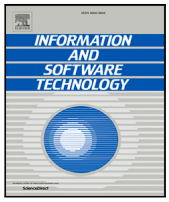
Information and Software Technology 164 (2023) 107318

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof



Revisiting the reproducibility of empirical software engineering studies based on data retrieved from development repositories

Jesus M. Gonzalez-Barahona *, Gregorio Robles

Escuela de Ingeniería de Fuenlabrada, Universidad Rey Juan Carlos, Spain

ARTICLE INFO

Keywords:
Reproducible research
Mining software repositories
Reproducibility
Validation studies
Empirical software engineering

ABSTRACT

Context: In 2012, our paper "On the reproducibility of empirical software engineering studies based on data retrieved from development repositories" was published. It proposed a method for assessing the reproducibility of studies based on mining software repositories (MSR studies). Since then, several approaches have happened with respect to the study of the reproducibility of this kind of studies.

Objective: To revisit the proposals of that paper, analyzing to which extent they remain valid, and how they relate to current initiatives and studies on reproducibility and validation of research results in empirical software engineering.

Method: We analyze the most relevant studies affecting assumptions or consequences of the approach of the original paper, and other initiatives related to the evaluation of replicability aspects of empirical software engineering studies. We compare the results of that analysis with the results of the original study, finding similarities and differences. We also run a reproducibility assessment study on current MSR papers. Based on the comparison, and the applicability of the method to current papers, we draw conclusions on the validity of the approach of the original paper.

Main lessons learned: The method proposed in the original paper is still valid, and compares well with other more recent methods. It matches the results of relevant studies on reproducibility, and a systematic comparison with them shows that our approach is aligned with their proposals. Our method has practical use, and complements well the current major initiatives on the review of reproducibility artifacts. As a side result, we learn that the reproducibility of MSR studies has improved during the last decade.

Vision: We propose to use our approach as a fundamental element of a more profound review of the reproducibility of MSR studies, and of the characterization of validation studies in this realm.

1. Introduction

In our paper [1] we proposed a method for assessing the reproducibility of empirical software engineering studies based on the analysis of data retrieved from software repositories (from now on "MSR studies"). That paper was based on an empirical study of the reproducibility of papers published in one of the main venues of the field, the International Conference on Mining Software Repositories, during the years 2004 to 2009 [2]. More than ten years later, we are revisiting both papers, reinterpreting their results in view of new approaches during this time, including a new analysis on the reproducibility of papers published in the same venue in 2023, and proposing some new ideas about how assessment of reproducibility could evolve.

Validation of research results by reproducing studies is an important part of the research process in science, and empirical software engineering is not an exception [3–5]. However, there is much concern about the quality of this kind of validation in the field, including the quality of the description of validation studies [6], and about the effort needed to produce validation studies due to poor support for reproducibility [7]. When we published our original study, these problems were already evident. We proposed to address them in the specific field of MSR studies by formalizing the assessment of their reproducibility aspects. For that, we characterized their reproducibility elements, and defined a process for producing reproducibility assessment reports based on the analysis of those elements. We expected that our proposal would make it easier to understand how reproducible a study is, because of how its methods (including software implementing those methods), datasets and other aspects are in this respect. Following this approach, we also characterized validation studies based on the elements they reused from the original study, and how they reused them.

In our approach, we were interested in *reproducibility with zero communication* between the research team performing the validation study, and the research team performing the original study [8]. We

* Corresponding author.
E-mail address: jesus.gonzalez.barahona@urjc.es (J.M. Gonzalez-Barahona).

https://doi.org/10.1016/j.infsof.2023.107318
Received 26 September 2022; Received in revised form 9 June 2023; Accepted 1 August 2023
Available online 22 August 2023

considered that this approach is the more convenient because of its transparency and consistency with the principles of science. In general, it is a problem if the research community needs to rely on private communication between researchers, exchanging behind the scene artifacts and description of processes that should be public in the published papers, or in their accompanying reproduction packages. This is also the approach used by some initiatives that have appeared during the last years, being the ACM Artifact Review and Badging [9] the most relevant of them in computing research.

The main contribution of this paper is the validation of the method presented in the original paper, for assessing on the reproducibility of MSR studies. The method includes a detailed and verifiable characterization of aspects influencing the reproducibility of MSR studies, based on the identification of elements with an impact on reproducibility, and the attributes determining the reproducibility of each of them. We also perform an analysis on 37 recent MSR papers, assessing their reproducibility, and comparing with the situation in 2004–2009. The structured reproducibility reports for all these papers can also act as examples of how the method can be applied to different kinds of MSR papers. The results of this study are a good proxy of the current state of reproducibility in MSR studies.

We expect that this paper helps to show that detailed assessment on reproducibility of MSR papers is possible, and beneficial both for the authors of the assessed papers, and for the MSR community at large. Since the effort in producing reproducibility reports is relatively low, once the method has been validated with current papers, it could be used for self-assessment, showing reviewers of a paper its reproducibility characteristics, but also during the review process, to improve reproducibility of the studies produced by the MSR research community.

We start the rest of this paper discussing some terminology issues in Section 2, because there is some confusion with different conventions for terms such as reproduction, replication or repetition, which is convenient to clarify before continuing. Then, in Section 3 we summarize our original study, and present its main results: the identification of reproducibility elements in MSR studies, the evaluation of their reproducibility attributes, the reproducibility assessment report based on them, and the characterization of validation studies. In Section 4 the most relevant related work is presented, with special attention to that which was still unpublished when we were writing our study. Section 5 offers a description of the empirical analysis of long papers presented in the International Conference on Mining Software Repositories (MSR 2023), for illustrating the current applicability of our proposal. After that, we discuss several aspects of our original proposal, how it could evolve in view of the approaches since it was published, and some ideas for the practical use of our results in current MSR studies, all of this in Section 6. Finally, in Section 7 we present some conclusions.

2. Terminology

Before entering into details, it is convenient to discuss two families of terms, because they may be confusing due to different uses, with different meanings, in the relevant literature. They are, on the one hand, *repetition*, *reproduction*, and *replication*, and on the other, *validation* and *reusable*. In this section we will present those different meanings, and also introduce which terms we will use in the rest of this paper. We will also discuss the differences between *experiment* and *empirical study*, since they are relevant for the rest of the paper.

2.1. Repetition, reproduction, replication

The use of the terms *repetition*, *reproduction*, and *replication* in the literature may be confusing, since they are used differently in different texts. In this paper, we adhere to the definitions in "ACM Artifact Review and Badging" [9][1] (from now on, "ACM Badges"), which we repeat here for convenience:

> Repeatability (same team, same experimental setup): "The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation".
>
> *Reproducibility* (different team, same experimental setup): "The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts".
>
> *Replicability* (different team, different experimental setup): "The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently".

These terms are defined following the International Vocabulary for Metrology [11]. A similar terminology is used in several recent studies on the matter, in software engineering and other computer science fields. For example, in software engineering, Shepperd et al. [6] highlight the difference between *reproducing* an experiment, which should be "as faithful as possible" to the original, and *replication*, which changes the experiment with the goal of "addressing confidence and generalisability". It also makes a difference between *internal replication* (with the replication team including authors of the original experiment) and *external replication* (replication run by an entirely different team). As another example, this one in machine learning, [12] defines *reproducibility* as "whether the reported experimental results can be obtained by other researchers using authors' artifacts [...] with the same experimental setup", and *replicability* as "whether the reported experimental result can be obtained by other researchers using their reimplemented artifacts with a different experimental setup". However, these two examples also show some differences with the ACM Badges terminology, which does not include the *internal replication* case, and uses a specific term for *internal reproduction* (*repetition*).

Earlier works use a somewhat different terminology. Considering three "classical" works on the matter (Basili et al. [3], Krein & Knutson [13], Gomez et al. [5]), all three use the term *replication* as a general term including repetition, replication, and reproduction (with the ACM Badges meaning), although then they make some differences according to how the setup changes with respect to the original experiment. When the setup does not change, both Basili et al. and Krein & Knutson use the term *strict replication*. Gomez et al. is more nuanced, using *literal replication* or *repetition* ("group I replication") when all details of the original experiment are kept, including the team running the experiment. When the setup of the experiment changes, Basili et al. uses the term *replication with a differential setup*. Krein & Knutson makes a difference between *differentiated replication*, *dependent replication* and *independent replication*, in growing order of variation with respect to the original experiment. Gomez et al. classify those (including cases when

[1] Unfortunately, version 1.0 of "Artifact Review and Badging" [10] used a swapped definition for *reproducibility* (different team, different experimental setup), and *replicability* (different team, same experimental setup), which caused some confusion.

Table 1
Different terminologies for replication, reproduction and repetition of software engineering experiments, according to changes in experimental setup and team: ACM Badges [9], Shepperd et al. [6], Basili et al. [3], Krein & Knutson [13], Gomez et al. [5].

	Same setup		Different setup		Different hypothesis
	Same team	Different team	Same team	Different team	
ACM Badges	**repetition**	**reproduction**	–	**replication**	–
Shepperd et al.	**reproduction**		internal **replication**	external **replication**	–
Basili et al.	strict **replication**		**replication** (different setup)		**replication** (different hypothesis)
Krein & Knutson	strict **replication**		differentiated **replication** dependent **replication** independent **replication**		–
Gomez et al.	literal **replication** (**repetition**) (group I)		operational **replication** (group II) conceptual **replication** (**reproduction**) (group III)		–

Table 2
Different terminologies for replication, reproduction and repetition of software engineering experiments, according to changes to the original experiment, ignoring changes to the experimental team: ACM Badges [9], Shepperd et al. [6], Basili et al. [3], Krein & Knutson [13], Gomez et al. [5].

	Same setup (exact)	Operational changes	Conceptual changes
ACM Badges	**repetition** **reproduction**	**replication**	
Shepperd et al.	**reproduction**	**replication**	
Basili et al.	strict **replication**	**replication** (different setup)	
Krein & Knutson	strict **replication**	differentiated **replication** dependent **replication**	independent **replication**
Gomez et al.	literal **replication** (**repetition**) (group I)	operational **replication** (group II)	conceptual **replication** (**reproduction**) (group III)

only the experimental team changed) as *operational replications* (“group II replications”) or *conceptual replication* (“group III replication”, or *reproduction*), also in growing order of variation. Basili et al. interested not only in classifying experiments, in experimental results in general, also mention the category *replication with a different hypothesis*, which would imply a complete redesign not only of the experiment, but of the research goals.

Therefore we can see how the term *reproduction* is used for experiments with conceptual changes with respect to the original (Gomez et al.) or for experiments with exactly the same setup of the original (ACM Badges and Shepperd et al.). The term *repetition* is the more consistent, used both in Gomez et al. and ACM Badges for all experiments with the same setup, including the experimental team. The term *replication* is used in some cases for all kinds of experiments (Basili et al. Krein & Knutson, and Gomez et al.), and in some others only for those with a different setup with respect to the original (ACM Badges and Shepperd et al.). A more nuanced review of other classifications found in studies about software engineering, and in other disciplines, can be found in [5], but for our purposes of showing the different terms and their uses, these examples are sufficient.

We have organized all these terms in Table 1, showing how they are classified in two dimensions: changes in the experimental setup, and changes in the team running the experiment. However, since the change in the experimental setup is not always a Boolean variable, we have also built Table 2, to show the same terms but now considering two types of changes: operational and conceptual. As explained in [5], there are many aspects of an experiment which can be changed. They identify several categories for aspects that could be changed (operationalization, population, protocol and experimenter), each with several possible variations. Therefore, “different setup” is closer to a continuum of options than to a discrete number of them. However, this table seems to be enough to show the relevant variations in terminology.

For our purposes in this paper, we are interested in studying the different aspects needed to facilitate, for different groups, performing experiments as similar as possible to the original one. Therefore, in ACM Badges terminology, we will study how to facilitate reproduction. It is fortunate that in our paper ten years ago we used the term “reproduction” with the same meaning we will be using in this one. In it, we stated: “In this paper we consider reproducibility as the ability of a study to be reproduced, in whole or in part, by an independent research team”.

That said, it is obvious that everything that a team makes to facilitate reproduction, will also facilitate their own repetition of the experiment (again, in ACM Badges terminology). If other teams are interested in running the original experiment with some variations (replication, in ACM Badges terminology), they will just select the elements facilitating reproduction that they want to keep, and modify others.

2.2. Validated, reusable

In some cases, we need a name for a kind of study that includes (in ACM Badges terminology) repetitions, reproductions and validations. Fortunately, ACM Badges terminology suggests a term exactly for this: *validation study*. This comes, in the ACM Badges classification of badges, from “Results validated”, a category which includes “Results Reproduced” and “Results Replicated”. Therefore, we find the name *validation* suitable for all kind of studies shown in Table 2.

A validation study may reuse some of the elements of the original study. ACM Badges defines the badge “Artifact Reusable” when “reuse and repurposing is facilitated”. Therefore, we can consider that *reuse* is applicable to elements that are reused as such from the original study, and that *reusable* is applicable to elements of an study that can be reused in further studies.

Our use of the term *validation* is similar to the use of the term *replication* in [3,5,13] (see Table 2). We preferred to use *validation*

instead to align with the ACM Badges terminology. With respect to the term *reuse* or *reusable*, although it is used in some texts of the papers mentioned in the rest of this section, we could not find anything close to a definition.

2.3. Experiment, empirical study

In addition, there is also some confusion on the use of the term *experiment* in MSR research. In many cases, the kind of studies presented in MSR papers do not have the characteristics needed to be considered as experiments, according to the definitions found in the empirical software engineering literature, and in the literature of other experimental fields [14]. This has important consequences with respect to their chances of finding causal relationships, but they are still empirical studies that can find other kinds of evidence.

The fact that many studies in the MSR field are not really experiments also affects the characteristics that are relevant when studying their reproducibility. For example, in some of them, there is a lack of manipulation or control, which means that there will be no description of them. This will be discussed later, when comparing with other studies on reproducibility which are more focused on experiments.

2.4. Use in this and the original paper

In our paper we adhere as much as possible to the ACM Badges terminology. These are the meanings we are using (along with a description of the terms used in our original paper):

Reproduction, reproducible We use the term *reproduction* to refer to a study which is in all relevant aspects identical to the original, and is performed by a different team. We use *reproducible* to refer to a study that can be reproduced. Thus, a study will be reproducible only if all relevant elements of the original study can be used for reproduction. We name those elements of a study that are relevant for reproduction as *reproducibility elements*. We will also use this term in *reproduction package*, which is a package with detailed descriptions and artifacts, intended to facilitate the reproduction of a study (this corresponds to a *replication package* in other works [8]).

In the original paper we used the terms *reproduction*, *reproducible*, and *reproducibility* in the same sense, but also, in some cases, in the sense of *validation* (see below). In it, we also used *replication package* instead of *reproduction package*.

Repetition, replication In general, we avoid these terms, because they are not needed for most parts of our paper and could cause some confusion. When used, *repetition* refers to a study with all relevant aspects identical to the original, performed by the same team, and *replication* to a study reusing only some elements of the original paper.

In the original paper we did not use these terms, except for *replication package* (see above).

Validation We use the term *validation* to refer to a kind of studies including repetitions, reproductions, and replications.

In the original paper we used the term *reproduction* instead of *validation* in some cases, usually noticeable by their context.

Reuse, reusable We use *reuse* to refer to the use of an element that was previously used or produced in the original study. Reusing an element may mean to actually use it as a part of performing a validation study (in the case of methods, parameters or datasets), or to use it for comparing final or intermediate results (in the case of datasets). *Reusable* will therefore refer to an element of a study that is suitable for reuse in a validation study. The term *reusable element* is stricter than *reproducibility element*: it implies not only that the element is relevant for reproduction, but that it can actually be used for reproduction.

In the original paper we did not make a difference between *reusable element* and *reproducibility element*.

Study In this paper we use the term *study* as a shorthand for *empirical study*, including not only experiments, but also other kinds of empirical studies such as observational studies. However, given the aims of the paper, we usually refer specifically to MSR studies.

In the original paper we also referred to *study* with this meaning.

3. Summary and main contributions (original paper)

In our original paper [1] we proposed a method for assessing the reproducibility of MSR studies. While presenting the method, we also identified a list of elements relevant when performing validation studies. For these elements, we identified a list of reproducibility attributes which could be used as a checklist, useful when building reproduction packages for those studies, and in general, for assessing the reproducibility of the study.

3.1. Main results

In summary, the main contributions of our original paper, all of them related to the reproducibility of MSR studies, were:

Reproducibility elements. We identify the elements of interest for the reproducibility of a study. These elements are: data sources, methods (retrieval, extraction, analysis), datasets (raw, processed, results), and study parameters. All of them are depicted in Fig. 1. We showed that these elements can be found in studies in the analyzed literature, including validation studies.

Reproducibility attributes. For each reproducibility element we identify several attributes that impact on reproducibility: identification, description, availability, persistence, and flexibility (see Table 3). Each attribute shows a dimension of how suitable the element is for the reproducibility of the study, affecting reproducibility in different ways. For example, availability of an element makes direct reuse possible (if it is available), or not (if not).

Reproducibility assessment. We propose an assessment on the reproducibility of a study, based on the reproducibility attributes of its reproducibility elements. Since these elements are determining which kinds of validation studies can be performed, and how difficult it will be to perform them, the reproducibility assessment is providing information about which kinds of validations can be done, and how difficult those validations will be. For example, if artifacts used for the extraction of data from the data sources are available and reusable, reproduction of the original procedure for data retrieval will require less effort than if artifacts have to be rewritten based on a description of the process. In addition, it will be very difficult to assess that an exact reproduction of the procedure was performed if the original artifacts are not available and usable.

Characterization of validation studies. We show how to analyze and classify validation studies according to which reproducibility elements they reuse, and how they are reused. An example of such characterization can be found in Table 6. In our paper we only identified, for each kind of validation study, the elements of the original study that were reused. In this table, we have also included which kind of reuse is intended, as we will discuss later (see Section 6).

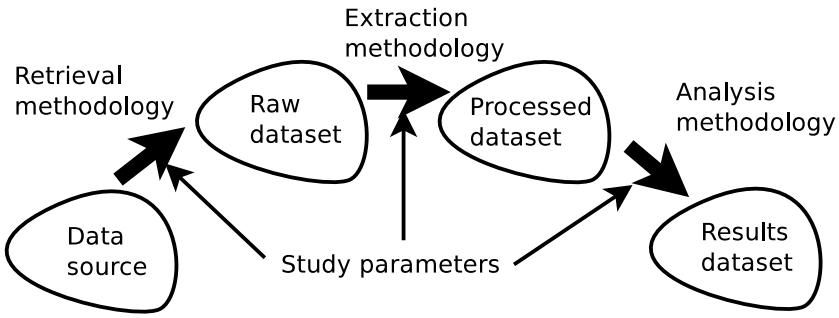


Fig. 1. Typical complete process for a study, as presented in [1], showing all elements with an impact on reproducibility.

Table 3
Attributes for each type of element (from the original paper).

	Data Source	Datasets	Parameters	Tools
Identification	X	X	X	X
Description	X	X	X	X
Availability	X	X		X
Persistence	X	X		X
Flexibility		X		X

Table 4
Assessment categories of elements with respect to their use in validation studies, and associated tags, suitable for their use in the reproducibility assessment report. Table constructed from descriptions in the original paper.

Category	Assessment	Tag
Usability	Usable for reproduction	U
	Usable with some difficulty	D
	Not usable for reproduction	N
	Irrelevant or non-existent	–
Persistence	Foreseeable persistent availability	+
Flexibility	Flexible usage	*

3.2. Reproducibility assessment

One of the main proposals of the original study is how to produce nuanced reproducibility assessment reports for characterizing the reproducibility characteristics of a study. This report is a detailed assessment on each of the elements with an impact on reproducibility, and is designed to be produced as a part of the review process, or maybe after it, in an independent reproducibility assessment process.

The reproducibility assessment for a study is based on the reproducibility attributes of each element. Those attributes are first determined for each element, and then evaluated according to their assessment category: usability, persistence, and flexibility. This assessment models how likely it is that the element will be useful in a validation study, either for being used as such, to reduce the effort in performing the new study, or to make the comparison with the original more exact. A list of assessments, along with a short tag used to represent them, is shown in Table 4. Some other assessment categories could be added to this list if they prove to be important for reproducibility.

Therefore, reviewers assessing the reproducibility of a study would start by identifying its reproducibility elements, and the artifacts associated to them. For each of the elements, they would determine the relevant attributes, and from them, they would produce the final reproducibility assessment report. An example of such report can be found in Table 5. If the process is followed, for example with papers in a journal, every published paper would show an assessment report for the study it presents.

3.3. Characterization of validation studies

Another proposal of the original paper is the characterization of validation studies. This is done by determining which elements of the original study were reused, and how. This characterization could be done as a part of the review process, or maybe as a self-characterization by the authors of the study. In any case, it provides details about what exactly can be validated with the study, and the elements from the original study that were reused, and with which purpose.

In the original paper, several kinds of validation studies were characterized, each with a different combination of reproducibility elements (see Table 6). This distinction allowed us to illustrate the differences between validation studies with different aims, such as repeating a study as similar as possible to the original one, to assess on its validity in the same conditions; changing data sources to search for more generality; or performing a different kind of analysis that proved to be more accurate.

4. Related work: new developments since the original paper

Our original paper was aimed at describing the relevant elements of an MSR study with impact on its reproducibility. Only if reproducibility elements were described with enough detail, the study could be validated. If key artifacts used in the production of the results were available, they could be used to lower the barrier for reproduction, and thus for validation of results. In this section we will review the most relevant related work to this approach, focusing on research and initiatives that were published after the publication of the original paper, and therefore are not mentioned in it, but also including some earlier references.

4.1. Steps of MSR studies

For identifying the relevant descriptions and key artifacts, we started by identifying the usual steps in MSR studies, each described as a method (see Fig. 1). If we map those methods to the phases in the experimentation process in empirical software engineering research [4,15] (scoping, planning, operation, analysis, interpretation, and presentation and packaging), we find that they correspond to the operation and analysis phases:

Operation phase. This phase correspond to the operation of the experiment, when the experiment is executed. In the case of MSR studies, there are two steps that are generally necessary in this respect: the retrieval of the data from the data sources, and the extraction of the data relevant for the study. Each of these steps will follow their collection of methods (retrieval and extraction method), and will produce their own data artifacts (the raw and processed datasets). Another relevant artifact is the data sources, which will be needed to perform full reproductions of the study. In this respect, it is important to notice that validation

Table 5
Example of reproducibility assessment report, with the final assessment for each of its reproducibility elements in the last column (table from the original study).

	Identification	Description	Availability	Persistence	Flexibility	Assessment
Data source	Partial	Detailed	Public	Likely	–	D+
Retrieval meth.	Partial	Source code	Public	Likely	Complete	D+*
Raw dataset	No	No	No	N/A	N/A	N
Extraction meth.	Partial	Source code	Public	Likely	Complete	D+*
Parameters	Complete	Complete	–	–	–	U
Processed dataset	No	No	No	N/A	N/A	N
Analysis meth.	No	Textual	No	N/A	N/A	N
Results dataset	No	No	No	N/A	N/A	N

Table 6
Characterization of several cases of validation studies, according to their use of reproducibility elements. U: element used for performing the reproduction study, O: element which may be optionally used for performing the reproduction study, C: element suitable for comparing results with the reproduced study, X: element which remains equal than in the reproduced study, but is not directly used in the reproduction study, -: element not relevant for the reproduction. Table built from information in the original study.

	Data source	Datasets			Methods			Parameters
		Raw	Processed	Results	Retrieval	Extraction	Analysis	
New study	–	–	–	–	–	–	–	–
Procedural validation	U	C	C	C	U	U	U	U
Analysis of processed dataset	X	X	U	C	X	X	O	O
Analysis of raw dataset	X	U	C	C	–	–	–	–
Reusing retrieval tools	O	–	–	–	U	O	O	O

studies may start with the raw dataset or even with the processed dataset, therefore not performing the whole experiment by retrieving data from the data sources (see our discussion on the terms *experiment* and *study* in Section 2.3).

Analysis phase. This phase corresponds to the analysis of the experiment, when statistical or other methods are applied to the data obtained after the extraction methods, to get statistical summaries, visualizations, etc. of it, which will be later interpreted. In the case of MSR studies, we also found this step, which we characterized as the analysis method. It starts from the processed dataset, and produces the results dataset. It is important to notice that, for a good evaluation of the results of a validation study, we will need details about the results of the analysis phase, and thus of the results dataset. In some cases, to validate the analysis itself, we may also need access to the processed dataset so that, for example, the analysis is repeated on it following a different method, and then both results datasets are compared.

However, when comparing our three steps (the method elements) with the phases described in [4] (below referred as *Phases in SE experiments*), three considerations arise:

Different levels of detail. The analysis of Phases in SE experiments is general enough to consider all kinds of experiments, including (and to some extent, even giving a preferential consideration to) experiments performed on human subjects. On the contrary, we were focused on MSR studies, which do not consider the specific case of experimenting with humans. MSR studies may deal with human behavior, or even with humans as individuals that affect software engineering processes, but they do not usually deal directly with humans in their experiments. MSR studies typically start by obtaining data from one or more data sources. Even when this data may be related to humans, humans are not directly involved. This means that we can be more specific in the identification of the steps, determining that the operation phase is in fact split in two very different steps (retrieval and extraction), each with their input and output datasets, since all MSR studies have some similarities. Maybe there is also the influence of the moment of publication: in 2012, MSR studies were still relatively novel, and were only starting to be considered as an important part of empirical software engineering research.

Experiments versus studies. Phases in SE experiments are referred to *experiments*. But our analysis was focused on empirical *studies* in general, which include, for example, observational studies. In fact, many typical MSR studies do not in fact perform an experiment [14], as defined in the literature on the matter [16]. This means that the kind of studies that we considered was different. In particular, we had into account studies that do not really include an empirical experiment, be that because there is only observation of events that happened (observational study), or because they are validations (maybe with different methods) that start from collected data. What our studies have in common is not their status as experiments, but their use of data sources about software development as the origin of the research. Therefore, they are at the same time more general, and more concrete. And it is this concreteness that helped us to identify their usual steps, and the intermediary datasets that they produce.

Different scopes. In our paper, we were concerned only with the description of the operational part of studies (operation and analysis), with the aims of analyzing how they could be reproduced. We understood that other phases (scoping, planning, or interpretation) do not have a direct influence on reproducibility, and therefore were out of scope for us. Looking at this from the future, maybe we could have identified other elements with influence on reproducibility. For example, the interpretation phase can be performed in part with data, producing summaries and analyzing correlations, which could be included as a separate step in our description of a study. However, we then considered that those other aspects could be included in the three identified steps, and in typical MSR studies we still think that is the case. That said, the only way of ensuring this is still right would be to repeat our study on more recent MSR literature, to check if our three steps are still relevant for all, or at least a very large fraction, of studies.

We can also compare our steps to those of the traditional KDD (from Knowledge Discovery in Databases) method for extracting knowledge from large datasets [17], in which our study was inspired. In KDD, data is first selected from the original data source, and then preprocessed and transformed. This transformed data is mined to find patterns, which are later interpreted to obtain knowledge. In our case, retrieval is similar to data selection in KDD, extraction is similar to preprocessing and transformation, and finally analysis is similar to mining. The

Table 7
Steps of MSR studies according to our original study, [18] ("Data Mining SE"), and [19] ("MSR Cookbook").

Our original study	Data Mining SE	MSR Cookbook
Retrieval method	Collecting/investigating data Determining SE task	Data acquisition & preparation
Extraction method	Preprocessing	
Analysis method	Adopting/adapting/developing algorithm	Synthesis
	Postprocessing/mining results	Analysis & interpretation
–	–	Sharing & replication

interpretation is out of scope for us, since it would correspond to the interpretation of the results of the study, which do not have an impact on reproducibility.

Other works, published in years close to our original paper, have similar descriptions of the typical process in MSR studies (see Table 7). [18] identifies five steps: collecting/investigating data to mine and determining the software engineering task to assist (steps 1 and 2); preprocessing data (step 3); adopting/adapting/developing a mining algorithm (step 4); and postprocessing/applying mining results (step 5). Of these, steps 1 and 2 broadly correspond to our "retrieval method", step 3 corresponds to our extraction method, producing the processed dataset, and steps 4 and 5 correspond to our analysis method, producing the results dataset. Hemmati et al. [19] identify 4 steps: data acquisition and preparation, dealing with data extraction and data modeling (step 1); synthesis, applying a mining/learning technique (step 2); analysis and interpretation of the results (step 3); sharing and replication (step 4). Of these, step 1 corresponds broadly to our retrieval and extraction methods, steps 2 and 3 correspond to our analysis method, and step 4 would correspond to the packaging and replication of the study.

4.2. Classification of validation studies

With respect to validation and reproduction, our original paper was based on several previous studies on the methods used to verify findings obtained from software engineering experiments. Among them, we relied mainly on the work of Gomez et al. (2010) [20], since we found it closer to our interest in MSR studies. Building upon the analysis of different ways of verifying experimental findings used in other disciplines, it identified three methods for verifying a finding, each of them fulfilling a particular verification purpose. Our approach, based mainly on our own analysis of MSR papers, was to identify reproducibility elements in MSR studies. We used Gomez et al. for analyzing, from the point of view of which reproducibility elements were needed, their classification of methods for verification. With this aim, we identified different cases of their classification:

Validation using the same method. This corresponds to the case when some or all of our reproducibility elements are maintained from the original study. We identified one category in this case: *procedural validation*, where all method and datasets elements from the original study were reused or used for comparing with the results of the validation study.

Validation reanalyzing existing data. In this case, datasets from the original study are reanalyzed using different methods. We identified two categories in this case: *new analysis based on the same raw dataset*, and *new analysis based on the same processed dataset*.

Validation using a different method. In our case, no element from the original study would be reused, so in our original paper this case was really out of scope. However, we identified a case, *new study reusing only the retrieval tools*, in which the tools used to retrieve the data were reused from a previous study, but all methods and datasets were really different.

In 2014, a more elaborated version of Gomez et al. (2010) was published by the same authors, Gomez et al. (2014) [5], with a more nuanced classification of verification studies. This classification is based on what changed in each kind of study with respect to the original study it is validating:

Literal replication. In this case, the aim is to run as exact a replication of the original experiment as possible. In the case of our original paper it would correspond to a *procedural validation*.

Operational replication. In it, some operational aspect is changed: the protocol, the operationalization, the population, or the experimenters. In our original paper, we capture both protocol and operationalization aspects in method elements and study parameters. The population, in the case of MSR studies, can be identified as the sample of the data source considered in the study. Experimenters are out of scope in our original analysis, since it does not make a difference from the point of view of reuse if the validation is done by the same or a different team. Therefore, we can consider that studies in which method elements, or study parameters, change, follow in this case. In our original study, we did not have a category for it, but it would be easily considered just by having into account changes in these elements.

Conceptual replication. In this case, new protocols and operationalizations are used by different experimenters to verify the results observed in the original experiment. This case is out of scope in our original paper, except for the case of *new study reusing only the retrieval tools*, as we already mentioned (and of course, *new study*).

It is worth noticing that Gomez et al. (2014) consider that, for performing a validation (*reproduction*, in their words), it is necessary to "execute an experiment". In their explanation, authors state explicitly that "this omits activities that some authors define as replication types like reanalyzing existing data using the same procedures, different procedures, or different statistical models to the baseline experiment". This detail is important, since it puts out of scope the *validation reanalyzing existing data* category described in Gomez et al. (2010). As we mentioned already (see Section 2.3), our original paper was interested in all kinds of MSR empirical studies, and therefore our categories *new analysis based on the same raw dataset*, and *new analysis based on the same processed dataset* are not considered in their new version of the study (see Table 8).

Validation studies (named as *replication studies*) is also one of the kinds of studies considered by Empirical Standards for Software Engineering[2] [21]. They differentiate between *reproduction study* (repeat the original study's data analysis on the original study's data) and *replication study* (repeat a study by collecting new data and repeating the original study's analysis on the new data). The former corresponds to *procedural validation* in our original paper, while the latter,

[2] Empirical Standards, Replication: https://acmsigsoft.github.io/EmpiricalStandards/docs/?standard=Replication.

Table 8
Comparison of classifications of validation studies in our original study, Gomez et al. (2010) [20] and Gomez et al. (2014) [5]. (*): this category was not considered in the original study.

Our original study	Gomez et al. (2010)	Gomez et al. (2014)
Procedural validation	Same method	Literal replication
Different method or parameters elements (*)	–	Operational replication
New analysis New analysis/same retrieval tools	Different method	Conceptual replication
New analysis/same raw dataset New analysis/same processed dataset	Reanalyzing existing data	–

which would amount to reusing methods but not datasets was not contemplated in our original paper. Both of them would be cases of *validation using the same method*, according to Gomez et al. (2010). According to Gomez et al. (2014), the former can be considered as a *literal replication*, but the latter has no clear equivalent: maybe a kind of *operational replication* where the population (which would be the dataset) is changed.

4.3. ACM badges

The most important development in the assessment of the reproducibility of computing studies is certainly the ACM Artifact Review and Badging [9] initiative (from now on, ACM Badges), which is being used as the basis for granting badges to papers in several major conferences in the area. As such, it is also used for the review process which leads to deciding on granting those badges or not.

The first characteristic of ACM Badges is that badges are granted to studies, by means of the artifacts accompanying them. Badges are of three types:

Artifacts Evaluated. This badge is granted "to papers whose associated artifacts have successfully completed an independent audit", which is basically a review process. This ensures a certain level of verification on those artifacts, but does not require their publication. This badge, alone, has little impact on reproducibility, except for, maybe, considering privately requesting authors for the artifacts, in case of performing a validation study. However, it ensures a certain level of quality, since it requires that artifacts are documented, consistent, complete and exercisable.

Artifacts Available. This badge is granted "to papers in which associated artifacts have been made permanently available for retrieval". This badge ensures availability of the artifacts to help when performing validation studies at any given time in the future. This badge can be granted on its own, but when combined with the previous one, it ensures that the reuse of the artifacts will be relatively easy.

Results Validated. This badge is granted "to papers in which the main results of the paper have been successfully obtained by a person or team other than the author". Therefore, it does not require reproducibility, but that at least one actual validation has been conducted, and its results were positive. There are two versions of it: *Results Reproduced* (the validation study was performed using some of the artifacts from the original study) and *Results Replicated* (the validation did not use any artifact from the original study). In both cases, the validation should be conducted by a different research team.

Even when our original study predates ACM Badges by several years, and to our knowledge it was not considered when defining their evaluation model, it is remarkable how similar both approaches are with respect to the aspects to evaluate. Table 9 displays an organized summary of aspects of artifacts that are evaluated for ACM Badges, and reproducibility attributes and other assessments that are described in our original study. It shows how some of our attributes have a direct correspondence with aspects evaluated in ACM Badges:

- Our identification and description attributes correspond roughly to the consistent and documented evaluations of ACM Badges.
- Our availability and persistence attributes, together, correspond to the available evaluation in ACM Badges, which also specifies these two dimensions.
- Our method does not include an attribute for the complete evaluation, but it corresponds to the assessment that all reproducibility elements are present, according to our list of reproducibility elements.
- Our method does not include an attribute for exercisable evaluation. We perform a usability assessment instead. Even when they are different, usability explores to which extent the reproducibility element could be used in a reproduction study, which in some sense goes in the same direction, even when it is a much weaker evaluation.
- We have not found an equivalent for our flexibility attribute, maybe because it is mostly useful in the cases of reuse, not necessarily with the aim of validation, thus being out of scope for ACM Badges.

ACM Badges only defines the characteristics of the badges, and not of the review process for granting them, which for now is left to the entities (conferences, journals) running those review processes. They also do not limit to which kind of papers they can be applied, only that they are "research papers published in ACM publications". It is interesting to compare ACM Badges with the proposals of our original study:

Scope. There are clear differences in scope. Our proposal was focused on MSR studies, while ACM Badges aims to cover any research in computing. This focus allowed our proposal to be much more specific, since the kinds of studies to cover was much more uniform. It is interesting to notice that maybe this makes our proposal adequate for fine-tuning ACM Badges for the case of MSR studies, being one of the "grassroots efforts to evaluate artifacts and formally test replicability" (as ACM Badges put it) in this area.

Reproducible elements. ACM Badges is only interested in artifacts, defined as "a digital object that was either created by the authors to be used as part of the study or generated by the experiment itself. For example, artifacts can be software systems, scripts used to run experiments, input datasets, raw data collected in the experiment, or scripts used to analyze results". This definition is very similar to what we considered as reproducible elements. Our dataset elements are clearly within the realm of that definition. Maybe the main differences are in method elements, which are any kind of description of a process, if possible codified in software. We include in it software built for other studies, such as retrieval software that can be adapted to a new study. We also consider configuration data, which can be used to tailor the tools and parameterize the study. In both cases, the software environment needed to run scripts and other research software is not directly taken into account. This leads to the consideration that maybe software support for reproducible research [22] should be included as a dimension of reproducibility.

Table 9
Summary of aspects evaluated for granting ACM Badges, and attributes of elements and other assessment aspects in our original study.

ACM Badges		Our original study	
Badge	Artifacts Evaluation	Reproducibility attribute	Other assessments
Artifacts Evaluated	Consistent	Identification	
	Documented	Description	
	Complete		All elements present
	Exercisable		Usability assessment
Artifacts Available	Available	Availability	
		Persistence	
	–	Flexibility	

Missing elements. ACM Badges is not very specific about which artifacts should be present for granting a badge. It just states (for the Artifacts Evaluated badge) that artifacts should be complete, meaning that "all components relevant to the paper in question are included". This aspect should be evaluated during the review process, but there is no checklist to ensure none is missing. This is reasonable given the may different studies considered by them, which may render it difficult to produce a detailed list of artifacts. Our proposal, being much more restricted, can specify all elements that are relevant for reproducibility in MSR studies. Therefore, it is easier to detect that some important element is missing at all, having a clear impact on the reproducibility assessment report.

General or detailed assessment. ACM Badges aim for general, binary badges: either the badge is granted, or not. On the contrary, our proposal produces a detailed assessment of the different attributes of each reproducibility element. This allows not only to know if a study is reproducible in some dimensions (functionality and availability), but also how far is a study from having good reproducibility, by producing information for all its elements.

Availability as a separate dimension. In the case of the ACM Badges, availability is considered as a different dimension than other characteristics of a reusable artifact, and is granted with a specific badge, the Artifacts Available badge. Other dimensions are evaluated for the Artifacts Evaluated badge. This distinction may make sense from a review point of view, since reviewers of the artifact had access to it, so they can assess its characteristics without it being public. However, from a reproducibility point of view, if the artifact is not available, its usefulness is very limited. In our proposal, we considered availability as just one of the dimensions to be evaluated.

Availability in a repository. The Artifacts Available badge requires that all artifacts are available from a publicly accessible archival repository providing unique identifiers for the artifacts. This is something that is not mentioned in our proposal, maybe because archival repositories for artifacts are relatively recent. However, this is a very important requirement, since it allows for both the preservation and the findability of the artifact in the future. Fortunately, there are several options for archiving artifacts in this way, which makes it reasonable to have it as a specific requirement. In our proposal we included the availability and persistence attributes, which are a less detailed way of expressing this condition.

Recognition of being validated. One very important aspect included in ACM Badges by design is the recognition for efforts leading to make a study more reproducible. One of the reasons for researchers not making their studies more suitable for reproduction is the lack of incentive, which makes that the significant amount of effort needed for that gets no recompense. ACM Badges go in the direction of providing incentives: researchers get an independent review that identifies their paper as more easily reproducible than others without the badge. This has two effects: on the one hand, reproducibility efforts are publicly highlighted. On the other, reproduction studies are more likely to happen, since other researchers know the bar for reproduction is lower. As we will discuss later, the lack of this incentive was maybe one of the reasons for our proposal to have a limited reach.

As a summary, our reproducibility assessment report has similar information to the results of a review to decide if a study should be granted an Artifacts Evaluated or Artifacts Available badge. The characterization of validation studies, which includes which elements of the original study were reused, and how, has also similar information to the review report that should be used to decide if the Reproduced Badge is granted. Thus, we can conclude that, despite its differences, the ACM Badges schema and our proposal both produce similar outputs. In the case of ACM Badges, the review report may remain private, with the badge granted acting as a very terse summary report. In the case of our proposal, the results of the assessment are more nuanced, and could be used as a basis for such ACM Badges review report.

5. Empirical study

To learn about the current status of reproducibility in MSR studies, we have analyzed all long papers (10 pages or longer) accepted for the 20th International Conference on Mining Software Repositories (MSR 2023). The main aim is to learn to which extent our proposal for analyzing reproducibility of MSR papers still holds more than 10 years later, and how it could be extended, if needed. At the same time, being MSR one of the most important conferences in the field, we will also show to which extent reproducibility has become a fundamental property of accepted papers.

5.1. Methods and datasets

Description of the methods used in the study, and the datasets produced:

Data sources: Proceedings of the 20th International Conference on Mining Software Repositories (MSR 2023), electronic version (PDF format), as provided during the conference. These papers are subject to copyright restrictions, and cannot be shared publicly.

Retrieval method: Download of all papers in the proceedings of the conference, using the browser. All papers are downloaded with the same name they have in the electronic proceedings.

Raw dataset: All papers downloaded, which correspond to all papers accepted for the conference.

Extraction method: Use the tool `pdftk dumpdata` to produce a list of all papers with their page count. Then, filter from that list all files longer than 9 pages. The files in the resulting list are the files to be considered for the analysis.

Parameters: The only parameter is the (minimum) number of pages of a file to be considered in the study (10).

Processed dataset: The list of files obtained after the extraction method, and the files listed in it.

Analysis method: Manually producing an assessment report for each paper in the processed dataset, see detailed information about the criteria used below. Then, produce some charts and tables with a Python notebook.

5.2. Criteria for producing the assessment reports

We produced assessment reports according to the guidelines provided in the original study, summarized in Section 3.2, with the following extensions:

Reused value: New value for the reproducibility attribute "Identification": "Reused". This value will be valid for datasets, and will mean that the corresponding dataset is a reused dataset, produced independently of the study presented in the paper. In this case, other attributes will not make sense for this element, since they are attributed of the reused dataset, and not of a real element produced for the study.

Reused assessment: New assessment for dataset elements, when "Reused" is value for the "Identification" attribute. It is also important to notice that in this case, previous reproducibility elements will not make sense: the elements that are needed to produce this dataset are not elements of this study, but of the process that produced the reused dataset.

In addition, to the reproducibility assessment, we also determined for each paper:

Reproduction package: If the paper identified a reproduction package, including reproducibility elements such as datasets, or source code for methods.

Reuse: If the study presented in the paper reused some dataset.

Third party reuse: If the study presented in the paper reused some dataset produced by a third party.

There are also some details that are important to mention, with respect to how reproducibility elements were estimated:

- We did not try to reproduce the papers, only check the information they provided, usually in their introduction, conclusions, and methods sections. We did our best to find mentions to their reproduction packages, and used information in them when available.
- We focused on datasets and the methods for producing them, as indicated in the original study. However, when machine learning models are involved in the study, we ignored the models themselves, considering as something "to validate" by the study, and not a data of the study itself. This decision is arguable, but should not have a large impact on the final results, since usually the reproduction of models could be considered similar to the one of the source code that runs them, which we considered as a part of the methods.
- We only considered "Persistence" as "Likely" (best value) for open archives designed as such, with preservation in mind, and not general archives of code or other artifacts, with a more industrial focus, and which are not committed to long term preservation of the archived items. For example, Zenodo or Figshare are in the former category, while GitHub or a personal website are in the latter.
- In the case of having several datasets or methods for a certain paper, we have analyzed all of them in the corresponding assessment report. But we have considered only the one that seemed more fundamental when aggregating data for all papers.

Table 10 shows a summary of the reproducibility assessment reports for all papers in our study (one row per paper, considering its most important element when there is more than one).

5.3. Results

When analyzing which papers provided a reproduction package, and which ones reused datasets (see Table 11), we found that almost all papers provided a reproduction package (more than 90% of them). Of those that did not provide one, in one case it was due, apparently, to ethics concerns, in another one it was due to an error in the link to the package (maybe the paper intended to provide a reproduction package, but the linked repository was empty). Only in one case the paper did not mention a reproduction package or a reason for not having it.

It is also interesting to mention to which extent it has become usual in MSR papers to reuse datasets not directly produced for the paper. We found that more than 40% of all papers used at least one reused dataset. Moreover, we also found that almost 30% of all papers used a dataset produced by third parties (about 75% of all reused datasets were produced by third parties).

When analyzing the reproducibility assessment of the MSR papers, we found that the level of usability of the different elements is in general high. As a summary Fig. 2 shows the assessment for all papers for all reproducibility elements.

For the papers in which the study used directly a data source, 86.2% of them were assessed as "usable" for reproduction as described in the paper or in the reproduction package (not all papers declare data sources, usually because they reuse datasets instead). Only 10.3% are "not usable" or not described at all (see details in Fig. 3).

For the different kinds of datasets (see a specific chart for them in Fig. 4), again for the papers whose methods used them, we found the better level of reproducibility for processed datasets. 63% of the papers using a processed dataset received the assessment of "usable". although for 30.6% no processed dataset is provided. However, raw datasets also a high level of reproducibility, because of the combination of "reuse" (37.8%) and "usable" (24.3%) datasets in papers. On the other hand, results dataset are non-existent in 70.3% of the cases, and "usable" in the rest.

For the reproducibility assessment of methods we also offer a specific chart in Fig. 5. In this case, the closer to results the method is, the most usable for reproduction we found it. In the case of the analysis method, 75.7% of the papers have a "usable" description of the analysis method. This number decreases to 67.6% of "usable" descriptions for the extraction method, and to 51.7% for the retrieval method. However, in this latter case, we do not count reused datasets, which could significantly increase the fraction of reusable descriptions, and in any case, for reproductions based on the raw dataset, would be good enough. It is also worth noticing that in this assessment, we considered as "usable" descriptions in the form of source code, or, in rare cases, if they were very, very detailed.

6. Discussion

In this section we will put our original study in the context of new developments, such as the proposal of ACM Badges, and the emergence of new approaches to reproducibility.

Table 10
Long papers accepted to MSR 2023: reproducibility assessment for their main elements.

Data source	Retrieval method	Raw dataset	Extraction method	Parameters	Processed dataset	Analysis method	Results dataset
U	U+*	R	U+*	U	U+*	U+*	U+*
-	-	R	-	D	D*	D+*	-
U	U*	U*	U*	D	U+*	U+*	-
U	U+*	R	U+*	U	U+*	U+*	-
U	U+*	-	U+*	U	U+*	U+*	U+*
U+	U*	U*	U*	D	U*	U*	-
		R	U*	U	U+*	U*	-
U	U+*	U+*	U+*	U	U+*	U+*	U+*
N	-	N	-		U+*	D+*	-
U	U*	R	U*	U	D	U*	-
U	U+*	U+*	U+*	U	-	U+*	-
		R	-	N/A	-	D+*	-
U	U+*	-	U+*	U	-	U+*	U+*
U	U+*	U+*	U+*	D	U+*	U+*	-
U	U*	U*	U*	U	U*	U*	-
		R	U+*	U	U+*	U+*	-
U	U+*	-	U+*	U	U+*	U+*	-
		R	D+*	N/A	U*	U*	U*
U	U+*	-	U+*	D	-	U+*	-
		R	D+*	N/A	-	D+*	-
U	D+*	-	D+*		-	D+*	-
		R	D+*	D	U*	U*	-
		R	D+*	D	-	D+*	-
U	D*	-	U*	U	-	U*	-
U	U*	-	U*	D	-	D+*	U*
U	D*	-	U*		U+*	U+*	-
U	D*+	-	D*+	D	U+*	U+*	-
-	-	-	D*		-	D*	-
U	D+*	U+*	U+*		U+*	D*	-
		R	U+*	U	U+*	U+*	U+*
U	D+*	-	U+*	U	U+*	U+*	U+*
U	D+*	U+*	U+*	U	-	U+*	-
U	U+*	R	U+*	U	U+*	U+*	U+*
U	D+*	R	D+*	U	U+*	U+*	U+*
D	D+*	-	D+*	U	R	U+*	U+*
U	D+*	-	U+*	U	U+*	U+*	-
U	D+*	U+*	U+*	U	U+*	U+*	-

Table 11
Long papers accepted to MSR 2023: reproduction packages and reuse of datasets.

	True	False	Total
Provides a replication package	34 (91.89%)	3 (91.89%)	37
Reuses at least one dataset	15 (40.54%)	22 (40.54%)	37
Reuses at least one dataset produced by a third party	11 (29.73%)	26 (29.73%)	37

6.1. Comparing with MSR before 2010

The state of reproducibility for papers accepted in the International Conference on Mining Software Repositories (MSR, then named International Workshop on MSR, and later Working Conference on MSR) for its editions of 2004 to 2009 was analyzed in [2]. That paper presented a basic reproducibility assessment for all papers (long and short) presented at the conference, based on the analysis of the availability of the "raw datasets", "processed datasets" and "tools and scripts" for the studies they presented. Even when the kind of papers analyzed, the assessment categories, and the conference itself are not exactly the same, we think that a comparison between our analysis of MSR 2023 and the results of that paper are worth discussing.

For this discussion, we compare our "raw dataset" and "processed datasets" reproducibility elements to the elements of the same name in the original study. For "tools and scripts" we compare with all of our methods elements, since those tools and scripts covered all methods (retrieval, extraction and analysis). Since the original study analyzed several editions, instead of focusing on one of them, we will use for comparison the data found in its Table 1, which provides numbers for the three elements for the total number of papers analyzed (see Table 12). For our current study, we have produced a summary of results for comparison (see Table 13). For methods, the "usable" characteristic is similar to "tools and scripts available" in the original study, since usually it means that source code is available. However, it is more difficult to compare our results with "tools and scripts partially available" or "tools and scripts not available", thus consider the comparison for "Partial" and "False" in the case of methods just as a rough approximation.

Let us analyze the comparison with the three elements of the original study:

Raw dataset. Reproducibility in this case is lower currently than it was in the original study (approximately, from 70% to 64%). However, the main difference is not evident in those tables: a large quantity of papers in the current study include reused dataset, a case which was not even a category in the original study. We can say that reusing raw datasets is now a very extended practice, leveraging a large quantity of high-quality datasets that have been made available. Therefore, the reuse

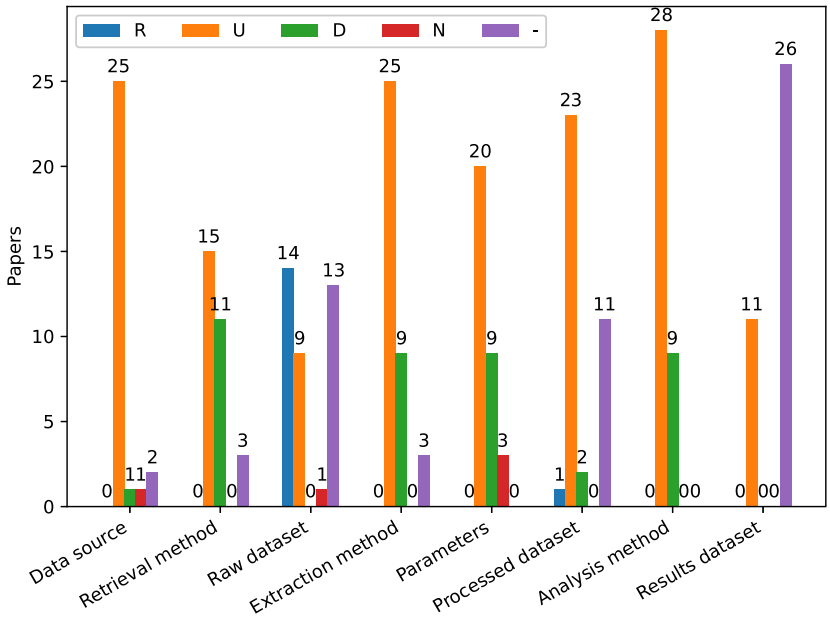


Fig. 2. Reproducibility assessment on elements with an impact on it, for all long papers accepted for MSR 2023. R: reused, U: usable, D: usable with difficulties, N: unusable, –: not available. Marks for persistence (+) and flexibility (–) are not included in this chart. For each element, only papers considering it are included. For example, if the study presented in a paper starts with a reused processed dataset, the following elements are not included: data source, retrieval method, raw dataset, and extraction method.

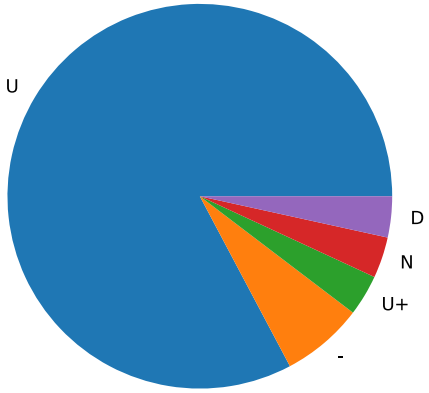


Fig. 3. Reproducibility assessment on the data sources reported in all long papers accepted for MSR 2023. U: usable, U+: usable and persistent, D: usable with difficulties, N: unusable, –: not available. Only papers considering data source are included.

Table 12
Assessment of reproducibility elements in the original study [2], as detailed in its Table 1 (reelaborated for convenience). For each element, "True" means the element is public ("Y" in the original paper), "False" that it is not ("N" in the original study), "Partial" that it is only partially public ("P" in the original study). In the original study, no detailed assessment on the different reproducibility characteristics was performed. Papers assessed as "N/A" in the original study (17) are not considered in this table, which means we consider a total of 154 papers of the 171 papers presented in all the analyzed MSR editions.

	True	Partial	False
Raw Dataset	108 (70.1%)	3 (1.9%)	43 (27.9%)
Processed Dataset	6 (3.9%)	6 (3.9%)	142 (92.2%)
Tools & scripts	27 (17.5%)	23 (14.9%)	104 (67.5%)

of datasets, which was just an idea around 2010, is now a very common practice. On the other hand, the decrease in reproducibility could be due to the most exigent standards that we had in our current study.

Table 13
Assessment of some reproducibility elements in our current study, for comparing with the original study. For each element, "True" means the element is "usable" or "reused", "False" that it is "non-usable" or "absent", "Partial" that it is "usable with difficulty". For each element, percentages are computed excluding papers not using the corresponding element. Since in our study we have three categories that roughly correspond to "tools and scripts" (all methods), here we include all of them.

	True	Partial	False
Raw Dataset	23 (63.9%)	0 (0%)	13 (36.1%)
Processed Dataset	24 (64.9%)	2 (5.4%)	11 (29.7%)
Retrieval method	15 (51.7%)	11 (37.9%)	3 (10.3%)
Extraction method	25 (67.6%)	9 (24.3%)	3 (8.1%)
Analysis method	28 (75.7%)	9 (24.3%)	0 (0%)

Processed dataset. Reproducibility of the processed dataset has increased tremendously during these years (approximately, from 4% to 65%). Maybe this is the single most interesting difference between both studies, and an interesting fact of increasing future reuse. The same way we see the reuse of raw datasets as an extended practice nowadays, maybe in the future we will see a more extensive reuse of processed datasets of high quality.

Tools and scripts. This is also an element whose reproducibility has increased clearly during the last years. From 17.5% in the original study, we are now in 51%–75%, depending on which kind of method we consider. And the more close the method is to the final results, the better reproducibility of the method we have. This is very important for the automatic reproduction of the original results, and also important for new studies with variations in the datasets, since the methods can be run automatically or semi-automatically.

So, in summary, we can say that in general reproducibility of studies has improved clearly. Only in the case of raw datasets there is some deterioration, maybe due to the fact that the original numbers were already good. But for other elements (processed datasets and methods), the situation has clearly improved. In addition, reuse is now clearly established for some kinds of datasets.

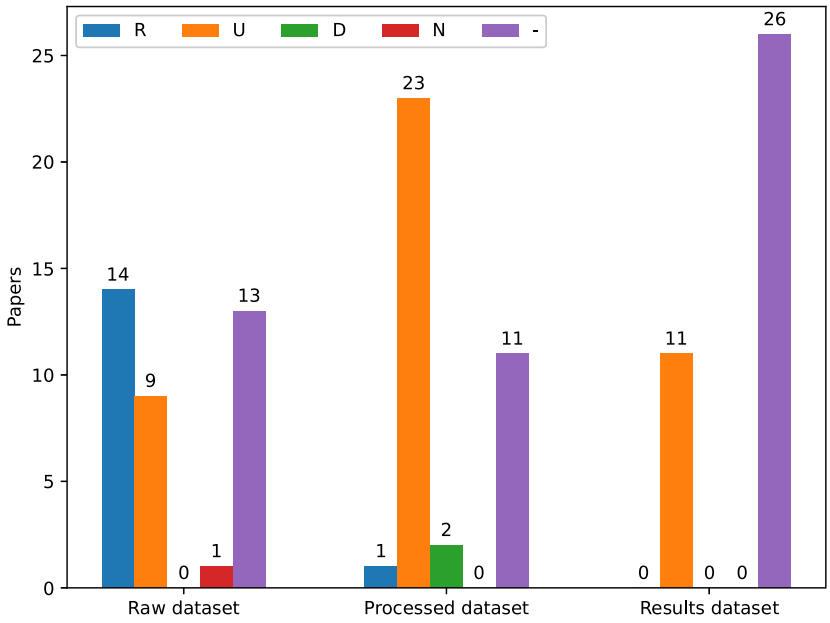


Fig. 4. Reproducibility assessment on the three kinds of datasets relevant for it, for all long papers accepted for MSR 2023. R: reused, U: usable, D: usable with difficulties, N: unusable, –: not available. Marks for persistence (+) and flexibility (–) are not included in this chart. For each kind of dataset, only papers considering it are included. For example, if the study presented in a paper starts with a reused processed dataset, the raw dataset is not included.

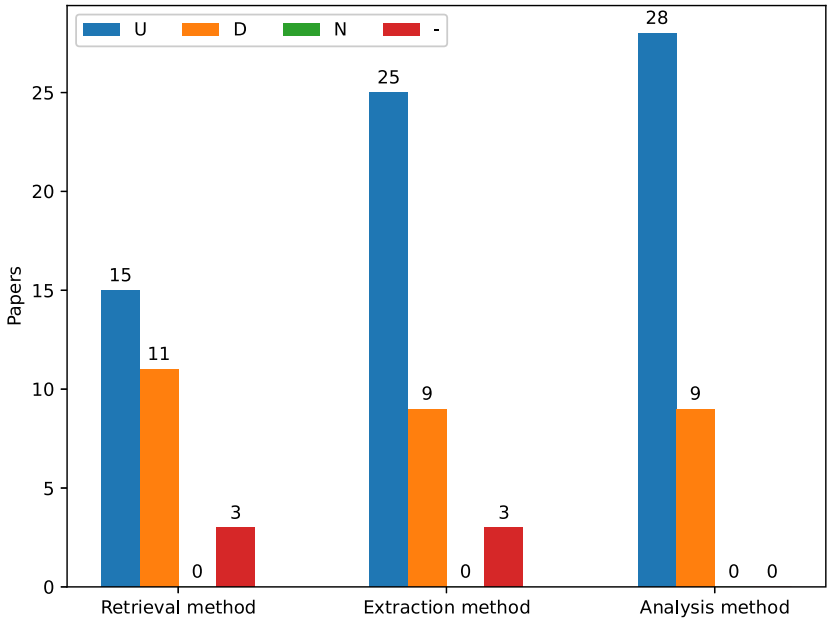


Fig. 5. Reproducibility assessment on the three kinds of methods relevant for it, for all long papers accepted for MSR 2023. U: usable, D: usable with difficulties, N: unusable, –: not available. Marks for persistence (+) and flexibility (–) are not included in this chart. For each kind of method, only papers considering it are included. For example, if the study presented in a paper starts with a reused processed dataset, the retrieval method, and the extraction method are not included.

6.2. Other approaches to reproducibility

During the years since the publication of our study, other approaches to reproducibility have also arisen, based not on the assessment of the reproducibility elements, but on facilitating the production of reproducible studies. Among them, we can mention, as good examples of these approaches, the following:

Reproducibility as code. Boa [23] is a well known example of this category, providing a scripting language that can express the process leading to the results of a study from the data sources. By running the script again, the same study is produced (usually, with fresh data). Boa also provides a software environment where those scripts can be run. To our knowledge, in the area of MSR studies, this is the more developed effort to support reproducible research with software, which as we mentioned above, was missing both in our proposal and in ACM Badges. Maybe including "is the process reproducible with code in a well-known platform?" as one more dimension when assessing reproducibility would make sense at this point. In any case, Boa allows for the automatic production of the method reproducible elements that we describe in our study.

Reproducibility as an archive. Software Heritage [24], GHTorrent [25] and World of Code [26] are good exponents of this approach. All of them archive data useful for MSR studies, with the double aim of preserving data for the future and making it easily available. In the case of Software Heritage, they archive all public source code, including all of its versions that they can find. For each version of each artifact (including whole repositories, but also finer grained elements such as files) they provide unique identifiers for each version, thus complying with the ACM Badges requirement of uniquely identifying the artifact. In the case of GHTorrent, they archive mainly metadata from GitHub, GitLab and other software development forges, and to some extent they have become the de-facto standard for that data. In both cases, their approach to helping reproducibility is by providing the whole archive for future researchers. World of Code is to some extent a mixture of both: they archive not only metadata, but also code for a very large collection of software projects, and are designed for supporting research studies. In some sense, they are also bridging the gap with the reproducibility as code approach. All of them are examples of how to automate the archival of the row datasets required to perform a study. In the case of Software Heritage and World of Code, if the data for the study is obtained from these archives, referencing the corresponding dataset element is straightforward, and it will automatically be identified and available in the future, as long as the archives remain active.

Reproducibility as data. Lean GHTorrent [27] is a good exponent this approach, which automates the production of a dataset with the relevant data for a set of GitHub repositories, as they were archived in GHTorrent when the study was produced. Since the data in GHTorrent evolves over time, this provides a data snapshot suitable for reproducing the same source data used by a certain study. With this tool, the production of the raw dataset described in our study is automatic, thus being a very good example of automating the production of reusable artifacts.

Reproducibility as a platform SmartSHARK [28] and GrimoireLab [29] are examples of this approach. They are complete platforms that can retrieve, store, analyzed and, to some extent, visualize, data retrieved from software development repositories. Their approach to reproducibility is based on recreating the exact same platform that retrieved and analyzed the data, by using its configuration files, although they also provide other helpers to produce useful reproduction packages. They can be also used to automatically or semiautomatically produce the dataset elements proposed in our study, and the scripts used to run them could encode most or all the method elements. Assuming the platform is available, this would mean a given study would be completely reproducible just by publishing those artifacts. In this sense, they go an step beyond the "reproducibility as code" approach, since they can also help with the provision of datasets, and allow for more generic analysis.

All of these approaches have in common that they help researchers to improve reproducibility of their studies, with less effort on their side. All of them can be combined with the assessing approach, by producing elements that will be later assessed. It is important to notice that in some cases, the assessment could be automatic. For example, if the element is stored in a preservation archive, the attributes for availability and preservation could automatically be assessed as true.

6.3. Other aspects

There are some other aspects worth mentioning with respect to our original study:

Lack of influence. Even when the original paper had a reasonable number of cites, we have not found any of them that uses our method for assessing the reproducibility of papers, nor for discussing our general approach. It is difficult to know the reasons, but we have some hypothesis. On the one hand, it seems there is relatively little interest in assessing the reproducibility of papers in detail. Even when during the last years checking if there a reproducibility package is becoming more and more common during the review of papers, and some conferences and journals have stated recommendations for checking the reproducibility packages, we are not aware of more detailed approaches to assessing reproducibility, and in particular, to structured approaches. In this context, there is no need to use our method, or any other detailed method, during review, or for self-assessment. On the other hand, if the method is not used, it is difficult to find reasons to discuss and improve it. Maybe this leads to a situation where the method is too difficult, or too unsuitable for assessing reproducibility, but there is little interest in refining it because there is very little experience in using it. Maybe a way for improving the situation in this respect is to provide some guidelines to improve reproducibility, and make easier to produce reproducibility assessment reports. We have addressed this idea in Section 6.4, by producing some guidelines based on our experience in the matter.

New kinds of studies. When we analyzed MSR papers more than ten years ago, some kinds of studies that are very popular today did not exist. Among them, the most prominent type is studies using machine learning (ML) techniques to analyze datasets. Usually, they use a part of the data as training set to produce a ML model capable of doing some kind of analysis, and a part of the data for evaluating it. Thus, these studies are not very different from 'traditional' ML studies, except for the fact that they use data from software development repositories. The analysis of the reproducibility of these studies fit perfectly well in our model. The raw dataset and the processed dataset are produced, as in other MSR studies, based on data from one or more data sources. The method for producing them can be described, and in some cases codified in software, producing the corresponding method elements. The ML model, along with the software needed to run and validate it, are also a codified-in-software description of the analysis method. Results can also be preserved as a results dataset. Study parameters are for example metaparameters of the ML model, and other configuration data. So, the assessment of the reproducibility of this kind of studies fit perfectly our model, which makes us think that other kinds of MSR studies that would be emerging in the next years could also fit well in it.

Exercisability. Our method does not require to actually exercise the reproducibility elements (test that they can actually be used for reproduction of the study). From some point of view, this could be an interesting property, because the effort needed for evaluation is much smaller than if exercising is required. However, that also makes our method less useful, since even when an element seems to be usable, it could really be unexercisable. It could be interesting adding a new dimension to the method to assess excercisability, at least in some cases, maybe as an optional assessment for those cases when the required effort is possible or required.

Modalities of reuse. When presenting the possibilities of reuse of reproducibility elements in different kinds of validation studies (Table 6), we introduced an aspect which is not presented in detail in the original paper: the different modalities of reuse of a reproducibility element. We have found this is a very important aspect when finding out how much difficult it will be to produce

a certain kind of validation study, and even if it will be possible at all. For example, to facilitate future validations of all the procedures of a study except for the data collection, raw data has to be available and usable. If it is not, the validation will require more effort, and will be more difficult, since data retrieval (which is not really a subject for validation in this case) will have to be performed. Even with a detailed description of the retrieval method, and with access to the artifacts that assisted in it, this may require considerable effort and time to understand the procedure. Even with all that effort and time, it will be impossible to ensure that the retrieved data is exactly the same as in the original study. And, what is worse, maybe changes (or disappearance) of the data source make it completely impossible to reproduce the raw dataset. But even in the case of studies that reuse very little, or nothing, of the elements of the original study, having some of them available can be very important for comparison. The best and easiest way of comparing results of a validation with those of the original study is having both available, in sufficient detail, and just compare them using any sensible statistics method.

Use of well-known datasets. Already in our original study we mentioned the existence of some well-known datasets, which in our schema could be used as raw or processed dataset, sparing the researcher from the retrieval work. When the aim of the study does not require obtaining fresh data from a data source, this is a very convenient validation study. It is worth noticing, however, that despite its name, in this case the study is not necessarily intended to reproduce any part of the original study: it may use the same dataset with completely different objectives (for example, to test a different hypothesis). This is important to notice, because it shows how reproducible elements are important for more than reproducing research: they also allow performing some studies with less effort. Our proposal acknowledges this fact by making it clear that datasets are reproducibility (and thus potentially reusable) elements by themselves. This kind of reuse, which is increasingly common in some fields, is unfortunately missed in many studies about reproducibility, because, as we said, is not really a reproduction practice.

Archiving infrastructure. As we commented above, the existence of archives designed for preserving reproducibility elements, is fundamental for their future reproducibility. In some cases, such as Zenodo, the archival of reproducibility elements can be done with the paper, thus ensuring that both are available in the future. However, another possibility is to archive elements in a specific repository for their kind, which supports their peculiarities. This is done, for example, by Software Heritage for software. It is important to notice that software development platforms, such as GitHub or GitLab, are not good preservation archives. Even when they have a very good record of availability and preservation of the artifacts they hold, there is no guarantee of preservation in the future. Unfortunately, there are no good preservation archives for artifacts other than source code, which means that preserving with the paper, or in open data repositories, is currently the best option to comply with ACM badges requirements, and with our availability and persistence attributes. When such infrastructure is used, assessment of availability and persistence attributes could be automatic. Besides, this difference between archives intended for preservation, and other that are not, could also be captured as a characteristic of reproduction packages, or of the reproduction elements included in them.

Bottom-up, detailed approach. In our original study, we presented a fine-grained approach to reproducibility, with a bottom-up approach. We start by identifying reproducibility elements, and their attributes that are important for reproducibility. Then, in the case of original studies, we define how to produce a detailed assessment on their reproducibility, and in the case of validation studies, a detailed assessment on how they reuse elements from the original study. It is important to notice that on top of these reports, shorter summaries could be produced, along the lines of ACM Badges, or even ACM Badges themselves. But these reports could be used for other aims. For example, authors could assess their own studies, with the aim of detecting which reproducibility attributes are subject to be improved. Or a researcher considering performing a validation study can use the reproducibility assessment report of the original study to evaluate the effort and feasibility of such endeavor.

6.4. Guidelines: reproducibility in MSR studies

These guidelines are intended to improve the assessment of reproducibility, to help people producing assessment reports, and in general to improve the reproducibility of MSR studies. They are based on our experience defining and analyzing the method for assessing reproducibility we presented in [1], and in the assessment reports we have written for many MSR papers, including those that can be found in the reproduction package for this paper.

When authors write papers presenting their studies, they could consider:

Clear definition of the study. Include a section in the paper (usually it is the *Methods* or *Description of the study* section) clearly describing the experiment or study, with all the detail needed for others to reproduce it. If the length of the paper does not allow for a detailed description, use the description file in the reproduction package. This helps to correctly fill in the assessment for these two attributes.

Identification of the reproducibility elements. When describing the study, include a clear identification and description of the reproducibility elements. If some of them are not relevant, explain why. This also helps to correctly identify those elements in the report.

Identification of the reproduction package. Ensure that the reproduction package is clearly identified, and easy to find, in the paper. For example, include a clearly visible, short section at the end of the paper, stating that datasets and source code are available in a reproduction package, with the corresponding link to it. Ensure that the link is persistent, when that is possible (for example, a DOI) and if possible, make it clickable in the PDF, and check it. This will help when writing the assessment report, by not missing the reproduction package, or not being able of downloading it.

Archiving the reproduction package. Whenever possible, ensure that the reproduction package is archived in an archive intended to guarantee long-term persistence, and FAIR principles,[3] such as Zenodo or Figshare. This will ensure that the package can be found in the future.

Single reproduction package. Put all reproducibility elements together in a single reproduction package, whenever possible. If some of the elements (such as source code) are stored in their specific repositories, include a copy of them as was used in the study in the reproduction package, and link to the repository from it. This will help people writing assessment reports, by having a single package to check, but also future reproducers, by having a specific version of all elements together.

[3] FAIR principles: https://www.go-fair.org/fair-principles.

Description of the reproduction package. Include a file in the reproduction package with a clear description of its contents, and the detailed procedure for reproducing the results from the data sources or raw dataset, if it is not described with enough detail in the paper. This will be fundamental for a quick and fair review of the reproducibility elements in the package.

Raw dataset. Raw dataset is important, even if very large: it is likely that if the data source is mined once again, results are different, which makes it difficult to replicate the study.

Processed dataset. Processed dataset is important: it is rare that the raw dataset is directly used, if it is just a "download" of the data source.

Results dataset. Results dataset is important: its presence makes it much easier to compare results (tables, figs, tables to produce figs).

Source code better than text. Even when the paper may describe methods with text, it is convenient to codify them as much as possible in source code, so that barriers to reproduction and to interpretation are lowered. In general, source code will get better reproducibility assessments than text: even if the code could not be executed, it is usually a more precise description of the process.

Reused dataset. If reusing datasets, cite the link in addition to the reference to the corresponding paper, that makes it much easier to find the dataset, and being sure it is exactly the one you used.

Self-assessment. You can produce a reproducibility self-assessment for your paper, and include it (for example, in a last section about reproducibility). In the reproducibility package you can include more details on it, to easy the work of reviewers and other people producing their own reproducibility reports. This way, reproducibility reports can also be part of the review process more easily, and could be used as the basis for third-party assessments.

6.5. Threats to validity

Since this paper was mainly conceived as a revisit of our original work, ten years, later, in several aspects we did not follow a systematic approach, which could affect the results. In particular, we can mention how this could affect to some aspects of our paper:

- Related work. Even when we did a reasonable effort for finding related work, based mainly on cites to our original papers and to other relevant papers that we knew, we did not perform a systematic literature review. This means that maybe some relevant works are not cited. However, our aim was not to present all related work, but only that which helps to put into context our original work, and how the field has evolved during the last decade in terms of providing frameworks for addressing the problem of reproducibility in empirical software engineering in general, and in MSR studies in particular.
- Terminology. Most of Section 2 is based on our interpretation of the mentioned studies. We could have missed relevant classifications, and misinterpreted some of them. Even when we did our best effort to identify specific definitions of the terms, in some cases the context is different, or we could miss some important context, that could render some of the equivalences wrong, or at least not exact. In addition, the terms that we selected for use in this paper were in part based on our own preferences: other authors could have selected some different terms or uses of the terms with equal basis.
- Reproducibility elements. In the original study we did an effort to ensure that our collection of reproducibility elements, and their characterization, was appropriate for all kinds of MSR studies and their peculiarities, based on the analysis of MSR papers of that time. In this paper, we have followed the same classification of elements, finding that it still can be applied to current MSR papers. However, there is a range of interpretation when mapping a paper to the list of reproducibility elements. Since the authors proposing the list of elements are the same doing the analysis on the papers, there could be some bias in their identification and mapping.

With respect to the study of the reproducibility of long papers accepted in MSR 2023, we did our best effort to find references to reproducibility elements in the papers and their reproduction packages. However, we could have missed some of them, or misunderstood some of them, thus producing wrong assessments. In some cases it is really difficult to find out if datasets are available, if source code for a method has been published, or even if some part of the code should be considered a part of the method or not. All of this, and some other details, could in general lead to wrong assessments of the reproducibility of papers.

In the comparison between the situation in 2004–2009 and in 2023, we have to acknowledge differences both in the data sources and in the selection criteria for the datasets. Considering the papers presented in the MSR Conference as the data source, it is important to understand that the conference has evolved significantly during this period, as have the papers presented at it. In 2004–2009 it was still a relatively small conference, evolving from its origins as a workshop, for an emerging research community. In 2023, it is a respected and important venue for an established and wide research community. This means that the standards for accepting papers, and even the aims of those papers, may be different. This could make the comparison difficult in some respects. From a more practical point of view, in the 2004–2009 study we considered all papers presented, including both short and long, in all tracks. This means that, for example, many short papers of the Data Showcase were included in the analysis. In 2023, we have only considered long papers, which are usually more mature, and subject to stricter review guidelines. This could also render the comparison of the reproducibility for both cases difficult.

6.6. Emerging ideas

While reconsidering our original study, and having into account what happened in this domain during the ten years that have gone by, we thought that there are some ideas worth considering when assessing both the reproducibility of MSR research studies, and the reuse of reproducibility elements by validation studies:

Improving usability. The method that we proposed in our paper, even when it has been cited by many other studies (107 different citations according to Google Scholar, or 65 according to Scopus, both checked on February 2023), has only being used to check reproducibility in a very limited number of cases. In a short fraction of the period that our study has been available, ACM Badges has obtained a considerable traction, being used in several major conferences on computing research. Of course, we cannot directly compare both approaches. Our paper was a research study, with actionable results but no incentive for adopting them. On the contrary, ACM Badges is a very interesting effort to increase the validation by third parties of computing research studies. It was directly designed by ACM so that ACM (and other) conferences could use it, thus providing authors with the incentive of being granted the corresponding badge if they fulfill its conditions. Looking retrospectively, we now think that incentives and being a part of an organized approach are

fundamental. Bona fide effort in reproducibility is something that most researchers take as granted, but assessing on the reproducibility of studies is something that is more difficult to afford. When we published the study, we expected that some papers would start to voluntarily use our reproducibility assessment as a self-assessment, to state publicly how reproducible it was. But this did not happen: either researchers used our assessment but kept it to themselves, or they did not use it at all.

Reflecting on the way our original paper was presented, we now think that we did not explain in it how it could be useful for more than just an academic curiosity. Thanks to the occasion of preparing this paper, we have reflected on this matter, and have decide to explicitly explain how our assessment method could be used for several tasks related to the reproducibility and the validation of studies (see Section 6.7.

Consequences of validation. The study by Gomez et al. (2014) [5] provides a detailed analysis of the functions of validation studies (referred as *functions of replication*). Using that analysis, we can infer the functions of different kinds of MSR validation studies, depending on which reproducibility elements are changed, and which validity threats are addressed with them. We have summarized this inference in Table 14. This table is of great importance to learn which elements to change depending on the threat we are addressing, when conducting a validation study to extend the validity or the original. It also means that by improving some reproducibility elements, some kinds of threats to validity can be more easily addressed in the future, by making the corresponding validation studies easier to perform.

The most extreme cases are when no element is changed, or when all of them are changed. In the former case, the validation study (which will be a repetition or a reproduction study, in ACM Badges terminology) will control for sampling error, limiting the possibility that results are obtained by chance. This will produce a better understanding of natural variations of results, and will address conclusion validity threats. In the latter case, a completely new study is performed, keeping the hypothesis to validate. This kind of validation will control for errors in the design of the study. When one or more of our method elements or study parameters are changed, we learn about the operational limits of our study, addressing construct and internal validity threats. In this case, we are changing how we obtain results from our original dataset. Of course, all datasets of the original study that are produced by steps further than the method changed will also change. However, they will still be useful for comparing the new resulting datasets with the original ones, to detect meaningful variations of final or intermediate results. Changing the data source or the raw or processed dataset in a validation study checks for the validity of different samples of objects. In this case, the same methods are reproduced on a different sample, which addresses external validity threats.

We do not considered the case of changes in the experimenter group, as we have not considered it a reproducibility element. However, it is interesting to notice that validation studies can be run by the same or by a different research team. If the validation study is run by a different research team, that will have an impact in controlling independence with respect to experimenters, addressing for example experimenter biases or malpractices. But it will not have an impact in the other validation functions described.

Reuse beyond validation studies. Most of the literature on reuse of research artifacts is linked to facilitating validation studies. However, there are interesting cases of reuse that go beyond this scenarios. We have already identified two of them in the case of MSR studies: the reuse of data retrieval and analysis tools, and of well-known datasets.

Data retrieval tools which implement a certain retrieval method, or better, that have the flexibility of implementing a family of retrieval methods, can be a byproduct of some research studies. In some cases, they are even intentionally designed to support a wide number of methods, intended to be used by a certain research team, or they evolve with the needs of several teams that help to maintain them over time. Since the retrieval method is usually encoded as a software tool, this is the most common case, but the same could happen for extraction or analysis methods. In fact, the approach *reproducibility as a platform* (see Section 6.2) intends to encode all methods in a single platform, that also produces the intermediate and final datasets. In addition to being very useful for performing validation studies, because they can easily automatically run variations of the original study, they are also useful for conducting new studies from scratch, with much less effort than necessary if tools are built from scratch. And since they run the studies automatically, they can easily instrumented to automatically produce datasets at the appropriate points in their processing. What is more interesting, these toolsets could also read datasets produced by them, or by other toolsets, making it much easier to reuse datasets from other researchers. But to be useful and trustable, these toolsets also need their own assessment. This leads to their from the point of view of our reproducibility assessment. In this case, he assessment would not be of a study, but of a toolset, assessing on which different methods can be configured, and which assistance will be provided for producing final and intermediate datasets. This would mean that researchers using one of those tools would know that, by default, several elements of their studies will be reproducible just because the toolset will ensure it.

Well-known datasets have been used in MSR studies since many years ago. Even when their use may mean that no real experiment is performed (again, see discussion in Section 2.3), they are very convenient to explore new methods. Instead of having to retrieve, and maybe process data, researchers can focus on the analysis of it, using different techniques. In many cases, the analysis is completely independent of the original analysis that produced the dataset. With such a popularity and usefulness, specific assessments for these datasets could be produced. Borrowing the idea of the *Results Validated* badge, we could grant *dataset reused* badges to datasets to recognize their usefulness to other researchers. Going one step beyond, we could also produce assessments of the suitability of datasets for different purposes. Those could be produced by the original authors of the dataset, but could be completed by others, producing different studies on them. To some extent, we could also borrow from the idea of a *model cards* [30], used my the machine learning community for ML models, summarizing the characteristics and known uses of datasets.

Extension to other software engineering studies. Even when our assessment model was based on the analysis of MSR studies, there are reasons to believe it could be extended beyond, to other empirical software engineering studies. As we have shown in several parts of this and the original paper, there are many aspects in which our approach is coincident, or very similar, to other more general approaches in the software engineering realm. Most of them identify (with different names) different kinds of data and methods, that maybe could be structured the way we did for MSR studies. If this can be done, it could mean the standardization of detailed information about reproducibility assessment for empirical software engineering studies. However, still more work is needed to know if this is feasible, even for a certain fraction of all the different kinds of empirical software engineering studies.

Table 14
Consequences of validation for different cases of variation of reproducibility elements, adapting the discussion in [5].

Variation of reproducibility elements	Function of validation	Validity threat addressed
All elements unaltered	Control for sampling error	Conclusion validity
Change method or parameters elements	Understand operational limits	Construct or internal validity
Data source or raw or processed dataset	Understand validity for different samples	External validity
All elements changed (or most of them)	Control for design error	Beyond study threats

Table 15
Self-assessment reproducibility report for the empirical study described in Section Section 5. Note: The datasource provides papers only to subscribers, with no permission for copying them, or redistributing them.

	Identification	Description	Availability	Persistence	Flexibility	Assessment
Data source	Complete	Detailed	Partial	Likely	–	D+
Retrieval meth.	Complete	Textual	Public	Likely	Complete	D+*
Raw dataset	No	N/A	N/A	N/A	N/A	–
Extraction meth.	Complete	Textual	Public	Likely	Complete	D+*
Parameters	Complete	Complete	–	–	–	U
Processed dataset	Complete	Detailed	Public	Likely	Complete	U+*
Results dataset	Complete	Detailed	Public	Likely	Complete	U+*

6.7. How to use our assessment method

We think our assessment method can help to improve the review (including the self-assessment) of MSR research studies, giving chances that conferences, journals and other venues interested in promoting reproducible research, production of reusable artifacts, and validation studies, could take advantage of. These are some ideas of how it could be used:

Artifacts review. Many conferences in the software engineering realm are including artifact review, either as an integral part of the review process, or as a separate process after papers are accepted. In some cases, this is linked with the ACM Badges initiative, which means that papers whose artifacts were reviewed, may get a badge. In all these scenarios, when papers are about MSR research, our guidelines could be the base for a detailed, homogeneous and comprehensive review of the reproducibility elements used in a study. In fact, they could be linked to granting ACM badges, since we already showed how their evaluation is not that different from our assessment. In any case, we suggest that reproducibility assessments are published for all papers with badges, so that people (including authors of the papers) can know in detail what to expect from the reproducibility elements of a paper.

Use in reproduction packages. Self-assessment on the reproducibility of a study can be very interesting, even if there is no reproducibility review. Authors could just use the criteria we presented in our original paper for evaluation the reproducibility attributes of each element of their paper, and include that information, prominently, in their reproduction package. This would help to make them more conscious about the reproducibility of their studies, and to tune expectations of other researchers checking the reproduction package. In addition, it would also help reviewers, be them performing a regular review or an artifact review: the self-assessment would direct their review efforts, maybe to check it, to complement it, or in some cases, to use it as such.

Use for describing validation studies. Validation studies need to specify which parts of the original study are reusing, and which ones are changed. In addition to (or maybe substituting) a textual description, a reproducibility characterization such as the one summarized in Section 3.3 could be useful to formally summarize the details of each reproducibility element. This would help reviewers and readers in general to quickly understand the kind of validation being performed. In addition, using the information in Table 14, they could quickly understand the function of the validation that can be expected, and therefore, the validity threat addressed.

Use for reusable tools and datasets. Following the idea of the model cards of ML models, tools and datasets suitable for use by third parties could have a descriptive card, with a part of it being a reproducibility report detailing which reproducibility elements of a MSR study includes or (in the case of tools) can automatically produce. This would help to quickly understand the advantages of using those tools if verification of a study is important.

7. Conclusions

The state of the art in assessing the reproducibility of empirical studies has improved clearly during the last decade. The analysis of the kinds of validation studies, and of the functions of validation in the field is also more detailed now. In addition, new approaches to improve reproducibility have also appeared. However, evaluation of reproducibility of MSR studies is still shallow, and there is still room for better characterization of validation studies. However, the reproducibility of MSR studies has clearly improved since the situation around 2010, and the reuse of some kinds of datasets has become commonplace.

After the analysis of the current situation, we propose the method described in our original paper for a more nuanced description of the reproducibility elements of a MSR study. It could be used in the review process of artifacts related to the reproducibility of studies, and for characterizing validation studies based on the reuse of elements from the original. In fact, we have shown how it can be used in assessing the reproducibility of recent MSR papers, ensuring that the method remains valid. Thanks to this assessment, we have also learned that the reproducibility of MSR studies have improved clearly since 2010, that reproduction packages have become the norm, and that reusing datasets and methods from third parties is usual.

We also propose the evaluation of the consequences of validation, depending on the characteristics of the reproducibility elements of a study. This proposal is based on the results of studies on the functions of validation studies. In addition, we also propose the evaluation of reusability characteristics of toolsets and datasets useful in MSR studies.

In summary, we think our analysis proves that the approaches in the original paper are still valid, and they could be used to improve reproducibility and validation of research results in the field.

Reproducibility and data availability

Datasets and source code of the methods for analyzing them are available in a reproducibility package.[4] In addition, we have run a self-assessment reproducibility report, provided in Table 15, for the empirical study described in Section 5, using the method described in [1].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. However, authors also declare that they know many of the autors of the MSR papers analyzed in the study presented, and have been coauthors (in other papers), and have participated in common research projects with some of them.

Acknowledgments

The research presented in this paper has been supported in part by the Government of Spain, through project "Dependentium" (PID2022-139551NB-I00).

References

[1] Jesus M. Gonzalez-Barahona, Gregorio Robles, On the reproducibility of empirical software engineering studies based on data retrieved from development repositories, Empir. Softw. Eng. 17 (1–2) (2012) 75–89, http://dx.doi.org/10.1007/s10664-011-9181-9.

[2] Gregorio Robles, Replicating MSR: A study of the potential replicability of papers published in the mining software repositories proceedings, in: 2010 7th IEEE Working Conference on Mining Software Repositories (MSR 2010), 2010, pp. 171–180, http://dx.doi.org/10.1109/MSR.2010.5463348.

[3] Victor R. Basili, Forrest Shull, Filippo Lanubile, Building knowledge through families of experiments, IEEE Trans. Softw. Eng. 25 (4) (1999) 456–473.

[4] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, Anders Wesslén, Experimentation in Software Engineering, Springer Science & Business Media, 2012.

[5] Omar S. Gómez, Natalia Juristo, Sira Vegas, Understanding replication of experiments in software engineering: A classification, Inf. Softw. Technol. 56 (8) (2014) 1033–1048, http://dx.doi.org/10.1016/j.infsof.2014.04.004.

[6] Martin Shepperd, Nemitari Ajienka, Steve Counsell, The role and value of replication in empirical software engineering results, Inf. Softw. Technol. (ISSN: 0950-5849) 99 (2018) 120–132, http://dx.doi.org/10.1016/j.infsof.2018.01.006, URL https://www.sciencedirect.com/science/article/pii/S0950584917304305.

[7] Lech Madeyski, Barbara Kitchenham, Would wider adoption of reproducible research be beneficial for empirical software engineering research? J. Intell. Fuzzy Systems 32 (2) (2017) 1509–1521.

[8] Sira Vegas, Natalia Juristo, Ana Moreno, Martín Solari, Patricio Letelier, Analysis of the influence of communication between researchers on experiment replication, in: ISESE'06: Proceedings of the 2006 ACM/IEEE International Symposium on Empirical Software Engineering, ISBN: 1-59593-218-6, 2006, pp. 28–37.

[9] ACM (Ed.), Artifact Review and Badging, 2020, URL https://www.acm.org/publications/policies/artifact-review-and-badging-current.

[10] ACM (Ed.), Artifact Review and Badging, 2020, URL https://www.acm.org/publications/policies/artifact-review-and-badging.

[11] Joint Committee for Guides in Metrology (Ed.), International vocabulary of metrology, third ed., in: JCGM, (200:2012) 2012, URL https://www.iso.org/sites/JCGM/VIM-JCGM200.htm.

[12] Chao Liu, Cuiyun Gao, Xin Xia, David Lo, John Grundy, Xiaohu Yang, On the reproducibility and replicability of deep learning in software engineering, ACM Trans. Softw. Eng. Methodol. (ISSN: 1049-331X) 31 (1) (2021) http://dx.doi.org/10.1145/3477535.

[13] Jonathan L. Krein, Charles D. Knutson, A case for replication: synthesizing research methodologies in software engineering, in: Proceedings of the 1st International Workshop on Replication in Empirical Software Engineering Research (RESER 2010), 2010, pp. 1–10.

[14] Claudia Ayala, Burak Turhan, Xavier Franch, Natalia Juristo, Use and misuse of the term "experiment" in mining software repositories research, IEEE Trans. Softw. Eng. 48 (11) (2022) 4229–4248, http://dx.doi.org/10.1109/TSE.2021.3113558.

[15] Claes Wohlin, Martin Höst, Kennet Henningsson, Empirical research methods in software engineering, in: Empirical Methods and Studies in Software Engineering: Experiences from ESERNET, in: Lecture Notes in Computer Science, LNCS 2765, Springer, 2003, pp. 7–23.

[16] Steve Easterbrook, Janice Singer, Margaret-Anne Storey, Daniela Damian, Selecting empirical methods for software engineering research, in: Forrest Shull, Janice Singer, Dag I.K. Sjøberg (Eds.), Guide to Advanced Empirical Software Engineering, Springer London, London, ISBN: 978-1-84800-044-5, 2008, pp. 285–311, http://dx.doi.org/10.1007/978-1-84800-044-5_11.

[17] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, The KDD process for extracting useful knowledge from volumes of data, Commun. ACM 39 (11) (1996) 27–34.

[18] Tao Xie, Suresh Thummalapenta, David Lo, Chao Liu, Data mining for software engineering, Computer 42 (8) (2009) 55–62, http://dx.doi.org/10.1109/MC.2009.256.

[19] Hadi Hemmati, Sarah Nadi, Olga Baysal, Oleksii Kononenko, Wei Wang, Reid Holmes, Michael W. Godfrey, The MSR cookbook: Mining a decade of research, in: 2013 10th Working Conference on Mining Software Repositories (MSR), 2013, pp. 343–352, http://dx.doi.org/10.1109/MSR.2013.6624048.

[20] Omar S. Gómez G., Natalia Juristo, Sira Vegas, Replication, reproduction and re-analysis: Three ways for verifying experimental findings, in: Proceedings of the 1st International Workshop on Replication in Empirical Software Engineering Research (RESER 2010), Cape Town, South Africa, 2010.

[21] Paul Ralph, Nauman Ali, Sebastian Baltes, Domenico Bianculli, Jessica Diaz, Yvonne Dittrich, Neil Ernst, Michael Felderer, Robert Feldt, Antonio de Filieri, Breno Bernarda Nicolau França, Carloa Alberto Furia, Greg Gay, Nicolas Gold, Daniel Graziotin, Pinjia He, Rashina Hoda, Natalia Juristo, Barbara Kitchenham, Valentina Lenarduzzi, Jorge Martínez, Jorge Melegati, Daniel Mendez, Tim Menzies, Jefferson Molleri, Dietmar Pfahl, Romain Robbes, Daniel Russo, Nyyti Saarimäki, Federica Sarro, Davide Taibi, Janet Siegmund, Diomidis Spinellis, Miroslaw Staron, Klaas Stol, Margaret-Anne Storey, Davide Taibi, Damian Tamburri, Marco Torchiano, Christoph Treude, Burak Turhan, Xiaofeng Wang, Sira Vegas, Empirical standards for software engineering research, 2020, http://dx.doi.org/10.48550/arXiv.2010.03525, arXiv preprint arXiv:2010.03525, arXiv:2010.03525 [cs.SE].

[22] Randall J. LeVeque, Ian M. Mitchell, Victoria Stodden, Reproducible research for scientific computing: Tools and strategies for changing the culture, Comput. Sci. Eng. 14 (04) (2012) 13–17.

[23] Robert Dyer, Hoan Anh Nguyen, Hridesh Rajan, Tien N. Nguyen, Boa: A language and infrastructure for analyzing ultra-large-scale software repositories, in: 35th International Conference on Software Engineering (ICSE), IEEE, 2013, pp. 422–431.

[24] Roberto Di Cosmo, Stefano Zacchiroli, Software heritage: Why and how to preserve software source code, in: IPRES 2017 - 14th International Conference on Digital Preservation, Kyoto, Japan, 2017, pp. 1–10, URL https://hal.archives-ouvertes.fr/hal-01590958.

[25] Georgios Gousios, Diomidis Spinellis, GHTorrent: GitHub's data from a firehose, in: 2012 9th IEEE Working Conference on Mining Software Repositories (MSR), IEEE, 2012, pp. 12–21.

[26] Yuxing Ma, Chris Bogart, Sadika Amreen, Russell Zaretzki, Audris Mockus, World of code: an infrastructure for mining the universe of open source VCS data, in: 2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR), IEEE, 2019, pp. 143–154.

[27] Georgios Gousios, Bogdan Vasilescu, Alexander Serebrenik, Andy Zaidman, Lean GHTorrent: GitHub data on demand, in: Proceedings of the 11th Working Conference on Mining Software Repositories, 2014, pp. 384–387.

[28] Fabian Trautsch, Steffen Herbold, Philip Makedonski, Jens Grabowski, Addressing problems with replicability and validity of repository mining studies through a smart data platform, Empir. Softw. Eng. 23 (2) (2018) 1036–1083.

[29] Santiago Dueñas, Valerio Cosentino, Jesus M. Gonzalez-Barahona, Alvaro del Castillo San Felix, Daniel Izquierdo-Cortazar, Luis Cañas-Díaz, Alberto Pérez García-Plaza, GrimoireLab: A toolset for software development analytics, PeerJ Comput. Sci. 7 (2021) e601.

[30] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru, Model cards for model reporting, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 220–229.

[4] Reproducibility package: https://doi.org/10.5281/zenodo.8022040.