Empir Software Eng (2016) 21:1302–1345
DOI 10.1007/s10664-015-9386-4

Experimentation with dynamic simulation models in software engineering: planning and reporting guidelines

Breno Bernard Nicolau de França[1] · **Guilherme Horta Travassos**[1]

Published online: 6 June 2015

Abstract Simulation-based studies (SBS) have become an interesting investigation approach for Software Engineering (SE). However, the reports on experiments with dynamic simulation models found in the technical literature lack relevant information, hampering the full understanding of the procedures and results reported, as well as their replicability. Apart from the limitations on the length in conferences and journal papers, some of the relevant information seems to be missing due to methodological issues not considered when conducting such studies. This is the case of missing research questions and goals, lack of evidence regarding the dynamic simulation model validity, poorly designed simulation experiments, amongst others. Based on findings from a previous *quasi*-systematic literature review, we propose a set of reporting guidelines for SBS with dynamic models in the context of SE aiming at providing guidance on which information the report should contain. Furthermore, these guidelines were evolved to support SBS planning by identifying potential threats to simulation study validity and in making recommendations to avoid them, through qualitative analysis and external evaluation. Finally, we conducted different evaluations regarding both the reporting and planning guidelines, apart from using them to support the planning of a SBS as regards software evolution. A set of 33 reporting and planning guidelines for different stages of the simulation lifecycle and focused on the experimentation with dynamic simulation models have been put together. The first assessments point to a comprehensive set of guidelines, supporting a comprehensive preparation and review of the plans and reports from the studies, apart from the planning of a SBS focused on software evolution, potentially reducing the threats to the experimentation with the validity of dynamic simulation models. The 33 guidelines cannot be understood as separate groups for reporting and planning as they overlap in many aspects. The

Communicated by: Maurizio Morisio

✉ Breno Bernard Nicolau de França
brenofranca@gmail.com

Guilherme Horta Travassos
ght@cos.ufrj.br

[1] COPPE/Universidade Federal do Rio de Janeiro, P.O. Box 68511, Rio de Janeiro, Brazil

main goal is to use the guidelines to support the planning of a simulation-based study with dynamic models so that experimenters may identify potential threats to validity and produce relevant information for a complete simulation experiment report in advance. Despite their initial contribution to increase the validity of SBS, the reporting and planning of simulation-based experiments with dynamic models still has to be discussed and improved in SE. Therefore, additional assessments of this set of guidelines are needed to strengthen the confidence in their completeness and usefulness.

Keywords Reporting · Planning · Guidelines · Software engineering · Simulation-based studies

1 Introduction

Many scientific research areas have succeeded in their adoption of Simulation-Based Studies (SBS) as both an alternative and supplementary strategy to support experimentation. Engineering, Economics, Biology, and Social Sciences are examples of disciplines where investigation has been done with the use of this approach (Müller and Pfahl 2008). Success cases can also be found in the automotive industry (Thomke 2003), and in Criminology (Eck and Liu 2008), amongst others.

Simulation-Based Studies involve the undertaking of several activities such as problem formulation, data collection, conceptual and executable model development, verification and validation, experimental design, output data analysis, and others (Alexopoulos 2007). This way, it is important to address the types of studies we are considering as simulation. For that, we adopt the definition of Banks (1999), which describes a simulation as '*the imitation of the operation of a real-world process or system over time. Simulation involves the generation of an artificial history of the system, and the observation of that artificial history to draw inferences concerning the operating characteristics of the real system that is represented*'. One expression merits attention in this definition: 'over time', which may also be replaced by 'dynamic', in the sense that it focuses only on behavioural aspects. In other words, it says we are just interested on how a system performs or phenomena take place and on how their variables change in time. Therefore, static approaches such as regression models or the Monte Carlo method are not covered in our research.

The success of simulation in other research areas may be the main motivation for the initiatives seen in Software Engineering (SE). Amongst the advantages attributed to simulation studies are: the possibility of replicating experiments, controlling variables of interest, and doing relatively less risky studies when compared to the amount of risk involved in performing in vivo or in vitro experiments. Nevertheless, there are also issues to observe in this sort of study. The credibility of simulation experiments is strongly related to the validity of the simulation model used as an instrument, as well as the supporting data. Besides, there is a high cost associated to model development. So, researchers should be aware of the tradeoffs involving both the development of a simulation model and the amount of investigation it may support.

Even with the many interesting achievements in simulation studies, as found in the SE technical literature, it is also possible to identify other uncovered issues. In a recent *quasi*-Systematic Literature Review (qSRL) (de França and Travassos 2013a), we characterize SBS in the context of SE, pointing at experimental aspects we believe should get more attention from the SE community. For instance, threats to simulation study validity, mainly related to model validity and a lack of a predefined experimental design; informal output analysis, causing untimely conclusions; and a poor quality of the reports found in the technical literature.

From the aforementioned issues, the lack of relevant information in reports is the first target of this research paper. By addressing such issue, we aim at improving the understanding through the guidance on what information is essential for a SBS with dynamic model report in SE. Essential information such as a precise context definition, clear goals and research questions, a description of the behaviours to be simulated, procedures for model assessment, experimental design, and output analysis are examples of what should build a study report (de França and Travassos 2012).

As an immediate benefit of a high quality report, we can point out that it improves the understanding of SBS, which are naturally complex, and enables their replication, based on full model description, the environment where a simulation has been performed, the publication of available input datasets, and the experimental procedure containing the steps performed.

The quality of experimental reports for primary studies is already a concern for the Experimental Software Engineering (ESE) community, which has proposed general (Kitchenham et al. 2008) and specific guidelines (Jedlitschka et al. 2008) (Runeson and Höst 2009) (Carver 2010). In this paper, we present the preparation of a set of guidelines for reporting dynamic simulation studies in SE, i.e., the result of the whole research, including the model development and use. The goal of this set is to improve the orientation of researchers in reporting activities. Apart from that, we present an evaluation of these reporting guidelines and analyse them against recent reports identified in the SE technical literature.

In addition, we started to work on these guidelines from the reporting to the planning perspective, aiming at reaching the results proposed in the reporting guidelines. However, planning studies without a better perspective of possible threats is a risky proposition. We therefore decided to extend the qualitative analysis of qSLR data aiming at identifying threats to validity for dynamic simulation studies in SE (de França and Travassos 2014b). The threats identified were linked to potential Verification & Validation (V&V) procedures supporting their removal, to highlight the potential benefits of applying V&V procedures prior to conducting simulation experiments. It allowed us to suggest how to improve result confidence by handling some threats.

As the reporting and planning guidelines overlap in many aspects, we suggest not to understand them as distinct subgroups, as they actually share the same discussions and examples. Even though, the research methodology presented in next section indicates how they have been identified in the different stages of our research.

As a proof of concept, we managed to apply these guidelines to organize a protocol for an actual simulation experiment involving a System Dynamics (SD) model for the observation of software quality decay (Araújo et al. 2012) in the context of software evolution. These results are presented in the next sections of this paper.

The remainder of this paper is organized as follows: Section 2 presents the relevant background research for the current results, along with the research methodology; Section 3 presents related work, including different sets of guidelines for empirical and simulation studies; Section 4 presents the reporting guidelines proposed for simulation studies in the context of Software Engineering; Section 5 presents the evaluations of the proposed reporting guidelines; Section 6 presents the planning guidelines as an evolution of the reporting ones;

Section 7 presents a proof of concept for the proposed planning guidelines; and Section 8 presents the conclusions and future directions of our research.

2 Background and Research Methodology

Our previous experience in dynamic simulation studies in SE lies mainly on two SD models. One is a scenario-based software project management model to support risk assessment (Barros et al. 2004). The other one regards software evolution, aiming at allowing the observation of software quality decay over successive maintenance cycles (Araújo et al. 2012). Both models were developed using the *Illium* tool, based on its own SD language (Barros et al. 2002). Besides, our interests and research goals are beyond this experience as it includes understanding how the SE community has developed and used dynamic simulation models. We started by capturing knowledge in SE computer simulations, looking for methodologies in the technical literature to support the development and use of valid simulation models. This way, we expected to increase our orientation in the planning and execution of SBS, by characterizing how different dynamic simulation approaches had been previously applied to the SE context.

The research methodology for this work starts with an initial investigation consisting of the definition of problem and research questions, as pointed in the previous section. Figure 1 shows the activities that form this research methodology.

With this initial investigation, we performed an ad-hoc literature review to capture the common terminology and understanding of dynamic simulation in SE. At this stage, both general purpose and SE simulation books were consulted, as well as relevant (high mention rate) simulation papers. This material was helpful to establish the scope of the review. We first found the IMMoS approach for simulation model development and use, as proposed by (Pfahl and Ruhe 2002), a methodology specific for SD models, focused on the software process and project issues. The part on model experimentation in IMMoS methodology intends to be general as it focuses on model development and therefore does not present detailed information and guidance on using simulation models. Based on such information, we performed a qSLR (de França and Travassos 2013a, b) to characterize SBS in the context of SE.

The qSLR research protocol followed the guidelines proposed by Biolchini et al. (2005), also adopting the PICO strategy (Pai et al. 2004). The main goal of the review is to characterize how different simulation approaches have been applied in SE studies. Such characterization involves identifying adopted approaches, SE domains, model validation issues, simulation procedures and experimental design and output analysis. Pre-defined selection and extraction

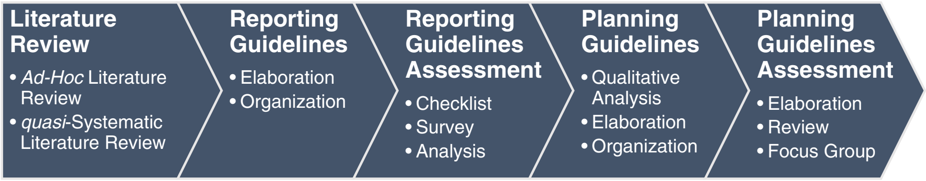


Fig. 1 Research methodology

procedures were defined and executed, followed by the quality assessment and the analysis of findings.

The review covers not only software process simulation, but also any study and model related to SE phenomena, including software product evaluation. It also includes model development and use (simulation experiments), being based mainly on three digital libraries: Scopus, EI Compendex, and the Web of Science, getting research papers until the execution date (April, 2011).

After information extraction from each of the 108 research papers selected [see Appendix A in (de França and Travassos 2013a, b)], it was possible to identify 88 simulation models, distributed among several SE domains. In brief, we identified 19 simulation approaches, 17 Software Engineering domains, 27 simulation tools or environments, 28 simulation model characteristics, 22 output analysis instruments, and nine procedures for the verification and validation of simulation models in the SE context. The most mature studies were identified using SD or Discrete-Event Simulation (DES) concerned with Software Process and Project Management. From the 19 simulation approaches, seven do not present any associated indication of validity. In other words, we could not observe any V&V procedure being performed on the models proposed using these approaches as the underlying simulation mechanism.

The experimental design is missing in 70.4 % of the research papers, and in 27.8 % of the papers it could mostly be inferred, as it was not explicit. We identified just two research papers that clearly explored experimental design issues (Houston et al. 2001) (Wakeland et al. 2004). This lack of detail on experimental design may contribute to the use of ad-hoc output analysis. Also, it reinforces the focus only on model development, disregarding model experimentation. However, it is an unexpected behaviour in simulation studies, as simulation models are expected to be proposed to perform experiments. We identified only 57 simulation experiments in the 108 research papers. It is unlikely that these experiments reach sound conclusions without proper experimental design and output analysis. It may be an effect of the lack of orientation on how to experiment with simulation models in SE.

Output analysis is mainly performed in an ad-hoc fashion. Although it is an expected issue, as the experimental design is missing, it also reinforces the concentration of efforts on model development, rather than on model experimentation (use). However, this is not a matter of choice, as simulation models should have been proposed to perform experiments. So, what are the conclusions of these 57 simulation experiments without a proper experimental design and output analysis? They may show a lack of orientation on how to experiment with simulation models, an even greater trait in the SE context.

Besides these methodological issues, we also identified some issues related to the lack of information in the reports, probably caused by the lack of orientation on what information should be used when reporting on this type of study. Such lack of information may lead to an impossibility to understand, replicate, audit, and also evaluate the quality of the results reaped in these studies. Thus, we move our research to propose what information should be available when reporting on SBS, i.e., the result of the whole research, including model development and use.

Using the findings from the review and existing reporting guidelines that focus on other types of study (empirical studies, controlled experiments, and case studies) and from other research areas (Medicine and Statistics), we organized a preliminary set of reporting guidelines (de França and Travassos 2012). This set (Section 4) was evolved through sequential evaluation initiatives, including a perspective-based review (section 5.1), using the instruments proposed by Kitchenham et al. (2008); a collaborative review (Section 5.2), which was structured as an online survey; and, finally set it all against technical literature reports (Section 5.3), as obtained from the systematic review update. All of these initiatives allowed us to get some feedback regarding the reporting guidelines' completeness and correctness.

Considering that we established a set of relevant information items to be reported, it is essential to understand when and how such information should be produced. Additionally, we saw that, if no plan is in place for the simulation study, the researchers are not likely to produce them, as much of it cannot be produced in retrospect. At this point of our research, we turned back to our primary goal, regarding orientation to support the planning and execution of SBS.

The next step in our research methodology aims at evolving the set of reporting guidelines in order to support the planning activities for simulation experiments. The additional guidelines do not mean embracing the task of simulation model development, but only the use of the simulation model or model experimentation (Balci 1990). This is justified by the existence of a methodological support for simulation modelling in the domain of SE, such as the IMMOS methodology (Pfahl and Ruhe 2002). For those readers interested on getting a broader view of the simulation process, please refer to (Balci 1990).

Back to the methodological issues, and aiming at supporting the elaboration of complete, coherent and effective simulation plans, the planning guidelines were developed based on findings from the qSLR, and additional information on the experimental design and threats to validity. For the experimental design information, we consulted the consolidated technical literature on Statistics from the classical (Montgomery 2008) to the simulation (Kleijnen et al. 2005) perspective. For the threats to validity, we did a qualitative analysis, using the Constant Comparative Method (Corbin and Strauss 2008), to identify reported threats to simulation studies in SE (de França and Travassos 2014b). Additionally, we analyzed how they could be mitigated using V&V procedures for simulation models and specific experimental design solutions. The result from this analysis is a subset of guidelines, shown in Section 6.

In the last stage of our methodology, we see the need for a qualitative evaluation of the planning guidelines, focusing on their usefulness and ease of use. As a qualitative evaluation we mean an observational study in which the subjects will have to apply the guidelines for a specific simulation experiment, and the researchers will collect and analyze data regarding such application. However, such a study requires great effort and considerable time, as well as subjects' availability.

3 Related Work

The Software Engineering community has produced some relevant initiatives regarding orientation on planning, executing and reporting empirical studies. These initiatives generally provide a set of important aspects to be considered when conducting empirical studies and, for each aspect, there is an associated discussion, as well as examples from the technical literature showing what and how a point should be addressed. For instance, some aspects from these guidelines involve the research context determination, the experimental design, data collection, and the presentation of results.

Kitchenham et al. (2002) proposed a preliminary set of guidelines to support researchers, reviewers and meta-analysts in the activities of design, conducting and evaluating of SE studies. This proposal has a large scope, presenting general guidelines that include many types of primary studies. Besides, authors point at the need for specific guidelines, for each research strategy.

This way, Jedlitschka et al. (2008) proposed guidelines for conducting and reporting controlled experiments in SE. In their guidelines they discuss different aspects such as redundant information, textual elements, and other aspects specific to controlled experiments, namely: experimental unit, instruments, study procedures, hypotheses, dependent and independent variables, and also experimental design, including the separation of groups of subjects, and the application of treatments. Finally, there are some general issues related to goals, data collection and analysis, but under the perspective of controlled experiments.

Runeson and Höst (2009) proposed similar guidelines for case studies in SE. In these guidelines, the authors address, apart from general aspects, specific characteristics of study cases aiming at presenting an in-depth view.

There are other initiatives such as guidelines for reporting on study replications (Carver 2010). We should point the need for such guidelines considering the heterogeneity and lack of reporting standards that may cause problems such as a misunderstanding of study results and conclusions, and difficulties to perform meta-analysis or any other method for aggregation.

For a perspective outside the SE community, we visited other research areas to expand our coverage. In Computer Simulation (Kleijnen 1975) (Balci 1990), Statistics (Ören 1981), Social Sciences (Rahmandad and Sterman 2012) and Medicine (Burton et al. 2006) we identified some orientations on which information should be considered when reporting on SBS.

Ören (1981) presents a series of concepts and criteria to evaluate the acceptability and credibility of SBS. The main concepts mentioned relate to data, to the model (both conceptual and executable) of the experimental design, and the methodology adopted to conduct the study. Balci (1990) presents guidelines for the success of SBS, organized according to the simulation model lifecycle and what the author calls "Credibility Assessment", a set of V&V activities concerning each lifecycle step. It is similar to the aspects presented by Shannon (1998), although Balci presents a more comprehensive process rather than isolated activities. In (Kleijnen 1975), the focus is on different techniques for the preparation and statistical analysis of the experimental design in simulation experiments. In Medicine, Burton et al. (2006) present a checklist emphasizing relevant issues for the elaboration of SBS research protocols. In general, there is a concern with simulation model validity and with a statistically adequate experimental design. In Social Science research, Rahmandad and Sterman (2012) published a set of reporting guidelines for SBS. In their proposal, they discuss three main aspects: model visualization for diagrams, model description for equations and algorithms, and simulation experiments design w.r.t. random numbers and optimization heuristics.

Ali and Petersen (2012) recently presented a consolidated process for conducting Software Process Simulation in industry in which they present some guidelines on how to perform the study for each activity. This way, it can identify some overlapping between their initial planning concerns such as using GQM for goal definition and the guidelines proposed in this paper. However, the remaining guidelines focus on model development rather than on model experimentation.

Some of the SE guidelines for empirical studies have reached a significant rate of usage. For instance, the guidelines mentioned above, as proposed by Kitchenham et al. (2002) are followed by Jedlitschka et al. (2008). Later, the case studies principles as proposed by Yin (2008) are followed by the specific guidelines from Runeson and Höst (2009). It is quite easy to identify SE research papers mentioning these guidelines as their methodological support. Actually, even secondary studies have been using these guidelines as reference for assessing rigour in research, referring to the precision or exactness of the research method use for its intended purpose, as proposed by Ivarsson and Gorschek (2011) and adopted in (Petersen 2011) (Barney et al. 2012). These guidelines intend to be drivers for research actions rather than mandatory recommendations. As they become mature, by identifying advantages in their use and influence on the quality of research protocols, their adoption tends to become natural.

The simulation guidelines presented in this paper do not represent any attempt at replacing the existing and common ones, but to consolidate shared aspects such as goals, design and analysis, by discussing them according to both dynamic simulation and SE perspectives. These perspectives clearly present particular characteristics, concerned with the advantages and drawbacks of using simulation to support experimentation for processes and products. In addition, raising issues on streamlining the validity to the context of dynamic simulation studies, something that has not been broadly discussed in SE. So, these guidelines also support the discussion on shared experimentation concepts in the simulation perspective, aiming at reducing the effort by simulation researchers and practitioners (mainly for those not well-experienced in computer simulation) on abstracting these concepts and then contextualizing them to SBS every time they need to successfully apply them.

4 Reporting Guidelines for Simulation Studies

In this section, we present an overview of the set of proposed guidelines concerned with the reporting on dynamic simulation-based studies in the context of SE research (Table 1). The terminology adopted can be consulted in the Glossary of Terms for Experimental Software Engineering.[1]

As a general suggestion, the audience the study is aimed at should be considered and its terms should be chosen accordingly. Also, this set of guidelines is organized in chained sections and this organization implicitly suggests a possible organization structure for the report. Finally, email addresses or other contact information should be provided to allow readers to ask researchers for further information or details on the study.

It is also important to point that each guideline should be taken by both its recommendation statement and the associated discussion and examples. The discussion and examples often try to put to together the perspectives of dynamic Simulation-Based Studies and the SE research area. The full descriptions of these reporting guidelines are available at (de França and Travassos 2014a, b). The next subsections briefly discuss the main aspects involved in the proposed guidelines.

The basis for these reporting guidelines is a combination of knowledge acquired in the technical literature and the authors' reasoning regarding dynamic simulation in SE. The sources of information come in most cases (SG3-6, SG9-17, and SG19) from a secondary analysis done in the qSLR (de França and Travassos 2013a, b) dataset. General reporting guidelines (SG1-2, SG8, SG18 and SG20-22) were added after the evaluation based on the approach proposed by Kitchenham et al. (2008), as presented in Section 5.1. One guideline (SG7) was extended from Ören (1981), including technical aspects as it originally discusses aspects such as costs, schedule and resources in a simulation project. Additionally, we identify on Table 1 (column *Refs*) examples of sets of guidelines that

[1] http://lens-ese.cos.ufrj.br/wikiese/index.php/Experimental_Software_Engineering_-_Glossary_of_Terms

Table 1 Simulation reporting guidelines overview

ID	Guideline statement	Refs
Report identification		
SG1	Proper title and keywords should objectively identify the simulation study report, as well as have a structured abstract summarizing the report contents.	A
From context to research questions		
SG2	Context where the simulation study is taking place should be described in full.	ABCD
SG3	Explicitly state the problem that drives the simulation study, so that research questions can be derived.	AFJ
SG4	Clearly state the simulation study goals and scope.	ACDGHJ
SG5	Present the research questions derived from established goals.	ABCD
SG6	Clearly state the null and alternative hypotheses from research questions.	AB
Simulation feasibility		
SG7	Present the justifications for considering simulation studies as the ideal or feasible strategy.	FGJ
Background and related work		
SG8	Present only essential background knowledge and also the related works	A
Simulation model and validation		
SG9	Have a detailed description of both conceptual and executable simulation models, as well as their variables, equations, input parameters, and the underlying simulation approach.	FGJI
SG10	Gather as much evidence as possible on simulation model (conceptual and execution) validity.	FGJ
Subjects		
SG11	Characterize the subjects involved in the simulation study as well as their training needs.	ABCD
Experimental design		
SG12	Experimental design (matrix), including independent and dependent variables and how levels are assigned to each factor should be reported.	ABCDEF
SG13	Describe the selected simulation scenarios and the criteria used to identify them as relevant.	EHI
SG14	The number of runs, along with the rationale to determine it should be reported.	EGHI
Intermediate experimental trial		
SG15	Describe which and how intermediate measures are stored between simulation trials to be used in the final analysis.	H
Supporting data		
SG16	Assess, whenever possible, the data used to support the simulation model development or SBS.	EFI
Simulation supporting environment		
SG17	Describe the simulation environment, including supporting tools, associated costs, and decision for using a specific simulation package.	GHI
Output analysis		
SG18	Procedures and instruments for output analysis should be reported as well as the underlying rationale.	ABCEHI
Threats to validity		
SG19	Always report the threats to study validity, limitations and non-verified assumptions.	ABC
Conclusions and future works		
SG20	Main results/findings should be identified and summarized, as well as the conclusions arising from the results.	ACDFH
SG21	Applicability issues should be addressed in the report, considering organizational changes and associated risks.	A
SG22	Point out future research directions and challenges after current results.	A

Refs: A (Jedlitschka et al. 2008); B (Kitchenham et al. 2002); C (Runeson and Höst 2009); D (Carver 2010); E (Kleijnen 1975); F (Balci 1990); G (Ören 1981); H (Burton et al. 2006); I (Rahmandad and Sterman 2012); J (Ali and Petersen 2012)

also share similar concerns for each reporting guideline we proposed, but not covering specifics from both simulation and SE. Please notice that the references on Table 1 do not represent the sources of the proposed guidelines. However, we used them to point that these aspects are considered in other science areas and other type of studies in the SE community.

4.1 Study Definition

Simulation-based studies may be performed both in virtuo and in silico environments (Travassos and Barros 2003). In virtuo experiments stand for studies where human subjects interact with a computerized environment, while in silico experiments stand for studies where both subjects and the environment are represented by computerized (simulation) models. These two kinds of environments are under the scope of the proposed guidelines. In both alternatives, the object of the study is always related to the simulation model. So, depending on the goal of the study, the object of the study may be the simulation model itself or the phenomenon/system/process, which the model abstracts over time. Some contextual factors rely on the collected data that supports the simulation model development and calibration. This is especially true in SE, where the context of software projects, the human nature of SE activities, and the amount of unknown variables can affect the results of the studies.

Contextual information (environment and pre-requisites) is also important when the results observed through a dynamic simulation study need to be implemented in the real context. In this case, the entire context assumed by the simulation model should be guaranteed or handled in the real context. Otherwise, it will be necessary to change the target processes, team training, incorporating new techniques or tools, and applying them to the right kind of systems/ applications.

Dybå et al. (2012) propose the use of a broad perspective approach for the so-called *omnibus* context (SG2). In brief, this proposal describes the context in such a way that the study report allows answering the following type of research question: “*What* technology is most effective for *whom*, performing *that* specific activity, on *that* kind of system, under *which* set of circumstances?”

Once the context information has been gathered, the problem (SG3) should then be stated and described as to how it was identified in such a context. Problems may arise from a specific critical situation or from repeated situations where the solution has a complex implementation or requires an expensive alternative. For problem statement, we adopt a template proposal[2] based on the following structure:

Statement 1 (Description of ideal scenario). However (or other adversative conjunction),
Statement 2 (The reality of the situation). Thus (or other conclusive conjunction),
Statement 3 (The consequences for the involved people).

Defining the goals (SG4) is the first step, after establishing the problem. It needs to be described in a clear way, leaving no doubt as to what is to be achieved, in the same way as it occurs with other Software Engineering studies, in which the definition of the goals uses the GQM approach (Basili 1992). Besides, non-structured goal definitions may prevent one from getting to the right point, but it is the way in which goals are described in simulation research papers in the SE technical literature. Therefore, the common goals for SBS should include:

[2] http://www.personal.psu.edu/cvm115/proposal/formulating_problem_statements.htm

developing a basic understanding (characterization) of a particular simulation model of the phenomenon, finding robust or optimum decisions, or comparing the merits of various decisions.

SBS goals should match the capabilities of the simulation model. In other words, the simulation model should be able to support the answers to the research questions through the output data, and its input parameters (variables or constants) should allow the desired scenario configuration.

In the context of Software Process Simulation with SD, the IMMoS methodology (Pfahl and Ruhe 2002) provides a more specific template that is similar to GQM goal definition, structured in five dimensions, as in Table 2. The scope dimension means the modelling boundary and granularity, similar to the object of study in GQM. Purpose, role and environment dimensions are identical to their alternatives in GQM. And dynamic focus represents the particular dynamic behaviour in the focus of interest, similar to quality focus in GQM.

According to Davis et al. (2007) and Kleijnen et al. (2005), research questions definition drives the simulation research, preventing the loss of focus by the researcher and the keeping of coherent methodological steps. This way, once following the GQM approach to drive goal definition, and deriving research questions (SG5), the next step is to define the metrics based on which the questions should be answered. The metrics definition allows one to 'ask' the research questions as hypotheses (SG6), which should be submitted to statistical tests.

Assuming such a study definition has been done and documented, it is important to assess the feasibility of the simulation as a candidate approach to solve or investigate the problem (SG7). As far as we know, Balci (1990) is the only resource available in the technical literature supporting this kind of analysis, suggesting the use of some questions as indicators, such as cost, time, benefits and the relationships amongst them, which naturally limit the field of observation. To overcome this limitation, we have discussed additional issues, shown below.

The goals of the simulation study should be beyond the getting of a value for an output variable. Simulation outputs also include a rationale, an explanation or a chain of changes in the system that produces the output values, often represented by high-order effects. Thus, simulation studies for SE should explain how the phenomenon (events and variables) occurs and what changes in processes, products or people may give a suitable solution. In this sense, we recommend additional questions to support the decision-making on the deciding to perform simulation studies. Therefore, one has to focus on more technical constraints regarding simulation model development and experimentation. The system or phenomenon under investigation should be observable in some sense. So what are the available instruments and procedures for data collection? Are the occurrence

Table 2 Goal definition templates from GQM and IMMoS

Dimension in GQM	Dimension in IMMoS
Object of study	Scope
Purpose	Purpose
Quality focus	Dynamic focus
Viewpoint	Role
Environment	Environment

risks (including loss of money or time, reaching an irreversible state of the system, safety) of the real phenomenon high? Also, data should be available to cover for statistical issues and the calibration of variables and equations involved in common approaches such as SD and DES.

4.2 Simulation Environment and Model Validity

The guidelines focus on the reporting of dynamic simulation experiments. Model development issues are out of scope, except those aspects at the frontier between model development and use. For the purpose of planning and reporting such experiments, it is important to know the model in detail (SG9). It is part of the required planning knowledge to understand the underlying simulation approach, the conceptual model, including its variables, parameters and associated metrics, as well as the underlying assumptions and calibration procedures. Furthermore, model description is useful to supplement the information on the experimental design and on how values for input parameters in each simulation run are determined.

The reader of the report expects diagrams, equations, and textual descriptions. Diagrams are useful for presenting the whole idea as well as the conceptual simulation model. Equations allow the possibility of replicating the model in other simulation tools. Finally, a text description supplements and clears any doubt about the previous ones.

The concern with model validity should also be addressed (SG10), as SBS validity is highly affected by the validity of the simulation model. It is a reflection of the nature of a computer-based controlled environment where the phenomenon under investigation is observed essentially through the execution of the simulation model. This way, the only possible changes are those to the input data or the simulation model. Thus, if the model used cannot be considered valid, invalid results will be obtained regardless of the mitigation actions taken to deal with other possible validity threats. In other words, the simulation model itself represents the main threat to study validity.

As mentioned in guideline SG10, the evidence on model validity means the experimenter should be aware of the initiatives (previous reports and research papers) to submit the simulation model to V&V procedures, and understand their results. In the case where such validation references are absent, these procedures should be performed to ensure model validity, exposing the results as well as the decisions that guided the validation process.

It is not usual to identify the use of performance measures such as bias, accuracy, coverage, and confidence intervals in SE simulation studies. The importance of such measures relies on the possibility of using them as benchmark criteria to compare simulation models and analyse the risks assigned to SBS conclusions. Burton et al. discuss how to calculate such measures (Burton et al. 2006).

Also, the opportunity to gather empirical evidence from the technical literature as one V&V 'procedure' is an important step when developing simulation models for experimentation, as such evidence does not rely only on expert opinion or ad-hoc observation of the phenomenon under study. Empirical evidence can support the existence of properties in the simulation model, as well as model assumptions.

The study environment (SG17) should be made clear when planning and reporting SBS. It entails the simulation model itself, datasets, data analysis tools, and simulation tools/packages. Besides, the characterization of human subjects (SG11) should be done, as it can influence the interpretation of in virtuo results. This way, the level of expertise, number of subjects per group (treatment and control, when applicable) and any other relevant characteristic should be included in the study plan and considered in the subjects' assignment process to the experimental units whether this is made randomly or not, for example. Additionally, the training sessions and their costs should also be planned. With computerized subjects, their behaviour model, configuration parameters, and process of assignment should also be considered when preparing the experimental design, if such behaviour can be clearly identified in the simulation model. Also, it is possible that the subjects' behaviour may be implicitly embedded in the simulation model when dealing with in silico environments.

4.3 Experimenting with Dynamic Simulation Models

Having reached a scenario of model understanding and validity checks, the experimental design issues should be considered for reporting and planning purposes (SG12). It involves the definition of a causal model, establishing a relationship between independent (or factors) and dependent variables, in a cause-effect nature. During the experiment, design factors may be held constant or allowed to vary. Additionally, interest factors may be: controllable, which can be measured and can vary; uncontrollable, just possible to measure; and noise factors, the ones we cannot measure and that 'naturally' vary.

The causal model should be derived from the research questions and should reflect part of or the whole simulation model. It is often represented by a design matrix that includes the factors and treatments for each factor. In this matrix, every row is called a design point or a scenario, which is a combination of different levels for each factor (Kleijnen et al. 2005).

It is also important to identify control and treatment groups when doing controlled experiments using simulation models as instruments. For instance, validated models under known conditions can be assumed as control and the new model (or new versions) to be evaluated or experimented (under the same conditions) can be assumed as the treatment. Another possibility is to use distinct datasets as factors, with the simulation model remaining constant. This way, different calibrations representing the different simulation scenarios can be compared and should be reported (SG13).

The number of simulation runs (SG14) should be based on the selected simulation scenarios and on the simulation model's deterministic or stochastic nature. Each selected scenario consists of an arrangement of experimental conditions where possible factors are assigned to one specific level. The more simulation scenarios involved in the study, the more simulation runs are needed. A detailed discussion on how to determine the number of simulation runs, based on factorial designs, can be found in (Houston et al. 2001) and (Wakeland et al. 2004).

When using stochastic models, the use of random variables should also be taken into account as a confidence interval should be estimated from the sample size to determine the number of simulation runs (or replications). Such a calculation can be found in (Burton et al. 2006). Replication is achieved by using different pseudo-random numbers (PRNs) to simulate the same scenario. In this case, the output is a time series, which has auto-correlated observations (Kleijnen et al. 2005).

In as much as simulation models play the most important role in simulation experiments, supporting data availability is crucial for the feasibility of the studies and also when reporting on them (SG16). Simulation models need to be calibrated, requiring data for the generation of equations and parameters, and to determine random variable distribution. Therefore, it is important to determine the type of data: real or synthetic (Ören 1981). If synthetic data has been used, some evidence should be presented to guarantee data validity, i.e., the report should answer questions such as ‘How far is the simulated data from real-system data?’ and show indicators of this gap.

Data collection should be planned to also avoid measurement errors, promoting the collection of data as soon as it becomes available. After the collection, quality assurance procedures should be carried out to verify their consistency and accuracy, avoiding the inclusion of outliers or incomplete data. If the simulation model needs to be calibrated, it is important to report whether it was calibrated or not, including the procedure used to do it and its results.

Another important aspect relies on raw data publication, despite its being rarely reported, basically for two reasons: (1) most papers report that it is not possible to present raw data as it is confidential, and (2) since simulation studies usually involve a large amount of data, it may not fit conference or journal paper formats. Even so, raw data should, when possible, be reported or made available by consulting the authors or by making it available as a downloadable source.

4.4 Simulation Results and Conclusions

In the context of Software Engineering, the output analysis of simulation-based studies (SG18) is mostly done by using charts (de França and Travassos 2013a, b). On the other hand, there are fewer cases where we can find statistical (hypothesis) tests or descriptive stats.

The simulation study protocol should contain the procedures and instruments to be used in the analysis of simulation results. Simulation runs often produce large volumes of data, distributed in different output variables. The output data analysis procedure and instruments should be adequately chosen, as statistical instruments (such as charts) and methods have many assumptions and restrictions.

Assumptions on the independence of variables and on how data is distributed should be carefully observed, to adopt the correct charts, statistical measures, and tests. Simulation experiments use such statistical measures for accuracy, for instance. Mean Magnitude of Relative Error and Balanced Relative Error are examples of such measures (Foss et al. 2003). Charts often assume that the data is organized in a particular way; for example, Sequential Run Charts (Florac and Carleton 1999) assume the data is chronologically ordered. Specific hypothesis tests assume normally distributed data or homoscedastic distributions. These properties should be guaranteed in order to use such instruments when doing output analysis. Also, the evidence that supports how these properties are reached should be given.

Output analysis concentrates efforts on understanding and quantifying trends for output variables. Still, it helps to check the statistical correctness of the results. However, simulation experiments need additional analysis, such as threats to validity, including the model and experimental design validity (SG19).

SBS protocols need, as any other empirical study, to mitigate and discuss possible threats to study validity. Common types of experimental validities are closely related to simulation model validity (de França and Travassos 2014b). So, such a model should be valid to ensure that the study can represent the actual phenomena. The SE community has discussed threats to validity, and most of the reported threats concerned with in vitro or in vivo experimentation have already been described in (Wöhlin et al. 2012). Most of them have to be considered when planning simulation studies, especially when considering in vitro experiments where the human nature may impose risks to the study. Still, new situations emerge for in silico experiments. Either known threats appear in a different outlook, or specific threats to such environments affect result validity. Here, we concentrate our perspective on these new situations.

According to Davis et al. (2007), simulation improves construct and internal validity, by accurately specifying and measuring constructs (and the relationship between them) and the theoretical logic that is enforced through the discipline of algorithmic representation in software, respectively.

Raffo (2005) and Garousi et al. (2009) mention model validity in a similar way. They take several perspectives into account, such as model structure, supporting data, input parameters and scenarios, and simulation output. We understand that these aspects are extremely relevant, but are not the only ones, as study validity goes beyond the simulation model (de França and Travassos 2014b). It is also important to consider the simulation's experiment design.

External and conclusion validity should be accomplished with the application of adequate statistical tests over the model outputs. In SBS, external validity concerns the possibility of reproducing empirical behaviours and consistent behaviours across different simulation studies. However, conclusion validity also relates to sample size, number of simulation runs, model coverage, and the degree of representation of the simulated scenarios for possible situations.

At the end of the report, the results/findings express the main contributions in a summary (SG20). The conclusions should be drawn upon the findings, establishing a link from the goals, using methods to achieve results that allow making conclusions. Additionally, the final discussion should include implications on the applicability (SG21) of the solution in real scenarios, e.g., practical use. Finally, the way ahead (SG22) should be mentioned in the report, pointing to further work and research challenges. It may also include hot topics and possible roadmaps for future research.

5 Reporting Guideline Evaluation

After defining a preliminary set for reporting SBS guidelines in the context of SE (de França and Travassos 2012), we took three approaches to its evaluation, blending multiple perspectives in the sequence they appear in the next subsections. The result from each evaluation was used as input for the subsequent one, as shown in Fig. 2.

These attempts at assessing the set of proposed reporting guidelines has two major quality focus points: completeness and correctness. As completeness, we mean they are capable of guiding researchers and practitioners in the identifying or providing of relevant and expected information on SBS reports. In other words, for each relevant and expected information item there is at least one reporting guideline covering it. As correctness, we mean that the contents

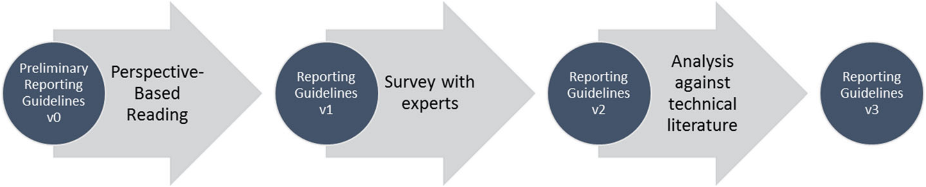

Preliminary Reporting Guidelines v0
Perspective-Based Reading
Reporting Guidelines v1
Survey with experts
Reporting Guidelines v2
Analysis against technical literature
Reporting Guidelines v3

Fig. 2 Reporting guideline evolution through subsequent evaluations

of the reporting guidelines are correct in terms of theoretical background and in accurately using concepts to discuss and exemplify each aspect.

5.1 Perspective-Based Review

With the preliminary version of the reporting guidelines, the authors did the first evaluation based on the approach proposed by Kitchenham et al. (2008). This approach is organized as a reading method inspired on both perspective-based and checklist reviews where, for each perspective, there is an associated checklist. The original perspectives are:

- Researcher: one who reads the report to find whether it offers important new information on a research area one is interested in;
- Practitioner/Consultant: one who provides information for use in the industry and wants to know whether the results in the report may aggregate value to one's company or clients;
- Meta-analyst: one who reads a report to obtain quantitative information that can be integrated with results from other equivalent experiments;
- Replicator: one who reads a report with the aim of repeating the experiment;
- Reviewer: one who reads a paper on behalf of a journal or conference to ensure that it is suitable for publication;
- Author: one who would be expected to use the guidelines directly to report on his/her experiment.

Amongst the mentioned perspectives, Kitchenham et al. (2008) proposed checklists for Researcher, Practitioner/Consultant, Meta-Analyst, Replicator, and Reviewer. However, we considered the Meta-Analyst perspective as quite questionable considering the SBS context, as studies on the same model should return the same results. Thus, the remaining checklists were applied.

We are aware of the validity threat regarding the guidelines' author playing the reviewer role. Nevertheless, we had no available experts to perform this review and, also considering the approach from Kitchenham et al. (2008), which requires a large amount of time from the experts, we chose to perform a first evaluation ourselves, using the checklists set in the original approach, without modifying them, attempting to reduce the associated bias, as it relies on an objective evaluation method by using the checklists (not proposed by the reviewers).

For each question in the checklists, we propose to use one of the following values:

- Attended: no issues identified, at least one guideline answers the question;
- Improvement Opportunity: any issue related to unclear sections, missing details, lack of theoretical foundations, excess of standardization, or document organization issues;
- Defect: any issue related to missing essential content or information, irrelevant recommendation for specific situations, ambiguous statements, and incorrect concepts.

The first version of the preliminary set had only 13 guidelines, organized into 11 sections. And the current version (Table 1) consists of 20 guidelines organized in 14 sections. This way, after the evaluation, there was a significant increase in the number of guidelines. Moreover, the results of the evaluation are shown in Table 3, according to the number of defects and improvement opportunities, for each adopted perspective.

Table 3 Evaluation results for proposed guidelines

Perspective	Number of questions	Improvement opportunities	Defects
Researcher	17	4	9
Practitioner/consultant	22	7	8
Replicator	9	3	1
Reviewer	7	1	3
TOTAL	55	15	21

The high number of defects occurred basically for two reasons: (1) the former version of the guidelines did not consider all the general reporting aspects, already used in other study strategies' guidelines; and (2) writing elements of reports such as title, structured abstract, and conclusions were not considered either. Just simulation-specific issues were taken into account in the first version. However, the improvement opportunities were related to a lack of detail or insufficient discussion for some guideline issues.

Additionally, not only new guidelines were introduced, but also some others were updated aiming at a better understanding and clarification of their contents. An example of update is given below:

First Version:
"The simulation (conceptual and execution) model must be tested against verification and validation procedures and the results should be reported. If the model has been validated before, a report of the results is expected. Also, the performance obtained in model validation should be presented."
Current Version:
"Present all possible evidence regarding the validity of the simulation model (conceptual and execution)."

It should be pointed that the current version of the statement is shorter and more objective than the first version. Therefore, the guideline details and outspread are discussed in a supplementary text, made available in the first version and that also evolved, along with examples of how such information was found in different situations in the SE technical literature.

5.2 Collaborative Review

With a more comprehensive and self-contained set of reporting guidelines, including discussions and examples from the technical literature, we released it as a technical report (de França and Travassos 2013b). We then organized a collaborative review as an online survey to obtain the opinion of simulation experts on the completeness and correctness of the proposed reporting guidelines, as it requires less effort and the experts can perform the tasks individually and remotely.

For this attempt, the main idea is to embrace the perspective of researchers and practitioners knowledgeable in SBS, both in Industry and Academia contexts. We expect to benefit from their feedback to improve the reporting guidelines. The following sections present the survey definition and results.

5.2.1 Survey Definition

The adoption of the approach by Kitchenham et al. (2008) in the online survey would not allow to get feedback on simulation-specific issues, as the checklists have no items for these simulation aspects. Therefore, we structured this survey (collaborative review) as a conference or journal review form. For that, we invited simulation experts, not only in the context of SE. The survey was released using the Web's *LimeSurvey* tool (www.limesurvey.org) and organized into five main parts:

- **Presentation**: Presents the study context, research goals, and instructions to join the survey, as well as the contact information.
- **Subject characterization**: Requires filling a four-question form regarding the subject's experience in both SE and simulation.
- **Guidelines' review**: Presents the link to download the technical report with the reporting guidelines, followed by a six-question form containing four closed questions on originality and novelty, technical soundness and contribution, presentation and readability, references to previous and related works, and two open questions on report strengths and weaknesses. This form is closely related to the correctness of the guidelines.
- **Feedback questions**: Six-question form on the need for formalized reporting guidelines for SBS, the possible recommendation of standard content-only or the including of an outline, future usage, adoption by publication venues in SE, and the possibility of either missing or extraneous information. For all questions, it is possible to comment the answers.
- **Acknowledgment**: message acknowledging the participation and bringing the study to its end.

The invitation or recruitment was done based on two approaches: by convenience and systematic. For the convenience approach, we used both the CNPq Lattes database (lattes.cnpq.br) and the ISERN (International Software Engineering Research Network) members list to look for both Brazilian and foreign researchers in SE with a background or experience in computer simulation. Then, we sent emails inviting 23 specialists to participate in the study.

In the systematic approach, we adopted the framework defined by de Mello et al. (2014). This framework consists of a systematic approach to define the adequate population and samples for SE surveys. In this case, we adopted the ResearchGate (www.researchgate.net) professional network as a source of sampling (SoS). This SoS has a meaningful constraint to send the invitations, allowing an account to send 20 invitations per day. We used three accounts to enable the execution and it took us 5 days to invite 300 members (assumed to be researchers). The criteria include researchers with a background on Software Engineering and simulation.

We ran one instance of the survey for each approach. It means the forms are the same. In both settings, the instances were open for 1 month long, as it included the full reading of the technical report (23 pages). After the deadline, we re-sent the invitations and extended the deadlines an extra month. The next section shows the results.

5.2.2 Results

After the deadline, we summarized the responses. During the extension period, we received two more answers from the sample by convenience and 32 more from sampling via the systematic approach, including incomplete participations. For the first sample (Lattes and ISERN), we got ten responses, but only two completed the survey. For the second sample (systematic sample), we got 54 responses, with 13 complete answers (Table 4).

From a quantitative perspective, the numbers are not promising. We selected the 15 completed responses for the analysis. The complexity of the task (read and review a 23-page technical report discussing guidelines for simulation in SE) may have influenced the low participation rate. Additionally, it is important to point that, except one subject, all others hold a PhD degree and reported experience on developing simulation models.

Both approaches returned responses with a similar quality, so we analyzed them as one single source. One of the subjects from the first sample also sent the reviewed technical report (pdf file) with comments by email. Table 5 has the number of answers from both samples: Lattes/ISERN and ResearchGate (RG). The consolidated results appear in the Total column. Questions 1 to 6 regard the reporting guidelines review and questions 7 to 12 regard the expert's opinion on the usefulness and application of the guidelines. Besides, only questions 5 and 6 are not mandatory. Thus, the total number of responses for mandatory questions should be 15.

In general, the results show this version of the reporting guidelines is comprehensive (questions 1, 2, 4, 5, and 12), understandable (question 3), useful (questions 7, 8, 9, and 10), but also has room for improvements (questions 6 and 11).

The contributions from the responses are important to reinforce the relevance of the proposed guidelines. In other words, subjects supported the proposed guidelines mentioning the need for more simulation studies in SE and for their systematization, as well as expressing agreement with the guidelines, even when having experience in other research areas (e.g., E-commerce, Physics). Specifically, they saw the guidelines as useful for young researchers, but not 'out of scope' for more experienced ones. Besides, subjects also point the aspects commonly missing in reported simulation studies, such as cost data, context information, underlying rationale for selecting simulation tools, and others. Additionally, the reporting guidelines bring principles from experimental statistics into the SE domain. Finally, the reporting guidelines' presentation was seen as concise, well-written and including well-chosen examples.

Subjects mentioned relevant improvements opportunities such as the need to emphasize the importance of the data used, as it is a critical step to ensure the 'health' of the data. In fact, the

Table 4 Response summary

Summary	Lattes and ISERN	ResearchGate
Invited	23	300
Total responses	10	54
Full responses	2	13
Incomplete responses	8	41

Table 5 Quantitative results for the review

#	Question	Lattes/ISERN	RG	Total
1	Originality and novelty			
	4: New and exciting idea	0	3	3
	3: Improves an existing idea in a significant way	2	9	11
	2: Nothing really novel	0	1	1
	1: Just rewrites or repeats known concepts or techniques.	0	0	0
2	Technical soundness and contribution			
	4: Excellent work and a major contribution	0	1	1
	3: Good solid work of some importance	2	10	12
	2: Marginal work but minor contribution	0	2	2
	1: Very questionable work and contribution	0	0	0
3	Presentation and readability			
	4: Very good	1	5	6
	3: Basically well written	1	8	9
	2: Readable	0	0	0
	1: Poor, needs considerable rework	0	0	0
4	References to previous and related works			
	4: Very good	2	4	6
	3: Good	0	6	6
	2: Average	0	2	2
	1: Poor	0	1	1
5	Strengths	2	13	15
6	Weakness	0	7	7
7	Need for formalized simulation reporting guidelines			
	Yes	2	9	11
	No	0	4	4
8	Standard content or standard outline			
	Only content	0	5	5
	Content and outline	2	8	10
9	Would you follow if they existed?			
	Yes	2	11	13
	No	0	2	2
10	Adoption of Empirical publication venues			
	Yes	1	10	11
	No	1	3	4
11	Missing information			
	Yes	0	4	4
	No	2	9	11
12	Extra (superfluous) information			
	Yes	0	0	0
	No	2	13	15

section "Supporting Data" refers to these issues, in a reporting perspective. According to Sargent (1999), data validity means the appropriateness, accuracy, availability of sufficient data, and all data transformations, such as data disaggregation, are correctly put. Unfortunately, there is not much that can be done to ensure that the data is correct. One should develop good procedures for (1) collecting and maintaining data, (2) testing the data collected using techniques such as internal consistency checks, and (3) screening the data for outliers and determining if the outliers are correct. Besides, model description is mainly influenced by the underlying simulation approach, which is already mentioned in the guidelines. One subject mentioned the existence of particular standards for reporting simulation models under specific approaches such as, for instance, System Dynamics (Sterman 2000) and Agent-Based Simulation (Grimm et al. 2010).

Additionally, one subject mentioned improvement in two major areas: validation and conclusions. For the validation part and later also discussing validity, he/she suggested the use of an underlying method. For instance, within the Air Traffic Management Community, the European Operational Concept and Validation Methodology (E-OCVM) as a departure point. The conclusion section is considered small, and terms such as risk and applicability offer room for multiple interpretation.

On the presentation side, they suggested a concrete table resuming the guidelines and more examples would help understanding, although they were concerned that this would unnecessarily increase the length of the reporting guidelines.

As negative aspects, some thought the list of references (bibliography) could be longer. However, the guidelines are not the review itself, but only one partial result. The references include all the outcomes from the systematic review and many additional sources outside Software Engineering.

Another aspect is that the reporting guidelines resemble a reformulation of ideas previously stated or perhaps that they do not really yet exist as established guidelines (or even standards) outside the realm of Software Engineering, and could have simply been re-used (after re-wording). In this sense, we are aware that the guidelines proposed share common concerns with other SE and simulation reporting guidelines (Section 3). These shared concerns were mainly added after the perspective-based reading (Section 5.1). Nevertheless, we understand the whole set of reporting guidelines as an original perspective, discussing simulation-related aspects and their issues faced in Software Engineering studies.

For the feedback questions regarding the use of the reporting guidelines, it is possible to see a positive direction in their usefulness, but with some limits. Regarding the adoption of the guidelines by researchers and reviewers, the subjects commented their use not as a standard, but as a recommendation or suggestion.

Finally, we could not see any theoretical or conceptual error, or even superfluous information. The possible lack of information mentioned earlier is mainly related to the importance of valid data. It reinforces the positive direction of the research and soundness of the guidelines proposed. However, one may still wonder if the contents of the reporting guidelines is obvious and, for that reason, the replies are dominantly positive. For that, we conducted an additional evaluation (Section 5.3), changing the perspective from the simulation experts to existing simulation study reports from the technical literature.

As this evaluation's result, the number of guidelines was added by two new items: applicability issues (SG21) and future research directions (SG22). These aspects have been included in the discussions, but we decided to highlight them into separate guidelines. Besides, we also included new examples and improved the discussions to clarify the concepts involved and the reasoning.

5.2.3 Threats to Validity

In this collaborative review, we worked intensively to present the tasks as a usual activity for researchers: the reviewing of conference or journal papers. This brings internal validity to the study as its instrument and concepts involved are familiar to the audience. Still on internal validity, we faced difficulties in recruiting subjects via the ResearchGate. The constraints imposed by the platform led us to accidentally recruit one unit from both samples. However, this unit answered in only one survey, not compromising the analysis.

From the perspective of conclusion validity, we obtained small sample sizes, having no room to apply statistical tests or determining confidence intervals. Even though, from the qualitative perspective, the comments and contributions are the valuable part of the feedback, pointing specific aspects that could be improved in the proposed guidelines. Furthermore, most comments show interest and expertise regarding the topic, which give us some confidence regarding subject's opinions.

As regards our constructs of interest, we captured correctness in items 2 (technical soundness), 3 (presentation and readability) and completeness in items 4 (previous and related work), 11, and 12 from Table 5. Items 5 (strengths) and 6 (weaknesses) contribute for both focuses. The remaining items were used to capture the perception of usefulness.

The subjects' characterization and comments showed different backgrounds. Some shared experiences of their work regarding simulation. Besides, we identified them realizing the application of the proposed guidelines to their research/engineering activities. It is important to embrace multiple perspectives of what we are taking as simulation studies and as SE issues. However, we have no ambition of assuming any generalization from these results, as it is based mainly on opinions and expected results and not on real application of the reporting guidelines. From this perspective, we have limitations on external validity.

As we are interested in more qualitative data, we also analyzed this study for descriptive validity, which regards the fact that the researchers are not making up or distorting the observed data, and interpretive validity, as to whether the inferences and conclusions follow the data, not being biased by the researchers during analysis. This way, we triangulate answers from different questions and comments, to ensure consistence amongst the items of the review. For instance, we cross-checked whether the answers for presentation and readability matched the comments on paper strengths and weakness. We also compared these last aspects to comments regarding missing or extra information.

5.3 Analysis Against the Technical Literature

The results after proposing and evaluating the reporting guidelines (as presented in the previous section) motivated an additional evaluation. It aims at understanding whether the reporting guidelines entail common aspects, often reported in current technical literature. For that, we updated the qSLR protocol (de França and Travassos 2013a, b) and analysed the outcomes against the guidelines proposed.

For this trial, we evolved the research protocol to concentrate on the use of simulation model or experimentation, rather than just the model development. So, we added a new

Table 6 Results from the updated review

Digital library	Number of records	Duplicated entries	Included
Scopus	261	1	10
Web of science	19	2	4
IEEE Xplore	172	59	6

inclusion criterion to the research protocol where every paper should contain at least one simulation experiment, excluding papers that have only proposal for a simulation model. Besides, we excluded the EI Compendex database from the sources of sampling as we could not apply the same search string used before, as it displayed unexpected behaviours and faults. Therefore, we included the IEEE Xplore digital library.

Due to these changes the time frame used to apply the search strings was different between libraries. From the Scopus and Web of Science ones we got papers from March, 2011 (year for the first trial of the review) to the date of the update trial (November, 2013). And from the IEEE Xplore we got papers until November, 2013, as we did not apply the search string in this library in the first trial.

The results from applying the strings to the search engines and after the selection procedure (based on the reading of titles and abstracts) are shown in Table 6.

The 20 papers included were read in full and we excluded most of them as we understood there are no studies in those papers, but just model proposals and examples of use. Thus, we remained with four papers (Andersson et al. 2002; Psaroudakis and Eberhardt 2011; Zhang et al. 2012; Uzzafer 2013). Additionally, we searched for simulation studies in the main conference (ICSSP[3]) and journal (SPIP,[4] including its new title JSEP), applying the same criteria from the qSLR. We found seven other simulation studies (Al-Emran et al. 2010; Birkhölzer et al 2010; Houston and Lieu 2010; Bai et al. 2012; Paikari et al 2012; Concas et al. 2013; Houston and Buettner 2013) in these venues.

These eleven research papers were read in full and we later analysed their reports based on the proposed reporting guidelines. For each guideline we assigned a three-value scale: Not Complied (0), Partially Complied (1) and Complied (2). The analysis consisted of searching for information in the reports that could satisfy each guideline.

It is important to note that we excluded reporting guidelines SG6, SG11, and SG15 from this analysis. Although we understand that these guidelines are strongly related to simulation studies, they are not entirely applicable to the studies selected for this analysis.

The SG6 relates to the establishment of hypotheses. So we understand that for characterization studies this may not be essential or necessary. Besides, SG11 relates to subjects description and it is not always applied to in silico studies. Finally, the SG15 focuses on intermediate trials and this is not common for the simulation approaches adopted in these studies, especially System Dynamics. It is often applied to stochastic simulation, where many replications are executed for the same inputs. Apart from that, when the simulation environment (or simulator) does not offer this kind of support, this is usually a concern.

The overall coverage for the reports studied as related to the reporting guidelines is shown in Fig. 3.

[3] http://www.icsp-conferences.org/

[4] http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1099-1670

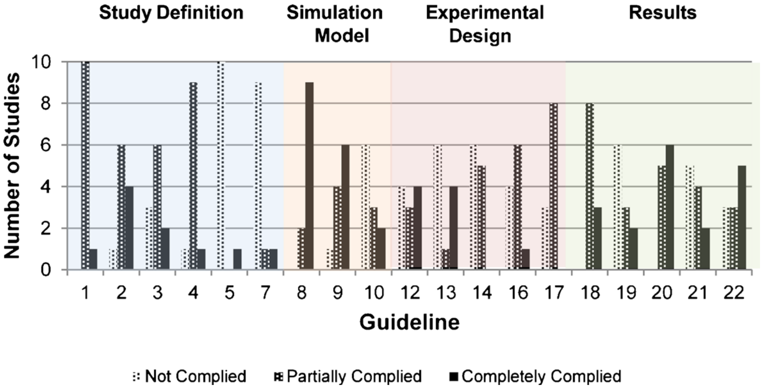


Fig. 3 Guideline coverage for different research papers

We roughly divided the set of guidelines into four sections for analysis purposes, in which the first one regards the initial planning, i.e., issues related to context, problem, goals, and research question definition. The second section regards the simulation model description, as well as its foundations and validity evidence. For the third section we focus on the experimental design issues, such as the definition of the variables of interest, the causal model, including the design matrix and simulation scenarios, as well as the number of simulation runs. The fourth and last part relates to the analysis of results and conclusions.

From Fig. 3 it is possible to see that all the guidelines seem to be reasonable, as all of them are completely reported at least once (in one study), except for reporting guidelines SG14 and SG17, which are respectively related to the number of runs and simulation environment. In the case of SG14 we could not identify the reasoning to determine the number of sufficient simulation runs. This aspect is important as establishing a loose number of runs may affect the output analysis effort and also the quantifying of the variance amount. It is also an issue not to have the complete supporting environment in terms of replication. Reports usually mention only the adopted simulation tool rather than the input/output analysis tools, preparations for calibration, runtime environment, and additional supporting technologies.

From an individual perspective (Figs. 4 and 5), no report mentioned the whole set of aspects covered by the reporting guidelines. Also, there is no report presenting a homogeneous distribution of the relevant information to be reported. It means that every report focuses on one or two specific groups of aspects when reporting on the simulation study.

It is possible to see a large variance on which kind of information the reports concentrate on, probably due to the lack of a standard or recognized methodology guidelines. For instance, in Fig. 5, the report made by (Houston and Buettner 2013) has a comprehensive description of the simulation model (SG8, SG9 and SG10) and also a good amount of information regarding the experimental design (from SG12 to SG17). The authors report what seems to be a case study supported by a discrete-event simulation model to investigate sources of variation in deliveries and how to improve delivery quality of an agile software project, where both the customer and the contractor had become concerned with the lack of predictability in contractor deliveries, in the Aerospace Corporation.

One could suppose that this report followed an adequate simulation methodology, considering the number of guidelines that could be (at least partially) applied and the results obtained,

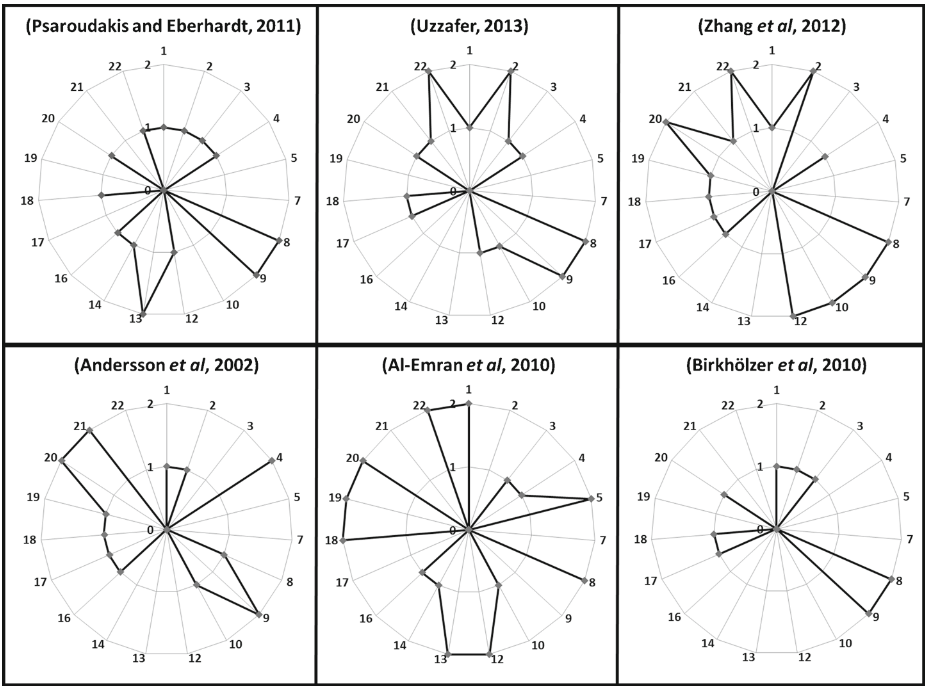


Fig. 4 Individual profile according to reporting guidelines

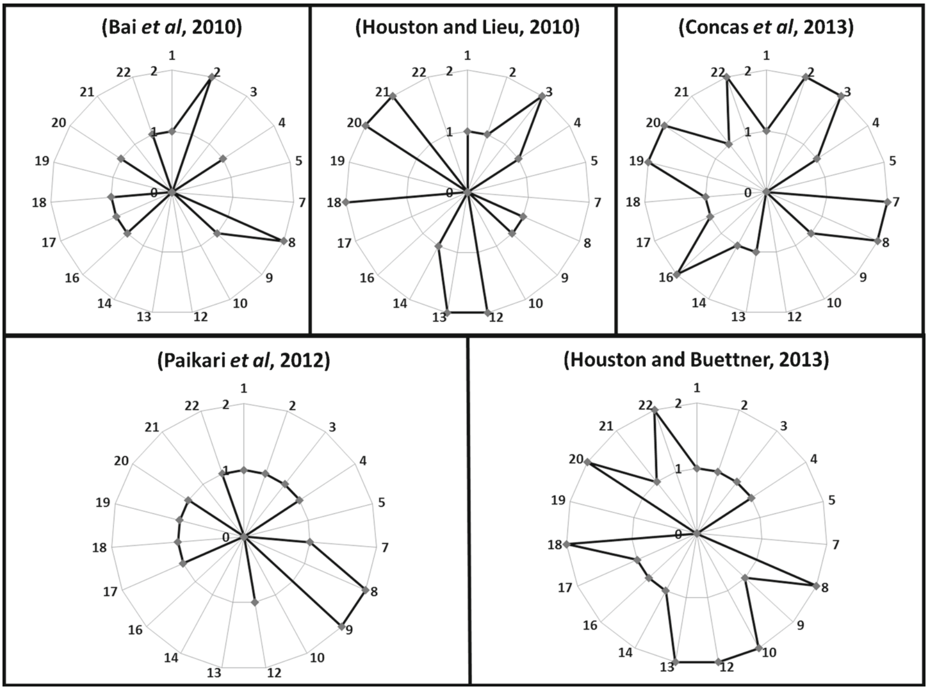


Fig. 5 Individual profile according to reporting guidelines

which are discussed in terms of applicability in practice. However, there are other reports, such as (Concas et al. 2013), where there is not much information on the simulation model and its validity, as well as the experimental design, but it also presents a good analysis of the results obtained. Indeed, for both reports, there is no explicit mention to the whole methodology adopted. So, it is not fair to judge the quality of a study based on the quality of the report.

As general behaviour, we can point at the model development vs. model experimentation perspectives. Usually, when a report focuses on the simulation model description, we can see a lack of information regarding the simulation experiment itself. Of course, this is a relevant aspect when considering the number of pages available, mainly in conference papers. However, the model description alone cannot show real contributions when it is not associated to a simulation experiment indicating its validity and usefulness. The opposite is also possible, i.e., when the experimental design is emphasized, with a resulting lack of data on model description.

The simulation model description should entail at least the conceptual model, the main factors, and response variables, as well as its equations. And it regards the full understanding of the study being conducted. However, in five of the 11 reports we could not identify one of these aspects. This not only reduces the understandability, but also compromises the possibility of replicating the study.

As regards model validity, V&V procedures are often just mentioned, when done; there should be evidence of the results that such procedure was executed, for example, the improvements regarding the issues found, the level of confidence in case of accuracy tests, the assumptions that could be verified and those that could not, amongst others. From the eleven papers analysed, six do not provide any information on model validity. Such lack of validity information compromises the credibility of the study and also the confidence on the results.

Considering the initial plan, some issues can be seen, such as the lack of precision in communicating the problem under investigation. The motivations and what is to be solved by simulation are not described in sufficient detail or not described at all. It resembles works where there is a solution and someone else is trying to find a problem that fits it. Goals and research questions seem to be used in an interchangeable way, e.g., the reports (except for one paper) present only the research goals, without presenting the research questions associated to the goals. It is possible to identify the general research goal and try to infer the research questions, but this is not clear anyway. Apart from that, the justifications for using simulation as an investigation approach are often neglected too, but in some of the reports it is possible to argue against the proper use of simulation for specific problems or goals. For specific purposes, other analytic methods could be applied. We generally argue that, without the data provided by these research questions, it is not possible to question the feasibility of using simulation as an adequate strategy.

Experimental design issues often involve scenarios, design matrices and the number of simulation runs. Simulation scenarios are mainly elicited ad-hoc, when the experimenters are not using DOE to plan the experiment. Also, it affects the determination of the number of simulation runs needed to perform the study. As the scenarios are not systematically identified, the number of runs tends to be lower due to the bias in scenario selection.

As in other research strategies, the reporting of raw data is also an issue due to disclosure agreements. The supporting data is another relevant criterion for the credibility and validity of simulation studies. The way in which this data is used to develop and validate the simulation model and how it looks like was not presented at all. Thus, without details on model calibration and applied statistics, it is not feasible to make any judgment as to whether it is adequate or not. However, there are several ways for reporting it, for instance, using a multiplier factor to mask the data. Apart from that, contextual information can be given without naming organizations and people.

Another aspect presented in a general idea is the simulation environment. We often identify indications of the simulation tool used and rarely a statistical package for data handling. However, it is not possible to repeat studies without further information.

We could also identify some issues regarding the simulation results. Simple comparisons of output variables are a common procedure, but experimenters have to be aware that these are not enough. No determination of effect is provided for the input factors involved in the experimental design, when related to the output variables. Apart from the report on the outcomes, often plotted in charts or tables, there are several missing discussions, such as threats to validity, conclusions, and the applicability of the results in the real world.

Threats to the validity of simulation studies are seldom discussed based on types of experimental validity. Authors usually refer to them as limitations and unverified assumptions, without discussing their consequences. Al-Emran et al. (2010) and Concas et al. (2013) present discussions according to the types of threats proposed in (Cook and Campbell 1979).

Finally, from a contribution point-of-view, the results are interesting, but the discussions are limited in explaining why these results occur and how they can be applied in practice. The explanation should be an answer for the research questions and be grounded on the experimental design and model description. For instance, the conclusions should state how input factors (and their interactions) affect the output variables, exposing the theoretical logic embedded in the simulation model through a chain of variables or events. Furthermore, this explanation should be reasonable as the attempts to validate the simulation model succeed and accrete confidence to the results. Apart from that, the conclusions seems to be based on a one-scenario design, without evaluating other possible interactions amongst input factors.

6 Planning Guidelines

In this section, we concentrate on the use of simulation models, also called model experimentation (Balci 1990), rather than the whole set of activities for model development. Their scope ranges from research goals to output analysis, assuming the model development had been done. Thus, the guidelines focus on the simulation model as an instrument for experimentation, not considering its development, except for those common features that include both development and application.

The lack of a planning perspective for simulation experiments can be compared to the perspective of performing extensive statistical tests during output analysis, causing methodological problems such as the desired results occurring by chance (Kitchenham et al. 2008). The comments by Davis et al. (2007) illustrate it better when stating that, without an intriguing question, simulation research relies on '*a fishing expedition, in which the researcher lacks focus and theoretical relevance and risks becoming overwhelmed by computational complexity*'. So, the lack of planning and concrete research questions can add bias to the results of the studies.

We concentrate on this topic due to the possibility of observing a lack of rigour in conducting such experiments (de França and Travassos 2013a, b) and potential threats to validity identified in simulation studies in SE (de França and Travassos 2014b).

In order to streamline the reporting guidelines on a planning perspective, we needed to accomplish one major goal that regards the possibility of anticipating potential threats to the validity of the simulation experiment. For that, we performed the procedure shown in Fig. 6.

First we identified these validity threats by applying a coding technique, namely Constant Comparison Method (Corbin and Strauss 2008), to the dataset produced by the outcomes of the qSLR blended with ad-hoc surveyed information regarding other science areas. We then analysed data extracted from 15 technical papers, consisting of simulation experiments from the qRSL reporting threats to validity and limitations, which allowed us to identify and classify 28 different threats to validity as related to SBS in SE, according to Cook and Campbell's categories (conclusion, internal, construct, and external validity) (de França and Travassos 2014b).

As the validity threats classified concern both model validity and experimental design, we performed analyses using the set of V&V procedures we acquired from the qSLR and knowledge made available on Design of Experiments (DOE), classic (Montgomery 2008) and simulation (Kleijnen et al. 2005). It allowed the matching of procedures and techniques supporting the mitigation of these validity threats and suggesting how they can be applied as planning guidelines (presented in the following subsections) in the context of SE. The planning guidelines linkages with the threats to validity are presented in Table 7. In this table we selected validity threats that emerge solely in SBS, as in virtuo studies may also involve common threats to in vitro ones, such as maturation and instrumentation effects (de França and Travassos 2014b).

Except for guidelines SG1, SG8, SG20, SG21, and SG22, which are exclusively focused on reporting aspects, all the others shown in Section 4 discuss both reporting and planning perspectives. This way, model experimentation should also involve aspects such as research context (SG2), problem formulation (SG3), goals (SG4), research questions (SG5), and hypothesis definition (SG6), which are clearly part of a study protocol, including simulation studies. The same can be said for the feasibility analysis of a simulation study (SG7), as well as model description (SG9), reflecting the full grasp of the observation instrument. The remaining

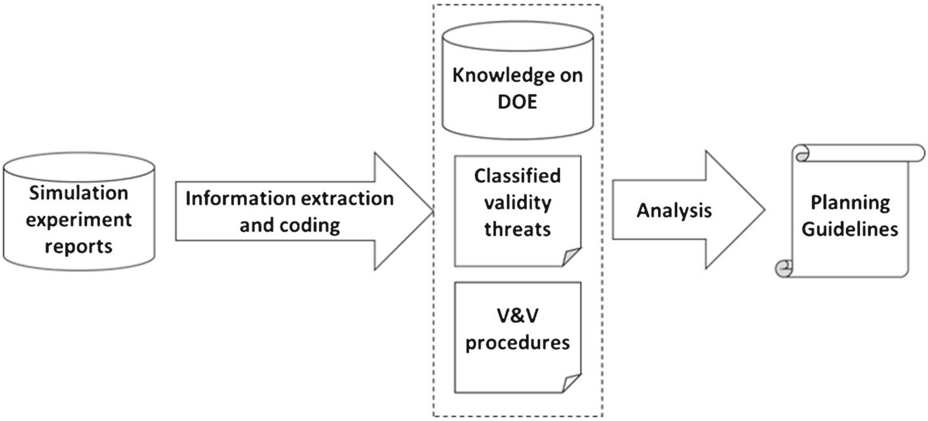

Simulation experiment reports
Information extraction and coding
Knowledge on DOE
Classified validity threats
V&V procedures
Analysis
Planning Guidelines

Fig. 6 General procedure to achieve the planning guidelines

Table 7 Association between planning guidelines and threats to validity

Guideline	Threats to validity
SG7	- Inappropriate application of simulation
SG23	- Simulation model simplifications (assumptions) forcing the desired outcomes
	- The simulation model does not capture the corresponding real-world building blocks and elements
	- Inappropriate cause-effect relationships definition
SG24	- Inappropriate cause-effect relationships definition
	- Simulation model not based on empirical evidence
SG25	- Hidden underlying model assumptions
	- Invalid assumptions regarding the model concepts
SG26	- Inappropriate real-world representation by model parameters
SG27	- Simulation results are context-dependent, since there is a need for calibration
	- Simulation may not be generalized to other same phenomena simulations
	- Inappropriate experimental design (missing factors)
SG28	- Considering only one observation when dealing with stochastic simulation, rather than central tendency and dispersion measures
SG30	- Different datasets (context) for model calibration and experimentation
SG31	- Not using statistics when comparing simulated to empirical distributions
SG33	- Inappropriate model calibration data and procedure
	- Simulation results differ from the outcomes of empirical observations
Other issues	- Naturally different treatments (unfair) comparison

aspects covered in the reporting guidelines will also be further discussed in this section, but focusing on the definition of valid scenarios and experimental design, aiming at avoiding potential threats to validity.

6.1 Model Validity

It is important to gather evidence on the validity of the simulation model. There are V&V procedures available to support it, but none of them can avoid all the potential threats to that validity alone. However, successfully applying some of these procedures together can help increase the confidence on the simulation results.

> SG23. Make use of Face Validity procedure (involving domain experts) to assess the plausibility of both conceptual and executable models and simulation outcomes, using proper diagrams and statistical charts as instruments respectively.

A common V&V procedure is Face Validity, which is a white box approach for reviewing both the simulation model and I/O matching. It enables the investigation of internal properties and behaviours of a simulation model, rather than dealing with it as a black box. This way, threats to construct validity, involving the mechanisms that explain the phenomenon captured by the simulation model, may be identified in advance by domain experts. For instance, experts can find both inappropriate definitions of cause-effect relationships and failures in capturing the corresponding real world building blocks and elements.

> SG24. Support model (causal) relationships, as much as possible, with empirical evidence to reinforce their validity and draw more reliable conclusions.

From an external perspective, it is sound to have the simulation model's causal relationships supported by empirical evidence. It also brings more external validity (Davis et al. 2007). This way, secondary studies can be done to search for evidence, if that is not known. At least for the core causal relationships, as bigger models may require greater effort.

> SG25. Always verify model assumptions, so that the results of simulated experiments can become more reliable.

Face validity can also be blended with Rationalism (Sargent 1999) to assess a model's assumptions on the underlying concepts. This last procedure uses logic deductions from model assumptions to develop the correct model, assuming the experts know whether the underlying assumptions stated are reasonable. However, when model assumptions are hidden or not clearly stated, no procedure can be applied. In these cases, procedures such as comparisons to reference behaviours and testing structure and model behaviour are more suitable. The expected behaviours can give insights on how hidden model assumptions are affecting their results.

The verification of model assumption also applies to simplifications that imposing an expected behaviour (Eck and Liu 2008). When the simulation model is to be developed, the modeller assumes, even implicitly, some issues regarding the phenomenon. For instance, the increase in a response variable directly caused by the presence of a given treatment. If these assumptions are embedded in the model, this may pose a threat to internal validity as they should not be directly coded in the model but be an effect of a chain of actions, events and conditions that generates such behaviour, affecting the response variable.

6.2 Experimental Design

In order to properly design the simulation experiment one needs a deep understanding of the simulation model at hand (see Section 4.2). As seen in (Balci 1990), different values and types of system parameters, input variables, and behavioural relationships, and auxiliary and response variables may represent system variants, as they constitute statistical design factors. This way, variables of interest should be embedded in the model. In any case, research goals and questions should drive the causal model definition for the experimental design, by taking the part of the model that reflects the concerns in the goal and the variables that can help in answering the research questions.

> SG26. Use results from Sensitivity Analysis to select valid parameter settings when running simulation experiments, rather than model 'fishing'.

Techniques such as Sensitivity Analysis are useful when selecting the groups of interest factors and level range. Once the more sensitive factors are determined, the number of levels for each factor and also the values they assume can be properly defined. Furthermore, a systematic way of defining the levels reduces the bias and avoids fishing for positive results.

> SG27. Consider using as factors (and levels), apart from the simulation model's input parameters, when designing the simulation experiment, as well as internal parameters, different sample datasets, and simulation model versions, implementing alternative strategies to be evaluated.

To identify design factors and levels, consider variables beyond the model's input parameters. Other possibilities include distinct datasets as factors, with the simulation model remaining constant. This way, different calibrations representing different simulation scenarios may be compared.

Each combination of all factor values is called a scenario. The design matrix is a table containing all the scenarios for the simulation experiment, which can fully describe the experimental design. However, there are several different designs that can be generated for the same set of factors. In Statistics, there is a mature discipline named DOE. We have no ambition to contribute with it, but to bring such knowledge and apply it to SE simulation experiments, considering the specific context as an immature field – lack of solid knowledge, unknown disturbing factors, hard-to-control environments, and so on. The application of DOE to simulation is not a new subject, even in SE. However, it is an interesting technique since, for real systems, which DOE was proposed to, it may be impractical or unfeasible to experiment with many factors and levels (more than 10 factors and 5 levels), and the same cannot be said for simulation experiments. Apart from that, Kleijnen et al. claim that DOE for simulation experiments is different as in simulation one is not limited by real-world constraints (Kleijnen et al. 2005).

Factorial designs are the most famous ones. They can be simply defined as a set of scenarios including all possible combinations for a set of factors, also called Full Factorial Designs. For instance, a full factorial design for k factors using two levels per factor is denoted as a2^k design, meaning the number of scenarios needed to determine effects from k factors and their interactions.

There are also variants proposed for a large number of scenarios and the simulation runs are time-consuming, as they grow exponentially with the number of factors and levels. So, it is possible to reduce the number of scenarios and still have an efficient estimator. In these cases just a fraction of the scenarios is executed and for this reason they are called Fractional Factorial Designs. Fractional designs can be defined as 2^{k-p}, where p is a value called power of the fraction, in which 2^{k-p} is greater than k. The value of p is also determined, considering the possibility of investigating interactions between factors and higher-order effects.

Some aspects are important to select an adequate design. Here, we give some of them, but not in an exhaustive list:

- Simulation goal, as designs for understanding are not the same for comparisons or optimizations;
- Experimental frame, of whether the area of interest is local or global, and it impacts on the range of levels;
- Number of factors and levels, as they exponentially increase the number of scenarios in full factorial designs;

- Domain of admissible scenarios, important as full factorial designs may generate inadmissible scenarios;
- Simulation model's deterministic and stochastic components, as they affect how to deal with variation in the experimental design. Stochastic simulations use pseudo-random numbers, which imply that each single replicate output is a time series with auto-correlated observations. So, the values of such observations cannot be aggregated;
- Terminating conditions, if it is steady state or a terminating simulation, with an event to specify the end of the experiment.

> SG28. When dealing with simulation models containing stochastic components, determine the number of runs needed for each scenario, to capture phenomenon variance.

Simulation models containing stochastic components naturally produce intrinsic noise in the output, due to the pseudo-random number generator. Thus, one single run of each scenario using those stochastic components cannot reveal the amount of variance in this noise. At the other end, the higher the number of runs, the greater the approximation of a desired accuracy level. So, given an accuracy level and an initial estimate from few model runs, it is possible to determine the number of runs required and avoid this threat to conclusion validity.

Another common approach to deal with experimental design issues is to investigate specific real scenarios. In this case, the experimenter needs to be aware of the relevance and adequacy of each selected scenario. The main drawback of this approach is to achieve, in an ad-hoc manner, the experimental design, potentially embedding some bias and with no opportunity to investigate side effects such as interactions between design factors. There are other types of design often applied to simulation experiments that provide successful results (Kleijnen et al. 2005) such as Central Composite Designs, Sequential Bifurcation, and Latin Hypercube Sampling. For the sake of space we are not going to discuss them here.

6.3 Planning Data Collection and Use

Apart from the aspects discussed in Section 4.3, it is important to mention the need for planning data collection, to also avoid measurement mistakes. So, promoting the collection of data as soon as it is made available for the target model variables and also to capture the contextual information associated to the quantitative data is relevant for simulation-based studies. After the collection, quality assurance procedures should be run to check their quality.

> SG29. Keep track of qualitative data along with quantitative data. It is also important to record data contextual information.

Contextual data is important to provide better and more accurate reasoning when doing output analysis and interpretation. On example can be seen in Section 7.8 where we support the explanations based on contextual data, as collected from one project team member.

> SG30. Make sure that both calibration and experiment datasets came from the same population.

The data used to calibrate the simulation model and to set model parameters in the experiment needs to share the same context in the sense that they are comparable. The values used for model experimentation have to be consistent, avoiding attempts to inappropriately generalize behaviours for different contexts. The use of cross-company data is an example of how it can impose a threat to internal validity on the simulation results.

6.4 Output Analysis

As simulation runs generate a considerable amount of data and also involve complex relationships amongst variables, it is possible to identify, prior to the execution, what the statistical instruments are to support the output analysis.

> SG31. Make use of proper statistical tests and charts to capture outcomes from several runs and to quantify the amount of internal variation embedded in the (stochastic) simulation model, increasing the precision of results.

There is a need for some common analysis such as main and interaction effects amongst factors, simulation confidence and accuracy, quantifications of variance and also comparisons with reference behaviours or alternative system configurations. For that, statistical charts and tests, along with descriptive statistics can help, but for every instrument there are assumptions that have to be assessed in the output data, such as normally distributed data, independent samples, homogeneous variance, and so on.

It is also important to take care of the perspective of the analysis, whether across different simulation runs (or replications) or within a single replication. Simulations from different replications are usually independent from each other, so it is possible to use means, standard deviations, and confidence intervals of measures from variables across replications, but not within a replication, i.e., calculating these values for variables using measures from different time steps.

6.5 Threats to Validity

For the discussion of threats to validity, we categorize them as done by Wöhlin et al (2012): conclusion validity, internal validity, construct validity, and external validity. Several potential threats that should be checked before the simulation experiment can be found in (de França and Travassos 2014b).

> SG32. Consider checking for threats to the simulation study validity before running the experiment and analysing output data to avoid bias.

Threats to conclusion validity involve the use of inappropriate instruments and assumptions to perform simulation output analysis, such as not using statistics (statistical tests or metrics) when comparing simulated distributions to empirical ones, considering only one observation when dealing with stochastic simulation, rather than central tendency and dispersion measures, independence between factors, amongst others.

As the experimental setting in SBS often relies on different input parameters configurations, the uncontrolled factors may be unreliably supporting data or distinct datasets (context) for model calibration and experimentation, human subjects manipulating the model when conducting in virtuo experiments, or bias introduction by the simulation model itself, when its assumptions force the desired outcomes. All of these affect internal validity.

Davis et al. (2007) claim that the nature of simulation models tends to improve construct validity, as it requires formally defined constructs (measurement) and algorithmic representation logic for the theoretical mechanism, which explains the phenomenon under investigation. However, it is possible to identify threats to construct validity in the context of SBS, such as: inappropriate cause-effect relationship definition, real-world representation by model parameters and model calibration data and procedure; hidden or invalid underlying model assumptions regarding the model concepts; and the simulation model not capturing the corresponding real world building blocks and elements.

> SG33. Be aware of data validity when comparing actual and simulated results: compared data should come from the same or similar measurement contexts.

In simulation studies, it is particularly interesting to know whether the results can be also observed in different simulation studies of the same phenomena [simulated external validity (Eck and Liu 2008)] or if it can predict real-world results [empirical external validity (Eck and Liu 2008)]. Threats to external validity can also appear as context-dependent results, as there is a need for calibration and simulation model not based on empirical evidence.

7 Planning Guidelines Application

To evaluate the guidelines proposed through a proof of concept, we planned and executed a simulation experiment focused on software evolution. In this section, the study plan and its results are described, besides the tracing between the plan's part and correspondent guideline, indicated through (SGxx) marks. We opted for this approach rather than an exhaustive discussion of guideline application.

The study motivation (SG2) converges on two aspects: (1) an initial feasibility assessment of the proposed planning guidelines for simulation experiments; and (2) the understanding of how a project manager can breakdown long-term releases of a large scale information system to control business processes in a research supporting organization. The project team is geographically distributed in two sites, following an iterative and incremental software development process, emphasizing V&V activities. This simulation experiment also intends to show how a software evolution simulation model (Araújo et al. 2012) can be used to support the answering of research questions regarding software maintenance.

Thus, the problem investigated (SG3) regards the software lifecycle at the time the information system changes from a development to a maintenance (corrective, evolutionary, or perfective) stage. Usually, maintenance cycles depend on a set of improvement requests from project stakeholders, which clearly identifies this moment (Kitchenham et al. 1999). This way, the project manager should be able to plan product releases observing the restrictions regarding product quality, time to market, and budget. However, these variables can depend on unpredictable or unknown factors, which can produce a sub/super estimated time for the maintenance plan. Thus, the project may go over schedule, needing actions such as increasing the number of human resources, with higher costs and possibly decay in product quality.

7.1 Goal and Research Questions

The goal (SG4) of this study based on GQM is:

> **To analyse** the evolution of an information system **for the purpose of** characterization as regards the duration of maintenance cycles as well as its effect on product quality **from the point-of-view of** the SE Researcher **in the context of** simulating quality decay for a large-scale information system with the use of a SD model as instrument.

For the goal defined, we derived two research questions (SG5):

Q_1: Which periodicity (shorter or longer cycles) performs better for the next 6 months after the last release?
Q_2: Which strategy (fixed or variable duration cycles) performs better regarding product quality?

7.2 Simulation Feasibility

Although we have presented some motivations to do this study, the use of simulation in this context can be justified (SG7) by the long-term analysis in which several variables of interest need to be timely controlled without imposing risks to the software project. Furthermore, we are interested in observing how these variables behave over time, and in their interactions considering not only first-order (i.e., effects of Periodicity on both Size and Complexity), but also higher-order effects (i.e., successive relationships and/or causal loops such as a loop involving Effort, Maintainability, and Reliability).

7.3 Simulation Model

Araújo et al. (2012) present (SG9) an infrastructure based on the Laws of Software Evolution to observe software quality decay throughout software development and maintenance

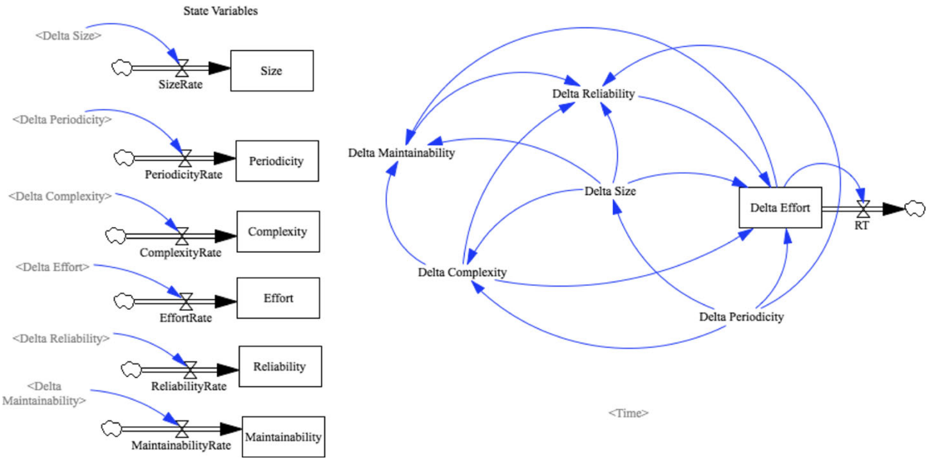


Fig. 7 Software Evolution Model adapted from (Araújo et al. 2012)

processes. The main idea is to get a better understanding of how the software system may be affected by several changes occurring in its lifecycle. In order to support the evolving systems' behaviour observation, an evidence-based logical model was defined and described through SD constructs to allow the simulation of successive maintenance cycles. The SD model for software evolution is shown in Fig. 7.

The model was developed over six state variables that represent the combined status for both project and product: **Periodicity**, the time interval between each release version of a produced artefact (e.g., software or documentation versions); **Size**, the magnitude of artefacts produced in each lifecycle stage of the proposed software (e.g., the amount of lines of code in the source code or the number of requirements in the requirements specification document); **Complexity**, the elements that can measure the structural complexity of an artefact (e.g., cyclomatic complexity of methods, or number of classes in the class diagram); **Effort**, the amount of work done to produce a version of some artefact (e.g., measured in terms of man-hours or equivalent unit); **Reliability**, the number of defects corrected per artefact in each software version; and **Maintainability**, the time spent in fixing defects. The only difference from the original model is that Periodicity is not determined by the simulation cycle, as it is a design factor.

In order to improve model validity, the authors collected evidence for each relationship amongst model variables from the technical literature (SG24). For the complete set of evidences, see (Araújo et al. 2012). Apart from that, the model was successfully evaluated using the procedure of Historical Validation (SG10) in which a dataset is divided into two pieces and the model is calibrated using the first eleven releases and then simulations are run to verify if the model can predict trends for each model variable according to the second part of the dataset (later eight releases). So, the model was able to predict the trends for the output variables. This is considered enough for the purposes of understanding of our study. The simulations are executed in the Vensim environment (SG17), which supports the simulation of SD models and has an academic version (PLE) with limited support for experimentation, but free of charge. Additionally, it offers interesting analysis tools, such as causal tree, output plotting on sequence charts and simulation traces.

7.4 Subjects

This is an in silico experiment (SG11). Therefore, subjects' characteristics are not taken into account and not explicitly represented in the simulation model. This way, the effects of the subjects from the real project are abstracted through the supporting data used to calibrate the model, which contemplate characteristics such as productivity and team expertise.

7.5 Experimental Design

The variables (SG12) of interest are Periodicity, as independent variable, and product quality in terms of Reliability and Maintainability, as dependent or response variables. For the periodicity factor, we will adopt low, medium, and high values, to understand how the response variables behave by increasing the periodicity. The level differences are meant to understand the effect of both small and large changes on the input parameter, i.e., whether factor sensibility is introducing bias. Additionally, we have a qualitative factor with two levels, from our research

Table 8 Design matrix for the simulation experiment

Scenario	Strategy	Periodicity
1	Fixed-duration	2
2	Fixed-duration	10
3	Fixed-duration	40
4	Variable-duration	Low mean (2) and variation (1)
5	Variable-duration	Medium mean (10) and variation (5)
6	Variable-duration	High mean (20) and variation (10)

question Q_1, regarding the strategy for the organization of maintenance cycles: fixed-duration or variable-duration cycles. Fixed-duration means that every cycle has the same periodicity. On the other hand, variable-duration means that each cycle may have a different periodicity.

In the causal diagram on the right side of Fig. 7 it is possible to see that there are first-order and other higher-order possible effects of Periodicity on both Reliability and Maintainability. So, in this experiment we will explore full factorial designs for questions Q1 and Q2 as shown in the design matrix (Table 8).

For the scenarios (SG13) regarding fixed-duration strategies, the model behaves deterministically, and therefore we need just 3 runs (SG14), one for each periodicity level. On the other hand, the experimental design involves the use of a stochastic variable for periodicity, using the strategy of variable-duration. This variable is assigned to a normal distribution, with different mean and variance for each scenario. The choice for a normal distribution was based on the *Kolmogorov-Smirnov* test, done on the collected data that presents a normal distribution for periodicity. In these scenarios (SG28), we use 100 runs for each one of the 3 scenarios, in a total of 300 runs for the variable-duration scenarios.

For each simulation scenario we defined an output dataset, resulting in six datasets. These simulation runs were executed in the Vensim PLE environment, by explicitly setting the input parameters for each scenario.

7.6 Supporting Data

The data and procedure used for model calibration came from (Araújo et al. 2012). It was collected from a large-scale software project, in which the software under development is a Web-based information system for automation of business processes of an organization responsible for supporting both in financial and administrative aspects of research projects. For this project, the development team adopted an iterative and incremental software development lifecycle, with strong emphasis on verification, validation, and testing techniques throughout the software development. Apart from that, geographically distributed teams took part in the development, using Java and Java Server Faces platforms. The development team is stable, with about 12 developers.

To support observation, 13 different system releases were considered (SG16). This historical dataset was available in version control system logs, bug tracking services, and effort registration spreadsheets, whose measurements are relevant to the observation of system evolution, and for each release it collected measures for the six variables mentioned in section 7.3.

The system releases came from corrective, adaptive and perfective maintenance activities (SG29). The perfective maintenance mainly regards, in this dataset, the enhancements regarding security, performance, maintainability, and graphical user interface. No new functionality was considered during these releases. So, our simulation results are limited to these types of maintenance. Additionally, users reported the corrected defects for each release, during the system's operating lifecycle.

7.7 Output Analysis

For output analysis (SG31), statistical charts are used, namely histograms and sequence run charts, to characterize response variable behaviour. Histograms are needed to check their distribution, while the sequence run is useful to understand how the values for these variables behave over time. Additionally, we use the sequence run to compare different scenarios by plotting their series on the same chart. For instance, to analyse the *Strategy* factor corresponding to research question Q_1, scenarios 1, 2 and 3 are compared against scenarios 4, 5 and 6, respectively. These comparisons keep the *Effort* factor constant on the base value, as it is not a variable of interest for this research question. Similar analyses are performed for the other factors or interactions w.r.t. each research question. Question Q_2 involves the use of a random variable, requiring the analysis of several runs. It implies the use of statistical measures of central tendency and dispersion when comparing the scenarios.

7.8 Simulation Results

As mentioned in Section 7.5, a total 303 simulation runs were needed to evaluate all the planned scenarios. After running these simulations, we could observe exclusively the context of this project dataset, that (SG20):

Shorter maintenance cycles lead to greater reliability As Fig. 8 shows, the shorter periodicity scenario (1) has a higher number of corrected defects over 6 months. This result is explained by two main reasons: shorter cycles are mostly related to corrective and adaptive

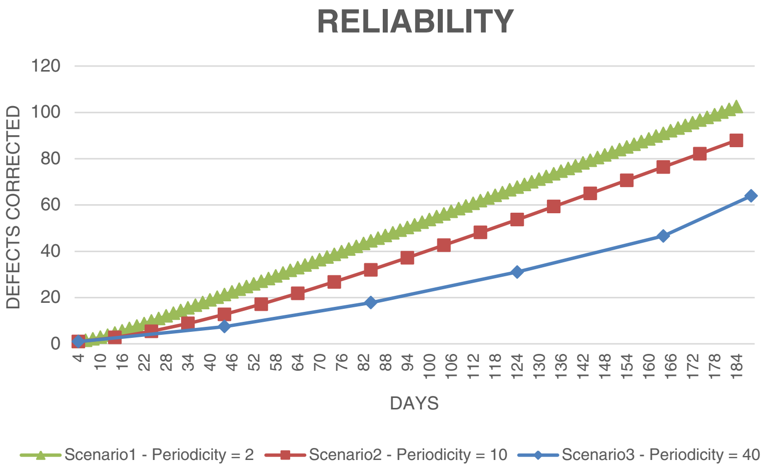


Fig. 8 Reliability output for fixed-durations

maintenance, and as there was no new functionality added, these maintenance cycles are always meant to correct defects, which is likely to improve system reliability. Moreover, shorter cycles are associated to critical defects. As the system was operational when the defects were reported, the most critical ones received the highest priority to be fixed, aiming at quickly delivering the releases that contained critical corrections.

Fixed-duration maintenance cycles are more reliable for shorter and medium cycles Based on previous results, the use of variable-duration cycles with a short mean and variance in their periodicity approximates the maintenance cycles from fixed-duration shorter ones, which we saw promotes more corrections. On the other hand, when adopting a high mean and variance for the periodicity, the variable-duration strategy does better than fixed, long cycles. It happens as it can also accommodate short cycles within the longer ones. Thus, in the case of some new project constraint or requests for new requirements, where the project manager needs longer releases, it would be better to intercalate them with shorter cycles (SG21).

Short cycles tend to decrease maintainability Releases in short cycles are usually associated with quick corrections, as mentioned before. In this case, successive short cycles accumulate more hours for corrections than longer cycles, in which the enhancements (not accounted for as corrections) are most likely to be performed. An increase in the effort to correct suggests a decrease in maintainability. However, we also can observe increasing trends for Size and Complexity over successive releases, which also explains the increasing trend in the correction effort. So, again, if there was an opportunity to perform improvements regarding any quality goal, it would be better to include such enhancements in the same release, amongst the corrections, rather than building a release only with quality improvements (SG21). Instead, we should point that, for short cycles where critical corrections have to be done, longer cycles need to be avoided; so, perfective maintenance waits for the next releases.

Stabilization of reliability and maintainability It is possible to observe that corrected defects and the effort to correct become stable (on average) in the long term when a fixed-duration is selected. However, we could not see the same behaviour for the variable-duration strategy. This behaviour suggests that fixed-duration cycles are more suitable for quality control. This way, the alternation between enhancement and correction releases should be done with caution, as some enhancements may generate new defects, penalizing conflicting quality attributes.

7.9 Threats to Validity (SG19 and SG32)

As a general limitation, the model adopted has a perspective that abstracts the process-level details and presents only the behaviour of continuous variables involved in its causal model. So, it is limited to how much explanation the experimenter can get from the model itself. Conversely, considering the scenarios investigated, it is possible to find some explanation in the contextual data.

In terms of construct validity, the choice of hours for corrections as a surrogate for Maintainability is troublesome as it does not take the effort for perfective maintenance into account, such as in refactoring, which also improves maintainability and is usually related to longer cycles.

The results focus on output variables trends, namely Reliability and Maintainability, explaining the general behaviour. However, we can also see both short enhancement cycles and long correction cycles in the initial dataset. This kind of behaviour is suppressed by the trends, obtained by linear regression to generate the model equations, and not generated by the model. Thus, it represents an external validity threat.

8 Final Remarks

In this paper, we presented a set of 33 reporting and planning guidelines for simulation experiments. These guidelines were organized based on studies found in a qSLR and information from other research areas. Basically, we have concentrated our discussions on how the conventional aspects of empirical studies should be considered in simulation experiments, mainly in SE. Our concern with the simulation model and study validity can be justified by the importance this model assumes (it is the main instrument for observation) and the bias the experimental design can promote in result interpretation. So, we hope our efforts can contribute to show the possible benefits well-designed simulation studies can offer to the SE field.

The main motivation for the reporting guidelines work arose from the opportunity to promote the quality of reported studies in Software Engineering, as it is one of the issues identified in the qSLRwe carried out. Additionally, we point that those issues found in the systematic review are still present in the studies reported in the current technical literature, as discussed in Section 5.3. We expect these reporting guidelines will help authors, researchers interested in simulation results, practitioners, and reviewers, whose information has to be presented when reporting simulation-based studies in the context of SE.

Specifically for authors, the contextual and planning information recommended by the guidelines indirectly motivates them to observing some specific features when planning simulation studies in SE. Researchers and practitioners can be aware of core information regarding the SBS results that may be used in their research work, respectively. Examples of such information are context information, threats to validity, conclusions, and applicability. Reviewers, members of conferences and those ones in editorial boards of journals should be able to quickly find the relevant contributions, as well as the evidence confirming the contributions and possible limitations of the SBS.

Although many topics in the reporting guidelines may seem the same of other disciplines and even other research strategies, their content allows the discussion of how they are, and how they should be, presented for SE studies. Some particularities can be seen as Software Engineering which, at least as a scientific field, is not in a mature stage yet. For instance, the lack of knowledge about relevant factors and variables for a given phenomenon, non-consensual terminology, lack of standard metrics, and others. Additionally, intrinsic characteristics such as the both quantitative and qualitative nature of SE phenomena, and the social and technical aspects involved as well.

In general, the current guidelines organization intends to provide a logical understanding sequence, by specifying the next step in a straightforward way. Such a sequence allows a reasonable reasoning flow from goals to output analysis, through discussions involving both model and experimental validity, which can provide support, according to our experience in decision-making. Additionally, the guidelines used can help avoiding some bias and assist in the construction of adequate scenarios.

The proposing of this additional set of planning guidelines does not mean for it to be a process or methodology to perform SBS. Processes for selecting the suitable simulation approach, V&V procedure, or analysis instruments are beyond the purpose of the guidelines. The specifics of any SE domain (such as software processes) or simulation approach (such as SD) are not covered either as this work has a general purpose.

Some of the guidelines proposed were not applicable to our simulation experiment in Section 7, which is the case of guideline SG6 (no initial hypotheses were raised), SG25 (model assumptions not explicitly defined), SG26 (Sensitivity Analysis was considered in the experimental design as it is a characterization experiment), SG27 (only one dataset considered) and SG33 (no comparison to actual data considered due to defined goals).

As our next steps, we plan a further evaluation of these guidelines to assess their contents from the perspective of researchers who need to report and get information from simulation-based studies when we hope to rely on the feedback from the SE community. We are also currently defining an empirical evaluation of this set of guidelines, as regards its usefulness, ease of use and completeness, using non-experienced researchers in simulation and observing how the planning guidelines can be helpful.

Acknowledgments The authors would like to thank the CNPq (Grants 141152/2010-9 and 304795/2010-0) for their support of this work.

References

Al-Emran A, Jadallah A, Paikari E, Pfahl D, Ruhe G (2010) Application of re-estimation in re-planning of software product releases. In: Proc. of International Conference on Software Process. Paderborn, Germany

Alexopoulos C (2007) Statistical analysis of simulation output: state of the art. In: Proceedings of Winter Simulation Conference. doi:10.1109/WSC.2007.4419597

Ali NB, Petersen K (2012) A consolidated process for software process simulation: State of the art and industry experience. In: Software Engineering and Advanced Applications (SEAA), 38th EUROMICRO Conference on (pp. 327–336). IEEE

Andersson C, Karlsson L, Nedstam J, Höst M, Nilsson BI (2002)Understanding Software processes through system dynamics simulation: a case study. In: Proc. of the 9th Annual IEEE International Conference and Workshop on the Engineering of Computer-Based Systems (ECBS)

Araújo MA, Monteiro V, Travassos GH (2012) Towards a model to support in silico studies regarding software evolution. In: ESEM 2012.

Bai X, Huang LG, Zhang H, Koolmanojwong S (2012) Hybrid modeling and simulation for trustworthy software process management: a stakeholder-oriented approach. J Softw Evol Process 24:721

Balci O (1990) Guidelines for successful simulation studies. In: Proc. Winter Simulation Conference (Dec. 9–12), pp. 25–32

Banks J (1999) Introduction to simulation. In: Winter simulation conference, Phoenix, AZ, USA

Barney S, Petersen K, Svahnberg M, Aurum A, Barney H (2012) Software quality trade-offs: a systematic map. Inf Softw Technol 54(7):651–662

Barros MO, Werner CML, Travassos GH (2002) A system dynamics metamodel for software process modeling. Softw Process Improv Pract 7(3–4):161–172

Barros MO, Werner CML, Travassos GH (2004) Supporting risks in software project management. J Syst Softw 70(1–2):21–35

Basili VR (1992) Software modeling and measurement: the goal/question/metric paradigm. Technical report. University of Maryland at College Park, College Park, MD, USA

Biolchini J, Mian PG, Natali AC, Travassos GH, (2005) Systematic review in software engineering: relevance and utility. PESC-COPPE/UFRJ, Brazil. Tech. Rep. http://www.cos.ufrj.br/uploadfiles/es67905.pdf

Birkhölzer T, Pfahl D, Schuster M (2010) Applications of a generic work-test-rework component for software process simulation. In: Proc. of International Conference on Software Process. Paderborn, Germany

Burton A, Altman DG, Royston P, Holder RL (2006) The design of simulation studies in medical statistics. Stat Med 25:4279–4292

Carver JC (2010) Towards reporting guidelines for experimental replications: a proposal. In RESER'10 (May 4), Cape Town, South Africa.

Concas G, Lunesu MI, Marchesi M, Zhang H (2013) Simulation of software maintenance process, with and without a work-in-process limit. J Softw Evol Process 25:1225–1248

Cook TD, Campbell DT (1979) Quasi-experimentation: design and analysis for field settings. Rand McNally, Chicago

Corbin J, Strauss A (2008) Basics of qualitative research: techniques and procedures for developing grounded theory. Sage, Newbury Park

Davis JP, Eisenhardt KM, Bingham CB (2007) Developing theory through simulation methods. Acad Manag Rev 32(2):480–499

de França BBN, Travassos GH (2012) Reporting guidelines for simulation-based studies in software engineering. In: Proc 16th EASE (Ciudad Real, Spain, May 14–15). IET, 156–160

de França BBN, Travassos GH (2013a) Are we prepared for simulation based studies in software engineering yet? CLEI electronic journal, 16:1:8. Available at: http://www.clei.cl/cleiej/papers/v16i1p8.pdf

de França BBN, Travassos GH (2013b) Reporting guidelines for simulation-based studies in software engineering. Technical Report RT-ES 746/13. Available at: http://www.cos.ufrj.br/uploadfile/1368206472.pdf

de França BBN, Travassos GH (2014a) Reporting guidelines for simulation-based studies in software engineering. Technical Report RT-ES 747/14. Available at: http://www.cos.ufrj.br/uploadfile/1409314364.pdf

de França BBN, Travassos GH (2014b) Simulation based studies in software engineering: a matter of validity. In: CIbSE/ESELAW. April. Pucón, Chile

de Mello RM, da SILVA, PC, Runeson, P, Travassos, GH (2014) Towards a framework to support large-scale sampling in software engineering surveys. In: Proc. of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '14). ACM, New York, NY, USA, Article 48, 4 pages. doi:10.1145/2652524.2652567.

Dybå T, Sjøberg DIK, Cruzes DS (2012) What works for whom, where, when, and why? On the role of context in empirical software engineering. In: ESEM'12. Sep 19–20, Lund, Sweden

Eck JE, Liu L (2008) Contrasting simulated and empirical experiments in crime prevention. J Exp Criminol 4: 195–213

Florac WA, Carleton AD (1999) Measuring the software process. Addison-Wesley, Reading

Foss T, Stensrud E, Kitchenham B, Myrtveit I (2003) A simulation study of the model evaluation criterion MMRE. IEEE Trans Softw Eng 29(11):985–995, **November**

Garousi V, Khosrovian K, Pfahl D (2009) A customizable pattern-based software process simulation model: design, calibration and application. SPIP 14:165–180

Grimm V, Berger U, DeAngelis DL, Polhill JG, Giske J, Railsback SF (2010) The ODD protocol: a review and first update. Ecol Model 221(23):2760–2768

Houston DX, Buettner DJ (2013) Modeling user story completion of an agile software process. In: Proc. of ICSS P'13, May 18–19. San Francisco, CA, USA

Houston D, Lieu M (2010) Modeling a resource-constrained test-and-fix cycleand test stage duration. In: Proc. of International Conference on Software Process. Paderborn, Germany

Houston DX, Ferreira S, Collofello JS, Montgomery DC, Mackulak GT, Shunk DL (2001) Behavioural characterization: finding and using the influential factors in software process simulation models. J Syst Softw 59:259–270

Ivarsson M, Gorschek T (2011) A method for evaluating rigor and industrial relevance of technology evaluations. Empir Softw Eng 16(3):365–395

Jedlitschka A, Ciolkowski M, Pfahl D (2008) Reporting experiments in software engineering. In: Shull F et al (eds) Guide to advanced empirical software engineering. Springer, New York

Kitchenham B, Travassos GH, Mayrhauser A, Niessink F, Schneidewind NF, Singer J, Takada S, Vehvilainen R, Yang H (1999) Towards an ontology of software maintenance. JSMRP 11:365–389

Kitchenham B, Pfleeger SL, Hoaglin DC, El Emam K, Rosenberg J (2002) Preliminary guidelines for empirical research in software engineering. IEEE Trans Softw Eng 28:721–734

Kitchenham BA, Al-Kilidar H, Babar MA, Berry M, Cox K, Keung J, Kurniawati F, Staples M, Zhang H, Zhu L (2008) Evaluating guidelines for reporting empirical software engineering studies. Empir Softw Eng 13(1): 97–121

Kleijnen JPC (1975) Statistical design and analysis of simulation experiments. Informatie 17(10):531–535

Kleijnen JPC, Sanchez SM, Lucas TW, Cioppa TM (2005) State-of-the-art review: a user's guide to the brave new world of designing simulation experiments. INFORMS J Comput 17(3):263–289. doi:10.1287/ijoc.1050.0136

Montgomery DC (2008) Design and analysis of experiments. Wiley, New York

Müller M, Pfahl D (2008) Simulation methods. In: Shull F, Singer J, Sjøberg DIK (eds) Guide to advanced empirical software engineering, section I. Springer, New York, pp 117–152

Ören TI (1981) Concepts and criteria to assess acceptability of simulation studies: a frame of reference. Simul Model Stat Comput 24(4):180–189

Pai M, McCulloch M, Gorman JD (2004) Systematic reviews and meta-analyses: an illustrated, step-by-step guide. Natl Med J India 17:2

Paikari E, Ruhe G, Southekel PH (2012) Simulation-based decision support for bringing a project back on track: the case of RUP-based software construction. In: Proc. of International Conference on Software and System Process. Zürich, Switzerland

Petersen K (2011) Measuring and predicting software productivity: a systematic map and review. Inf Softw Technol 53(4):317–343

Pfahl D, Ruhe G (2002) IMMoS: a methodology for integrated measurement, modelling and simulation. Softw Process Improv Pract 7:189–210

Psaroudakis JE, Eberhardt A (2011) A discrete event simulation model to evaluate changes to a software project delivery process. In: IEEE Conference on Commerce and Enterprise Computing, pp. 113–120

Raffo D (2005) Software project management using PROMPT: a hybrid metrics, modeling and utility framework. IST 47:1009–1017

Rahmandad H, Sterman JD (2012) Reporting guidelines for simulation-based research in social sciences. Syst Dyn Rev 28(4):396–411

Runeson P, Höst M (2009) Guidelines for conducting and reporting case study research in software engineering. Empir Softw Eng 14:131–164

Sargent RG (1999) Validation and verification of simulation models. In: Winter simulation conference

Shannon RE (1998) Introduction to the art and science of simulation. In: Medeiros DJ, Watson EF, Carson JS, Manivannan MS (eds) Proceedings of the 1998 Winter Simulation Conference

Sterman JD (2000) Business dynamics: systems thinking and modeling for a complex world. Irwin/McGraw-Hill, Boston

Thomke S (2003) Experimentation matters: unlocking the potential of new technologies for innovation. Harvard Business School Press, Boston

Travassos GH, Barros MO (2003) Contributions of in virtuo and in silico experiments for the future of empirical studies in software engineering. In: WSESE03, Fraunhofer IRB Verlag, Rome

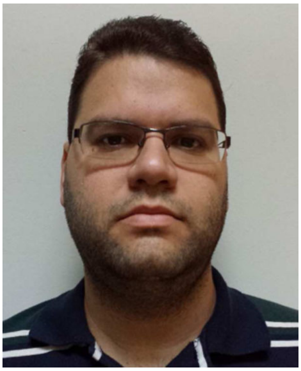
Uzzafer M (2013) A simulation model for strategic management process of software projects. J Syst Softw 86: 21–37

Wakeland WW, Martin RH, Raffo D (2004) Using design of experiments, sensitivity analysis, and hybrid simulation to evaluate changes to a software development process: a case study. Softw Process Improv Pract 9:107–119

Wöhlin C, Runeson P, Host M, Ohlsson C, Regnell B, Wesslén A (2012) Experimentation in software engineering: an introduction. Springer, New York

Yin RK (2008) Case study research: design and methods, vol 5. SAGE Publications, Newbury Park

Zhang H, Klein G, Staples M, Andronick J, Zhu L, Kolanski R (2012) Simulation modeling of a large-scale formal verification process. In: ICSSP 2012, Zürich, Switzerland



Dr. Breno Bernard Nicolau de França is a researcher at COPPE/UFRJ. He received his bachelor and master degrees from UFPA, and he holds a D.Sc. in Systems Engineering and Computer Science from COPPE/UFRJ. He is member of the Experimental Software Engineering Group at COPPE/UFRJ. Further information at http://www.cos.ufrj.br/~bfranca.

Dr. Guilherme Horta Travassos is a professor at COPPE/UFRJ and a CNPq (Brazilian Research Council) Researcher. He holds a D.Sc. in Systems Engineering and Computer Science from COPPE/UFRJ, with a post-doc in Experimental Software Engineering at UMCP/USA. He leads the Experimental Software Engineering Group at COPPE/UFRJ and is a member of ISERN, ACM and Brazilian Computer Society. Apart from that, he is an associate editor of Elsevier – IST and also takes part in the editorial board of World Scientific - IJSEKE and Springer-JSERD. Further information at http://www.cos.ufrj.br/~ght.