 2409–2452.

Ljungberg, A., Åkerman, D., Söderberg, E., Lundh, G., Sten, J., Church, L., 2021. Case study on data-driven deployment of program analysis on an open tools stack, in: Proceedings 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), pp. 208–217.

Llerena, L., Rodriguez, N., Castro, J.W., Acuña, S.T., 2019. Adapting usability techniques for application in open source software: A multiple case study. Information and Software Technology 107, 48–64.

Márquez, G., Taramasco, C., Astudillo, H., Zalc, V., Istrate, D., 2021. Involving stakeholders in the implementation of microservice-based systems: A case study in an ambient-assisted living system. IEEE Access 9, 9411–9428.

Mockus, A., Fielding, R.T., Herbsleb, J.D., 2002. Two case studies of open source software development: Apache and Mozilla. ACM Transactions on Software Engineering and Methodology 11, 309–346.

Mourão, E., Pimentel, J.F., Murta, L., Kalinowski, M., Mendes, E., Wohlin, C., 2020. On the performance of hybrid search strategies for systematic literature reviews in software engineering. Information and Software Technology 123, 106294.

Muñoz, M., Rodríguez, M.N., 2021. A guidance to implement or reinforce a DevOps approach in organizations: A case study. Journal of Software: Evolution and Process , e2342,doi:https://doi.org/10.1002/smr.2342.

Nashaat, M., Ghosh, A., Miller, J., Quader, S., Marston, C., 2019. Mlean: An end-to-end development framework for predictive models in b2b scenarios. Information and Software Technology 113, 131–145.

Pohlmann, U., Hüwe, M., 2019. Model-driven allocation engineering: specifying and solving constraints based on the example of automotive systems. Automated Software Engineering 26, 315–378.

Rahad, K., Badreddin, O., Mohsin Reza, S., 2021. The human in model-driven engineering loop: A case study on integrating handwritten code in model-driven engineering repositories. Software: Practice and Experience 51, 1308–1321.

Rainer, A., 2010. Representing the behaviour of software projects using multi-dimensional timelines. Information and Software Technology 52, 1217–1228.

Ralph, P., 2021. Empirical standards for software engineering research – case study and ethnography.
`https://acmsigsoft.github.io/EmpiricalStandards/docs/?standard=CaseStudy`. This is an online resource, edited by P. Ralph.

Razavian, M., Tang, A., Capilla, R., Lago, P., 2016. In two minds: how reflections influence software design thinking. Journal of Software: Evolution and Process 28, 394–426.

Robson, C., 2002. Real World Research : A Resource for Social Scientists and Practitioner-researchers. Wiley Publishing.

Robson, C., 2017. Small-scale evaluation: Principles and practice. Sage Publications.

Runeson, P., Höst, M., 2009. Guidelines for conducting and reporting case study research in software engineering. Empirical Software Engineering 14, 131.

Runeson, P., Höst, M., Rainer, A., Regnell, B., 2012. Case Study Research in Software Engineering: Guidelines and Examples. Wiley Publishing.

Sharp, H., Dittrich, Y., de Souza, C.R.B., 2016. The role of ethnographic studies in empirical software engineering. IEEE Transactions on Software Engineering 42, 786–804.

Sharp, H., Robinson, H., 2004. An ethnographic study of XP practice. Empirical Software Engineering 9, 353–375.

Sheikh Bahaei, S., Gallina, B., Vidović, M., 2021. A case study for risk assessment in ar-equipped socio-technical systems. Journal of Systems Architecture 119, 102250.

Spinellis, D., Avgeriou, P., 2019. Evolution of the Unix system architecture: an exploratory case study. IEEE Transactions on Software Engineering 47, 1134–1163.

Staron, M., 2020. Action Research in Software Engineering - Theory and Applications. Springer Nature.

Stol, K.J., Fitzgerald, B., 2014. Two's company, three's a crowd: a case study of crowdsourcing software development, in: Proceedings of the 36th International Conference on Software Engineering, pp. 187–198.

Stol, K.J., Ralph, P., Fitzgerald, B., 2016. Grounded theory in software engineering research: A critical review and guidelines, in: Proceedings of the 38th International Conference on Software Engineering, pp. 120–131.

Tsai, W.L., 2021. The impact of project teams on cmmi implementations: a case study from an organizational culture perspective. Systemic Practice and Action Research 34, 169–185.

Verner, J.M., Sampson, J., Tosic, V., Bakar, N.A.A., Kitchenham, B.A., 2009. Guidelines for industrially-based multiple case studies in software engineering, in: Proceedings Third International Conference on Research Challenges in Information Science, pp. 313–324.

Wang, Y., Redmiles, D., 2016. Cheap talk, cooperation, and trust in global software engineering. Empirical Software Engineering 21, 2233–2267.

Wohlin, C., 2021. Case study research in software engineering—it is a case, and it is a study, but is it a case study? Information and Software Technology 133, 106514.

Wohlin, C., Runeson, P., 2021. Guiding the selection of research methodology in industry–academia collaboration in software engineering. Information and Software Technology 140, 106678.

Yang, C., Liang, P., Avgeriou, P., Eliasson, U., Heldal, R., Pelliccione, P., Bi, T., 2017. An industrial case study on an architectural assumption documentation framework. Journal of Systems and Software 134, 190–210.

Yin, R.K., 1981. The case study crisis: Some answers. Administrative Science Quarterly 26, 58–65.

Yin, R.K., 2003. Case study research: Design and methods. 3rd ed., Sage Publications.

Yin, R.K., 2018. Case study research and applications: Design and methods. 6th ed., Sage Publications.

Zannier, C., Melnik, G., Maurer, F., 2006. On the success of empirical studies in the international conference on software engineering, in: Proceedings 28th International Conference on Software Engineering, pp. 341–350.