COPPE
UFRJ
Instituto Alberto Luiz Coimbra de
Pós-Graduação e Pesquisa de Engenharia

GUIDELINES FOR EXPERIMENTATION WITH DYNAMIC SIMULATION MODELS IN THE CONTEXT OF SOFTWARE ENGINEERING

Breno Bernard Nicolau de França

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia de Sistemas e Computação.

Orientador: Guilherme Horta Travassos

Rio de Janeiro
Maio de 2015

GUIDELINES FOR EXPERIMENTATION WITH DYNAMIC SIMULATION MODELS IN THE CONTEXT OF SOFTWARE ENGINEERING

Breno Bernard Nicolau de França

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

Guilherme Horta Travassos

Prof. Guilherme Horta Travassos, D.Sc.

Cláudia Maria Lima Werner

Profª. Cláudia Maria Lima Werner, D.Sc.

Alvaro

Prof. Álvaro Luiz Gayoso de Azeredo Coutinho, D. Sc.

Márcio de Oliveira Barros

Prof. Marcio de Oliveira Barros, D.Sc.

Prof. Rafael Prikladnicki, D.Sc.

RIO DE JANEIRO, RJ – BRASIL

MAIO DE 2015

França, Breno Bernard Nicolau de

Guidelines for Experimentation with Dynamic Simulation Models in the Context of Software Engineering / Breno Bernard Nicolau de França – Rio de Janeiro: UFRJ/COPPE, 2015.

XIII, 163 p.: il.; 29,7 cm.

Orientador: Guilherme Horta Travassos.

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2015.

Referências Bibliográficas: p. 130-138.

1. Experimentation Guidelines. 2. Dynamic Simulation Models. 3. Simulation-Based Studies. 4. Experimental Software Engineering. I. Travassos, Guilherme Horta II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

À minha família, pelo amor e compreensão.

À Patrícia, pelo apoio quando mais precisei me dedicar ao doutorado.

Agradecimentos

Em primeiro lugar à minha família pelo apoio, amor e incentivo neste período em que estive distante. Em especial à minha mãe, Maria do Carmo, pela eterna preocupação com meu bem-estar mesmo de longe e ao meu irmão, Bruno.

Ao meu pai, Mair, que apesar de não estar mais entre nós, sempre lutou muito para me dar condições de chegar até aqui.

À minha namorada, Patrícia, que sempre me deu apoio mesmo sabendo que teríamos que abrir mão de estar juntos em determinados momentos.

Ao meu orientador, prof. Guilherme Horta Travassos, que mesmo com muitas responsabilidades sempre esteve presente e comprometido com a qualidade final deste trabalho. Ainda, gostaria de mencionar que além da tese, os ensinamentos foram muitos, do ponto de vista ético, científico e profissional.

Aos professores Cláudia Werner, Álvaro Coutinho, Márcio Barros e Rafael Prikladnicki por participarem da minha banca de defesa de doutorado e oferecerem contribuições para sua melhoria.

Aos meus amigos da COPPE, em especial do Grupo ESE, alguns já seguindo sua carreira longe COPPE, pela companhia, discussões, contribuições e todo o dia-a-dia que compartilhamos bons momentos durantes esses anos.

Ao pessoal administrativo do PESC, Claudia Prata, Maria Mercedes, Solange Santos, Sônia Galliano e Gutierrez da Costa pela atenção e disponibilidade.

Ao CNPq, pelo apoio financeiro (processo 141152/2010-9) durante o doutorado.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

DIRETRIZES PARA EXPERIMENTAÇÃO COM MODELOS DE SIMULAÇÃO DINÂMICOS NO CONTEXTO DA ENGENHARIA DE SOFTWARE

Breno Bernard Nicolau de França

Maio/2015

Orientador: Guilherme Horta Travassos

Programa: Engenharia de Sistemas e Computação

Estudos Baseados em Simulação (EBS) têm se mostrado uma interessante abordagem de pesquisa para a Engenharia de Software (ES). Entretanto, é possível identificar a falta de informações relevantes nos relatos desse tipo de estudos encontrados na literatura técnica, dificultando o entendimento dos procedimentos e resultados apresentados, bem como sua repetição. Além das limitações de espaço nas publicações, algumas informações não são apresentadas, aparentemente, devido às questões metodológicas não abordadas na condução dos estudos. Estas e outras questões foram identificadas na condução de uma *quasi*-Revisão Sistemática da Literatura que, após a evolução de uma versão preliminar, resultou na proposta de um conjunto de 30 diretrizes para planejamento e relato de EBS no contexto da ES, cujo objetivo é prover direcionamentos sobre aspectos a serem tratados em diferentes estágios de um EBS, concentrando-se na utilização de modelos de simulação dinâmicos e na identificação e mitigação de potenciais ameaças à validade desse tipo de estudo. O conjunto proposto foi organizado com base em resultados de sucessivos estudos experimentais, utilizando diferentes estratégias de pesquisa. Os resultados das primeiras avaliações indicam que o conjunto de diretrizes proposto é coerente e completo em relação aos aspectos que um EBS deve considerar no planejamento e relato. Ainda, um estudo de observação permitiu caracterizar as diretrizes quanto ao apoio na elaboração e revisão de protocolos para EBS, indicando resultados positivos para eficácia e percepção de utilidade, mas com possibilidades de melhoria principalmente relacionadas à facilidade de utilização. Assim, foi proposta uma nova versão deste conjunto, a qual necessita de avaliações adicionais, sobretudo da comunidade de ES na discussão e aplicação de EBS.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

GUIDELINES FOR EXPERIMENTATION WITH DYNAMIC SIMULATION MODELS IN THE CONTEXT OF SOFTWARE ENGINEERING

Breno Bernard Nicolau de França

May/2015

Advisor: Guilherme Horta Travassos

Department: Computer Science and Systems Engineering

Simulation-based studies (SBS) have become an interesting investigation approach for Software Engineering (SE) research and practice. However, the reports on experiments with dynamic simulation models in the technical literature lack relevant information, hampering the full understanding of the reported procedures and results, as well as their replicability. Apart from the limitations on conference and journal papers length, some of the relevant information seems to be missing due to methodological issues not considered when conducting such studies. These issues were identified in a *quasi*-Systematic Literature Review and, after evolving the preliminary set, lead to a set of 30 planning and reporting guidelines for SBS in the context of SE. This set of guidelines aims at providing orientation regarding relevant aspects to be considered in different stages of the SBS lifecycle, focusing on the use of dynamic simulation models and on the identification and mitigation of potential validity threats. The development of the guidelines is based on results from successive experimental studies, adopting different research strategies. Preliminary evaluation results indicate a complete and coherent set of guidelines as to aspects that should be considered in SBS planning and reporting. Furthermore, an observational study allowed characterizing the simulation guidelines w.r.t. the support to the elaboration and review of SBS protocols, indicating positive results regarding their effectiveness and usefulness, and improvement opportunities mainly related to ease of use. Therefore, we proposed a new version of this set of guidelines, which requires additional assessment, especially from the SE community on the discussion and application of SBS.

INDEX

INDEX OF FIGURES

INDEX OF TABLES

1 Introduction

In this chapter, we present the problem and context of this thesis, as well as the motivation and research questions supporting the investigation. Furthermore, we establish the objectives to be accomplished in order to answer the research questions and how they will be performed through an evidence-based methodology.

1.1 Motivation

Computer simulation is a mature technology with wide application range. Many science areas have benefited from simulation[1] as a supporting tool for analysis and comprehension of systems, processes or phenomena of interest. Engineering, Economics, Biology, and Social Sciences are examples of such areas (MÜLLER and PFAHL, 2008). Additionally, it has been used to support experimentation in both academia and industry.

THOMKE (2003) reported the adoption of this sort of study as an alternative strategy to support experimentation in the automotive industry. In this field, the prototypes used in crash-tests are associated to high costs and to long periods needed to build them. Besides, it is difficult to perform some analysis since they are completely destroyed after crash. In these cases, the use of simulation allows the development of models at lower costs and time than prototypes. Models demonstrated to be very useful to perform feasibility assessments and preliminary tests, even the prototypes being just closer to reality. Furthermore, simulation allows the easily changing of experimental conditions (variables) when investigating different scenarios.

Criminology is another field where researches have taken place with the support of computer simulation to understand how the crime patterns rise (ECK and LIU, 2008). Data on crime are usually unreliable, regardless of the efforts made to improve their quality, since it is inherent to the phenomenon. The sources of crime information are unreliable due to conditions they are imposed to. Actually, simulations can support experimental manipulations that are unfeasible or unethical to conduct on real subjects. Besides, it is very important to describe how the phenomenon really comes about (i.e.,

[1] In this text, both “simulation” and “computer simulation” terms are used interchangeably.

through which mechanism), since statistical explanations do not provide information on how the outputs are generated.

As a wide-range tool, the term simulation varies in meaning from one research community to another. In order to clearly set up the scope under investigation, we adopted the following definition from BANKS (1999):

> *"Simulation is 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".*

This definition is interesting for our scope as it uses the expression over time, which may be replaced by dynamic, in the sense it focuses on behavioural aspects. In other words, it states that we are just interested on how the systems or phenomena perform and how the values of their variables change over time. Therefore, static approaches such as regression models or the Monte Carlo method are not going to be covered in this research.

Simulation-Based Studies (SBS) often involve several activities such as the system observation or data collection, model development (coding), model verification and validation (V&V), experimental design, output data analysis, and implementation of results (ALEXOPOULOS and SEILA, 1998). Generally, such activities comprehend most part of the SBS lifecycles presented in the technical literature (BALCI, 1990) (MARIA, 1997) (BANKS, 1999) (SARGENT, 1999) (BIRTA and ARBEZ, 2007). The lifecycle represents an iterative process and each iteration may encompass both the model development and its use.

On one hand, in order to develop a simulation model, the conceptual model should be coded into a simulation language, which could be based on a simulation approach[2] like System Dynamics (SD), Discrete-Event Simulation (DES), or Agent-Based Simulation (ABS). The simulation approach abstracts the essential characteristics and behaviors the model has to fit. On the other hand, the systematic use of simulation models is called model experimentation or simulation experiments. Besides, it requires the definition of research plans or protocols describing how to perform the simulations.

Among the benefits credited to SBS, it is possible to highlight the low cost and risks associated with the virtual environment where simulations are being executed. This

[2] It is also called simulation paradigm in the technical literature.

is interesting for those scenarios involving real systems that are expensive, safety critical, time-consuming or cause irreversible effects (BIRTA and ARBEZ, 2007). On the experimentation perspective, we can also point out the high degree of control in these environments, as well as the possibility of developing and testing hypotheses or theories, replicating experiments, and enabling the execution of a myriad of combinations for the variables of interest. Conversely, simulation studies may also involve high costs and effort concerned with the model development, generating a tradeoff between developing one single model to perform just one single simulation experiment. Moreover, assumptions and simplifications may not be suitable for specific research contexts or goals, increasing the need to provide evidence on model validity.

Computer simulation is an alternative strategy for SE experimentation. It does not mean that such strategy can replace other types of study such as controlled experiments, case studies and so on. Actually, it is useful for supporting knowledge acquisition and decision-making in the cycle of SE experimentation. Simulation-Based Studies require knowledge from previous observations (*in vivo* and *in vitro* experiments) so that one may create a conceptual model representing a certain SE phenomenon or behavior (TRAVASSOS and BARROS, 2003). Once this knowledge is captured, large-scale observations can be performed, using the controlled environment to understand or characterize the phenomena and possibly explain them through simulation traces. This way, a simulation modeler can design and perform SBS to understand how interventions in software processes and projects affect costs, schedule or quality, for example. Still, it may be possible to characterize the team arrangement for software projects in distributed settings.

Although it may be an interesting approach to evolve the SE research, there are limitations that SBS are imposed to consider due their inherent characteristics. Simulation models require not only knowledge, but require data for calibration, validation and experimentation. Such data come from observations (as mentioned before, through in vivo or in vitro studies) and are constrained to their observational context. Therefore, simulation is feasible when the research goals cannot be achieved by other empirical or experimental strategies, however it must exist enough knowledge and data to support it. Furthermore, simulations are recommended for characterization studies involving the combination of many factors and levels, with possible interactions among factors, long-term observations of software development and maintenance projects, and when risks regarding the real phenomenon are unacceptable in the field.

The Software Engineering (SE) community has presented interesting initiatives on simulation models and studies. The abstractions over SE phenomena involve different domains and perspectives. Such advances concentrate more on software process and project issues, aiming at understanding or improving them in different contexts (DE FRANÇA and TRAVASSOS, 2013b). In addition, it is possible to observe simulation studies concerned with software products, e.g., software architecture decisions regarding quality attributes. However, it is still possible to observe issues in the technical literature, as it will be discussed in the next section.

1.2 Problem Definition

A simulation-based study makes use of a simulation model as the instrument to observe the phenomenon under investigation. It allows understanding, and even optimizing, processes and systems (in the broader sense) with certain control of input parameters, anticipating possible scenarios and configurations representing the system's variants. Moreover, simulation experiments can be performed faster and less costly than *in vivo* or *in vitro* studies (TRAVASSOS and BARROS, 2003).

Besides the advances already achieved by the SE community, it is still likely to observe lack of methodological support to conduct simulation experiments in the context of SE. Such problem includes issues on simulation model validity, inappropriate experimental design for simulation experiments, lack of concerns regarding validity threats and relevant information on the simulation studies report.

Both model validity and experimental design may impose threats to SBS validity when not properly performed. Lack of validation experiments, models not being capable of reproducing reference or empirical behaviors, and unbalanced designs not capable of performing fair comparisons among simulation scenarios are examples of such threats. Furthermore, experimenters are not aware of common threats to simulation study validity and lack relevant information such as research questions, experimental design or evidence regarding the model validity when reporting this sort of study. Therefore, these issues affect the credibility and confidence of studies' results, contributing to reduce the use of SBS as supporting tool for SE research and development. Additionally, they hamper the full understanding of simulation studies, as well as the possibility of replicating their results.

Mostly, methodological support on simulation in the context of SE rely on processes for model development, and particularly on software process simulation (PFAHL and RUHE, 2002) (ALI and PETERSEN, 2012). However, the use of simulation models to perform experiments intends to be generic in such approaches, needing further orientation on how to design, execute and analyze simulation experiments.

The Empirical Software Engineering (ESE) community has already proposed, evaluated and applied guidelines for different research strategies, such as systematic literature reviews (SLR) (KITCHENHAM, 2004), controlled experiments (JEDLITSCHKA, CIOLKOWSKI and PFAHL, 2008), case studies (RUNESON and HÖST, 2009), and replications (CARVER, 2010). Current significant usage of existing guidelines includes not only the methodological support to conduct primary and secondary studies (BORGES *et al.*, 2014), but also to distinguish and assess rigor 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) and (ALI, PETERSEN and WÖHLIN, 2014).These guidelines intend to be drivers to research actions, rather than mandatory recommendations. As they become mature, by identifying advantages on their use and influence on the quality of research protocols, their adoption tends to be natural.

In this direction, we advocate the need for guidelines to conduct simulation experiments, similar to other research strategies. For that, we organized our research as described in the following subsections.

1.3 Research Questions

The research questions for this thesis are derived from the problems stated in Section 1.2, establishing the main directions and scope to be investigated.

Q_1: Which are the relevant aspects to be concerned with when conducting simulation-based studies in the context of Software Engineering?

Q_2: How to conduct simulation-based studies in the context of Software Engineering in order to accumulate evidence regarding the study's validity to increase its confidence and credibility?

Q_3: Which information should compose research protocols (plans) and reports for simulation-based experiments in the context of Software Engineering in order to provide their full understanding and to enable their replicability?

These three research questions are concerned with methodological support. The aspects under investigation regard the organization of planning and reporting issues, including simulation model validity, experimental design, and output analysis. We highlight these three aspects, since they concentrate on methodological issues that usually lead to the occurrence of validity threats. However, other relevant aspects should also be addressed for a coherent study plan and we discuss later in the following chapters. In the next section, we describe our objectives as milestones to reach as we progress in our research.

1.4 Research Objectives

The goals for this research are derived from the problems stated in Section 1.2, and established to answer the research questions defined in Section 1.3. Research goals are structured in general and specific goal (also called objectives). The general research goal consists on proposing an approach to support the planning and reporting of simulation-based studies in the context of SE, focusing on their experimental validity.

The objectives bellow are established as steps that should be accomplished to answer the research questions (Q_1, Q_2 and Q_3):

- For Q_1:
    - O_1: To characterize SBS in the context of SE by organizing a body of knowledge;
    - O_2: To identify which are the relevant information composing a SBS report;
- For Q_2:
    - O_3: To identify potential threats to simulation studies validity;
    - O_4: To identify Verification and Validation (V&V) procedures for simulation models;
    - O_5:To identify experimental designs and output analysis instruments applicable for simulation experiments;
- For Q_3:
    - O_6: To organize a set of guidelines to support the planning of simulation experiments in the context of SE;
    - O_7: To organize a set of guidelines to support the reporting of simulation experiments in the context of SE;

Apart from the presented objectives, we have a general goal of evaluating all proposed sets of guidelines, including their different versions, in an iterative process. These evaluation cover different aspects concerned with the guidelines validity.

Finally, the achievement of each of the seven objectives in isolation does not answer the research questions. However, their joint perspective along with the evolutionary methodology (next section) compose the required knowledge to support the answers.

1.5 Research Methodology

The research methodology for this work starts with an initial investigation consisting on defining the problem (Section 1.2) and research questions (Section 1.3). Figure 1-1 presents the stages composing this research methodology.

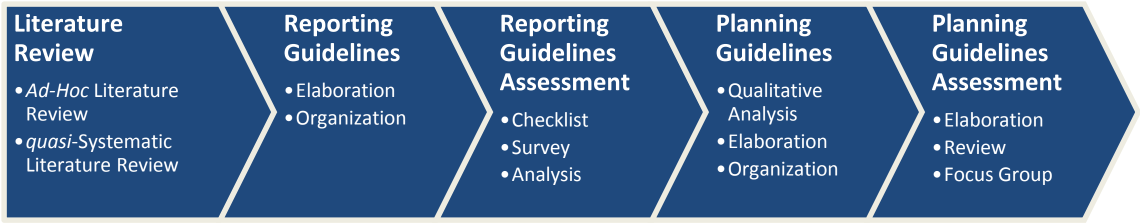


Figure 1-1. Research Methodology (DE FRANÇA and TRAVASSOS, 2015).

With this initial investigation, we performed an *ad-hoc* literature review in order to capture the terminology and understanding regarding simulation in SE. At this stage, both general purpose and SE simulation books were consulted, as well as relevant (high-cited) simulation papers. This material was helpful to establish the review's scope. Based on such information, we started the elaboration of a research protocol to undertake a *quasi*-Systematic Literature Review (qSLR) (DE FRANÇA and TRAVASSOS, 2012) (DE FRANÇA and TRAVASSOS, 2013b) for a broader, repeatable and systematic procedure.

The research protocol followed the guidelines proposed by BIOLCHINI *et al* (2005), 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 the adopted simulation approaches, SE domains, model validation issues, simulation procedures and experimental design, and output analysis. For that, three digital libraries were selected as source of information. Pre-defined selection and extraction procedures were defined and executed, followed by the quality assessment and the analysis of findings. See Chapter 2 for details regarding the qSLR.

From the qSLR, we observed common issues across different reports: lack of simulation model description and validation information, no description of how simulation experiments have been performed, no discussions regarding the output analysis and threats to validity. Moreover, this lack of information hampers the full understanding of the simulation study results and the ability to reproduce them as well. This way, we searched for guidelines concerning with what should be a reasonable set of information to be reported on simulation studies. We could not identify such information contextualized for SE Research. However, we did find some related discussions regarding other research areas.

Using the findings from the review and additional reporting guidelines concerning with 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 was evolved through sequential evaluation initiatives, including a checklist-based review, using the instruments proposed by KITCHENHAM *et al* (2008); a collaborative review, which was structured as an online survey; and, finally analyzed against technical literature reports, obtained from the qSLR update.

The checklist-based review (details in Section 4.2) was originally published as a method inspired on the perspective-based reading using checklists to guide the reviewers. Each checklist represents a particular perspective. The aim of this approach is to evaluate the reporting guidelines using different reviewers and reaching consensus. Our main goal in this assessment is to align our perspectives to the existing guidelines on Empirical Software Engineering (KITCHENHAM *et al.*, 2002) (JEDLITSCHKA, CIOLKOWSKI and PFAHL, 2008) (RUNESON and HÖST, 2009), and keep a broader sense of the study report.

After the checklist-based review, we still needed external evaluation of these reporting guidelines, based on the opinion of knowledgeable people in both Simulation and Software Engineering. Thus, we structured a collaborative review (details in Section 4.3), very similar to conference reviews, using an online survey platform. Our intention with this study is to get feedback regarding the completeness and correctness of the guidelines set.

Finally, after the two former assessments, we analyzed the reporting guidelines against the technical literature (details in Section 4.4). For that, we updated the qSLR, essentially using the same research protocol, to verify whether new study reports somehow comply with the proposed guidelines.

All these initiatives enabled us to get feedback regarding the completeness and correctness of the reporting guidelines. Thus, the current version (reporting only) was made fully available in (DE FRANÇA and TRAVASSOS, 2014a).

Considering we have established a set of relevant information to be reported, it is essential to understand the stage in the SBS lifecycle and how such information should be produced. Additionally, we understood that, if not planned for the simulation study, the researchers are not likely to produce them, since most of them cannot be produced in retrospect.

The next step in our research methodology concerns with evolving the set of reporting guidelines in order to support planning activities in simulation experiments. The additional guidelines do not mean to embrace the simulation model development, but only the simulation model use, also called model experimentation (BALCI, 1990). This is justified by the existence of methodological support for simulation modeling in the context of SE. Some examples include (PFAHL and RUHE, 2002) (ALI and PETERSEN, 2012).

Aiming at supporting the elaboration of *complete*, *coherent* and *effective* simulation plans, the planning guidelines were developed based on the results of a secondary analysis of the outcomes identified through the qSLR and additional resources. Such analysis followed a qualitative approach, aiming at identifying potential threats to SBS validity in the context of SE, as well as possible mitigation actions (DE FRANÇA and TRAVASSOS, 2014b).

- By *complete* we mean that the plan should encompass all the aspects considered in EBS reports, enabling to produce the expected information.

- *Coherent* plans have a strong linkage among different aspects, for instance, the alignment of problems to the goals and research questions, the clear derivation of the experimental design from the research questions and hypotheses, the output analysis reflecting the experimental design and the analysis of threads to validity concerning model validity issues.
- By *effective*, we mean executable plans, with output analysis and conclusions including reasoning and explanations for the results, limited by known imposed threats to the simulation experiment validity.

At the final stage of our methodology, we performed a qualitative evaluation of the planning guidelines, including a Focus Group focusing on their usefulness and ease of use (DE FRANÇA *et al.*, 2015) (see Chapter 6).The aim of this first assessment is to show the feasibility of the proposed set of guidelines to support simulation experiments in SE.

In general, the main contributions of this work include (1) the body of knowledge resulting from the qSLR, (2) the proposed reporting and (3) planning guidelines for simulation studies, (4) the list of identified threats to simulation studies validity, and (5) the performed evaluations for both sets of guidelines, as described in (DE FRANÇA and TRAVASSOS, 2015). Therefore, the resulting set of simulation guidelines (statements only) including planning and reporting perspectives is presented in Table 1-1.

Table 1-1. Simulation Guidelines Overview.

ID	Guideline Statement
Identification	
SG1	Proper title and keywords should objectively identify the simulation study, and a structured abstract should summarize its contents
From Context to Research Questions	
SG2	The context where the simulation study is taking place should be captured in full
SG3	Explicitly state the problem motivating the simulation study, so that research questions can be derived
SG4	Clearly state the simulation study goals and scope
SG5	Derive the research questions from the established goals
SG6	Clearly state the null and alternatives hypotheses from the research questions
Simulation Feasibility	
SG7	Present justifications for considering simulation studies as the ideal or feasible observation strategy
Background and related work	
SG8	Present only essential background knowledge and the related works
Simulation Model Specification	
SG9	Have a detailed description and understanding of both conceptual and executable simulation models, as well as its variables, equations, input parameters and the underlying simulation approach
Simulation Model Validation	
SG10	Gather all evidence regarding the simulation model (conceptual and execution) validity
SG11	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

ID	Guideline Statement
SG12	Support model (causal) relationships, as much as possible, with empirical evidence to reinforce their validity and draw more reliable conclusions
SG13	Always verify the model assumptions, so the results of simulated experiments can get more reliable
Subjects	
SG14	Characterize the subjects involved in the simulation study as well as their training needs
Experimental Design	
SG15	Describe the experimental design (design matrix), including independent and dependent variables and how levels are assigned to each factor
SG16	Use Sensitivity Analysis to select valid parameters settings when running simulation experiments, rather than model “fishing”.
SG17	Consider as factors (and levels) not only the simulation model’s input parameters when designing the simulation experiment, but also internal parameters, different sample datasets and simulation model versions, implementing alternative strategies to be evaluated
SG18	When adopting ad-hoc design determine the selected simulation scenarios and explain the criteria used to identify them as relevant
SG19	When dealing with simulation model containing stochastic components, determine the number of runs required for each scenario, along with its rationale, in order to capture the phenomenon variance.
Supporting Data	
SG20	Assess, whenever possible, the data used to support the simulation model development or experimentation
SG21	Keep track of contextual information (including qualitative data) along with quantitative data
SG22	Make sure that both calibration and experiment datasets came from the same population
Simulation Supporting Environment	
SG23	Set up and describe the simulation environment, including the supporting tools, associated costs, and decision for using a specific simulation package
SG24	Determine which and how intermediate measures are stored among simulation trials to be used in the final analysis
Output Analysis	
SG25	Determine which statistical procedures and instruments support the output analysis, as well as the underlying rationale, quantifying the amount of internal variation embedded in the (stochastic) simulation model to augment the precision of results
SG26	Be aware about data validity when comparing actual and simulated results: compared data must come from the same or similar measurement contexts
Threats to Validity	
SG27	Consider to check for threats to the simulation study validity before running the experiment and analysing output data to avoid bias, as well as to report non-mitigated threats, limitations and non-verified assumptions
Conclusions and Future Works	
SG28	Main results/findings should be identified and summarized, as well as the conclusions arising from the results.
SG29	Applicability issues should be addressed in the report, considering organizational changes and associated risks.
SG30	Point out future research directions and challenges after current results.

1.6 Thesis Organization

This chapter presented the main concepts, motivation, methodology and directions for this thesis. Next chapters are organized as follows. Chapter 2 presents the results from the technical literature review, delineating the scope and challenges for this thesis. Chapter 3 presents related work regarding reporting guidelines for different research strategies and encompassing several research fields. Besides, this chapter presents an overview of the proposed reporting guidelines for SBS in the context of SE. Chapter 4 presents three successive assessments performed to evolve the reporting guidelines, including discussions on how they improved the set of guidelines, as well as their limitations. Chapter 5 discusses threats to SBS validity, as well as an overview of the planning guidelines for simulation-based experiments. Chapter 6 presents the observational study performed to evaluate the planning guidelines. Chapter 7 presents the current full version of the guidelines for simulation-based experiments in SE. Finally, Chapter 8 presents our conclusions and future work.

2 Literature Review

In this chapter, we present the conducted technical literature reviews aiming at characterizing simulation-based studies in the context of Software Engineering. The findings compose the essential knowledge basing all contributions of this thesis.

2.1 *Ad-hoc* Review

According to the research methodology presented in Section 1.5, we performed an *ad-hoc* technical literature review in order to capture the basic terminology concerned with computer simulation, as well as some initial understanding about the simulation studies' lifecycle.

Since computer simulation is a well-established and steady research field, this review is based initially in textbooks such as (ABDEL-HAMID and MADNICK, 1991), (SEVERANCE, 2001), (BIRTA and ARBEZ, 2007), (SOKOLOWSKI and BANKS, 2009) and (MADACHY, 2007). These textbooks generally classify the simulation approaches as continuous and discrete-time ones, where discrete-event simulation is an example of discrete-time simulation, and System Dynamics (SD) is a relevant approach for continuous-time simulation.

Several characteristics concerned with simulation practices were explored with respect to systems modeling and statistical output analysis, as well as the associated advantages and disadvantages, some of them already mentioned in Section 1.1.

Apart from the textbooks, simulation papers were also consulted, including tutorials from the Winter Simulation Conference (WSC, wintersim.org), where general modeling and simulation issues such as model verification and validation are discussed under both theoretical and practical perspectives. Seminal research papers regarding simulation studies in SE were also considered (LUCKHAM *et al.*, 1995) (MADACHY, 1996) (DRAPPA and LUDEWIG, 2000). Some of these works were used as control papers (see next section) for the qSLR.

2.2 *quasi*-Systematic Literature *Review*

Systematic Literature Reviews (SLR) started to be used in SE in the early 2000s, as a building block of the Evidence-Based paradigm (KITCHENHAM, DYBA and

JORGENSEN, 2004). SLR are also mentioned as "a means of evaluating and interpreting the available research relevant to a particular research question, topic area, or phenomenon of interest" (DYBÅ, DINGSøYR and HANSSEN, 2007). Earlier works on this topic used to name all reviews performed with some systematic process as SLR. However, many of them did not follow specific and fundamental aspects or characteristics usually expected in systematic reviews, such as comparison among the outcomes w.r.t. their quality and possibilities of synthesis or aggregation. In this context, the term *quasi*-systematic literature review (TRAVASSOS *et al.*, 2008) appeared as a definition for reviews following SLR guidelines, but not covering at least one aspect (i.e., no comparison), which is the case of the literature review presented in this chapter. Thus, the "*quasi*" term stands for the unfeasibility of comparing outcomes due to lack of knowledge on the field or specific domain of investigation, also limiting the definition of quality profile for the available evidence, based on a hierarchy of evidence in SE.

In the context of SE, the systematic review presented in (ZHANG, KITCHENHAM and PFAHL, 2008) characterizes Software Process Simulation Models (SPSM) by tracing the research evolution from 1998 to 2008. The authors highlight their main results: the need for adjustment in categories for classifying SPSM to better capture the diversity of models available in the technical literature; improvements on the efficiency of SPSM promoted by research; and more realistic simulation models using hybrid approaches.

Assuming broader scope and goals, we undertook a qSLR aiming at characterizing simulation-based studies performed in the Software Engineering research area (DE FRANÇA and TRAVASSOS, 2013b), rather than focusing on software process simulation models as in (ZHANG, KITCHENHAM and PFAHL, 2008). Next sections present an overview of the research protocol and main results.

2.2.1 Research Protocol

The research question driving the protocol elaboration is "How different simulation approaches have been applied to Simulation-Based Studies in the context of Software Engineering?"

In order to answer this research question, we structured the search string using the PICO strategy (PAI *et al.*, 2004), composed by the following dimensions:

- Population: Simulation-based studies in Software Engineering;

- Intervention: Simulation models used as instruments[3];
- Comparison: None;
- Outcome: Purpose, characteristics, SE domain, experimental design and other relevant aspects regarding the simulation study.

Selected sources of information are digital libraries available via Portal CAPES[4]. Sources were selected according to criteria: (1) allowance of executing searches using logical expressions, (2) application of the search string to the publication title, abstract and keywords, and (3) include relevant conference and journal papers from the SE research area. This way, Scopus, EI Compendex and Web of Science were selected. These libraries encompass publications from the main venues regarding computer simulations, software engineering, and related areas, such as ACM, IEEE, Elsevier, Springer and WILEY. This way, the following search string was adapted and submitted to the search engines:

POPULATION: (("simulation modeling" OR "simulation modelling" OR "in silico" OR "in virtuo" OR "simulation based study" OR "simulation study" OR "computer simulation" OR "modeling and simulation" OR "modelling and simulation" OR "simulation and modeling" OR "simulation and modelling" OR "process simulation" OR "discrete-event simulation" OR "event based simulation" OR "system dynamics" OR sampling OR "monte carlo" OR "stochastic modeling" OR "agent based simulation" OR "state based simulation") AND ("software engineering" OR "systems engineering" OR "application engineering" OR "software development" OR "application development" OR "system development"))AND

INTERVENTION: ("simulation model" OR "discrete event model" OR "event based model" OR "system dynamics model" OR "agent model" OR "state model"))

For the search string, we purposely suppressed the PICO's Outcome dimension. The justification consists in the fact that, usually, authors do not include terms like "characteristics", "domain", "experimental design", "validation" and other relevant terms in the publication's abstract when reporting SBS. Therefore, we considered such information

[3] The instrumentation provides means for performing an experiment and to monitor it, without affecting the control of the experiment (WÖHLIN *et al.*, 2012).

[4] Digital libraries containing scientific publications can be accessed from Brazilian institutions through the CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) portal, available at: http://www.periodicos.capes.gov.br.

only in the information extraction stage. Secondary studies aiming at characterizing SE may face compatibility issues on adopting the PICO strategy, as it forces a causal structure.

In order to calibrate the search string we tested it on the Scopus search engine, using the controls papers (Table 2-1) captured in the previous *ad-hoc* literature review (section 2.1). These research papers were used as control as they satisfy the inclusion criteria and help to answer the research question.

Table 2-1. Control Papers

Reference
Martin, R.; Raffo, D. Application of a hybrid process simulation model to a software development Project. Journal of Systems and Software, Vol. 59, Issue 3, 2001, pp. 237-246
Khosrovian, K.; Pfahl, D.; Garousi, V. GENSIM 2.0: A customizable process simulation model for software process evaluation. LNCS, Vol. 5007, 2008, pp. 294-306
Drappa, A.; Ludewig, J. Simulation in software engineering training. Proc. ICSE 2000, pp. 199-208
Madachy, R. System dynamics modeling of an inspection-based process. Proc. ICSE 1995, pp. 376-386
Al-Emran, A.; Pfahl, D.; Ruhe, G. A method for re-planning of software releases using discrete-event simulation. Software Process Improvement and Practice, Vol. 13, Issue 1, Jan 2008, pp. 19-33
Al-Emran, A.; Pfahl, D.; Ruhe, G. DynaReP: A discrete event simulation model for re-planning of software releases. LNCS, Vol. 4470, 2007, pp. 246-258
Luckham, D. C.; Kenney, J. J.; Augustin, L. M.; Vera, J.; Bryan, D.; Mann, W. Specification and analysis of system architecture using rapide. IEEE TSE, Vol. 21, Issue 4, April 1995, pp. 336-355
Arief, L. B.; Speirs, N. A. A UML tool for an automatic generation of simulation programs. Proc. Second WOSP, 2000, pp. 71-76
Choi, K.; Bae, D.H.; Kim, T. An approach to a hybrid software process simulation using the DEVS formalism. Software Process Improvement and Practice, Vol. 11, Issue 4, July 2006, pp. 373-383

The basic selection procedure is based on the paper’s title and abstract. For that, a set of inclusion and exclusion criteria is established *a priori*. So, only papers available in the Web; written in English; discussing simulation-based studies; belonging to a Software Engineering domain; and those mentioning one or more simulation models, should be included. Papers not meeting all of these criteria should be excluded.

Three researchers (R1: the author of this thesis, R2: an external researcher experienced in SBS in the context of SE, R3: the advisor of this research) were involved in the selection of potential relevant papers. R1 did the searching, retrieving the papers, saving their references (including abstract) in the reference manager (JabRef tool), adding a field to represent the paper status (I - Included, E - Excluded, D – Doubt), and removing possible duplicates. After that, the first selection (based on the inclusion criteria – reading title and abstract) was done and the status for each paper was assigned. Next, R2 received the Jabref file (in BibTeX format) with the references and status information, and reviewed the included and excluded papers. In the case of updating any paper status, R2 set up paper status as D2 and tried to solve queries from R1 by setting the former D's as I2 or E2. Lastly, R3 went through the same R2 procedures, though tagging them as I3 or E3. The papers remaining in D after three reviews were included for post-analysis in the information extraction stage. All papers included in the selection stage could be later excluded in the extraction stage, where they are integrally read to improve understanding, clearing any doubts and allowing a better decision on their inclusion or exclusion according to the criteria.

After the selection stage, information of interest was extracted according to the form presented in Table 2-2, also stored in the Jabref tool. These information were proposed as they are meant to answer the research question for this review.

Table 2-2. Information Extraction Form (DE FRANÇA and TRAVASSOS, 2013b)

Field *[Extracted Information]*
Paper identification *[Title, authors, source, year, publication type]*
Simulation approach name
Simulation model purpose *[Objective for the model to be developed]*
Study purpose *[Objective for the study to be performed]*
Software Engineering Domain *[Application area]*
Tool support *[Does the model used have any tool support? If so? Which ones?]*
Characteristics related to the model *[Examples: discrete, continuous, deterministic, stochastic]*
Classification or taxonomy for the characteristics *[Does it have any characteristic classification?]*
Simulation model advantages
Simulation model disadvantages
Verification and validation procedure *[V&V techniques used to evaluate the simulation model]*
Analysis procedure *[Output analysis methodology applied to the study]*
Study strategy *[Controlled experiment, case study, among others]*
Paper main results *[Applicability of the approach, accuracy of results]*

With all extracted information from the selected papers, the analysis of results should take into account their quality, considering the research goal of the systematic review. For that, we defined a set of quality criteria based on the extraction form (Table 2-3). The scores are in a [0-10] scale in which we weighted each criterion according to the relevance for the characterization of the SBS.

2.2.2 Results

The application of the search string in the three search engines by April 2011 resulted in 1,492 records (906 from *Scopus*, 85 from *Web of Science* and 501 from *EI Compendex*), removed 546 duplicates, resulting in 946 records for the selection procedure execution. Initially, 150 research papers were included for the extraction and analysis stages. However, 28 of them were excluded after detailed reading and 14 research papers were unavailable (even after soliciting them to the authors), remaining 108 research papers for the analysis (APPENDIX F).

Table 2-3. Quality assessment criteria (de FRANÇA and TRAVASSOS, 2013)

Criteria	Value
[Approach] Does it identify the underlying simulation approach?	*1 pt*
[Model Purpose] Does it explicitly mention the simulation model purpose?	*1 pt*
[Study Purpose] Does it explicitly mention the study purpose?	*1 pt*
[Domain] Is it possible to identify the SE domain in which the study was undertaken?	*1 pt*
[Tool] Does it explicitly mention any tool support?	*0,5 pt*
[Characteristics] Does it mention the characteristics on the simulation model?	*0,5 pt*
[Classification] Does it provide any classification or taxonomy for the characteristics mentioned?	*0,5 pt*
[Advantages] Does it present the advantages of the simulation model?	*0,5 pt*
[Disadvantages] Does it present the disadvantages of the simulation model?	*0,5 pt*
[V&V] Does it provide any verification or validation procedure for the simulation model?	*1 pt*
[Analysis] What are the statistical instruments used in the analysis of the simulation output?	*1 pt*
[Study Strategy + Exp Design] Is it possible to identify the study strategy in which the simulation model is used as an instrument as well as the experimental design?	*1,5 pt (0,5 for strategy + 1pt for exp. design)*

By analyzing these 108 papers, we identified 88 simulation models, distributed into 17 SE domains. In summary, we found 19 simulation approaches, 27 supporting tools, 28 model characteristics, 22 output analysis instruments, 9 verification and validation procedures, all of them in the SE context.

Among the studies, the most dominant combination concerns with Software Project Management issues using System Dynamics. Besides, it is possible to identify two secondary studies, 57 primary studies and 49 examples of use [assertions or informal studies (ZELKOWITZ, 2007)].

Aiming at characterizing simulation studies, we concentrated more effort on their analysis rather than simulation model descriptions. It is possible to observe the lack of rigor regarding planning activities, Verification and Validation - V&V (before executing simulation experiments) in order to assure minimal confidence on the simulation results, and output analysis procedures. In general, such activities are performed *ad-hoc* for the identified SBS. However, it is also possible to observe particular studies using systematic procedures to perform one or another activity, but no case presents a full systematic process or method to conduct the simulation study.

Specially about the planning issues, i.e., the experimental design, it is possible to observe that the so-called control, often claimed as an advantage of simulation studies, is poorly explored and inadequately used in some situations. There seems to be a misunderstanding of the term “control” in simulation studies. The simple act of varying input parameters does not mean that the object of study is under control, but only a possible configuration of parameters describing a specific scenario. It is quite different from the control in controlled experiments, for instance, where a baseline is defined (control group) under which it is possible to perform a fair comparison among two or more factors.

In this characterization, we also observed some reporting issues. For instance, reports do not present information on how the phenomena, systems or processes are observed to acquire data. Still, it is not possible to identify whether any system (or process) under observation considers its real environment or just its specifications. In the same way, no study protocol or plan could be identified.

Generally, we assigned these issues to the reports rather than to the studies, because we have no access to additional information regarding the actual study. Thus, except for some studies that we can identify lack of proper methodology, they were assessed only based on their report.

2.2.3 Quality Assessment

Based on the extracted and analyzed information, we applied the quality assessment criteria from Table 2-3 to the 108 research papers focusing on the reported information. Therefore, it is possible to observe lack of relevant information. Figure 2-1 presents the general assessment results, where each evaluation criterion is scaled in terms of coverage regarding the 108 research papers.

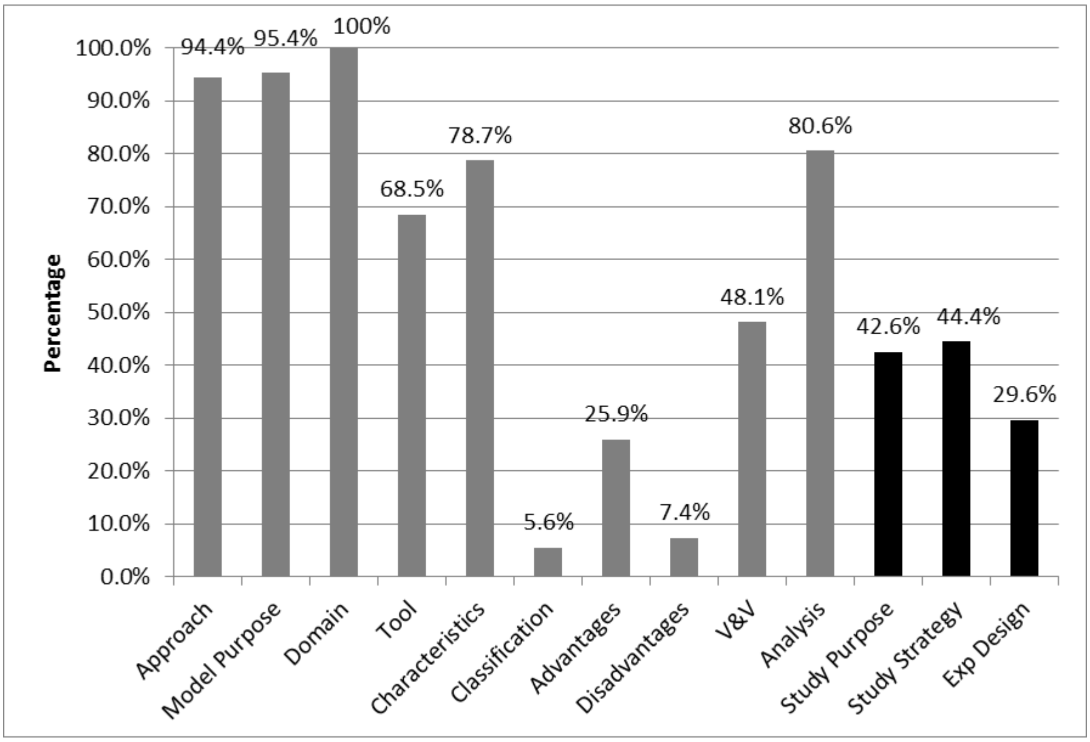


Figure 2-1. Quality Assessment Results (DE FRANÇA and TRAVASSOS, 2013b)

Only 5.6% of the studies report some classification for the model characterization. These classifications are usually based on the underlying simulation approach. Besides, most of the studies (51.9%) do not mention any V&V procedure applied to the reported simulation model. Moreover, analysis instruments are mentioned in 80.6% of the reports, but with no concern of describing how they were selected and used. Furthermore, darkest bars highlight specific information regarding the simulation experiments. In this case, it is possible to observe that only 29% present some information (not necessarily complete information) about experimental design issues.

Concerning the assigned scores (based on values from Table 2-3), the mean value is 6.16 in a [0-10] scale, with standard deviation of 1.41. The minimum score is 3.0 and maximum is 9.0. Considering each criterion concerns one kind of relevant information, possibly not presented in the report, we can understand such values as lack of relevant information in the reports, since they hamper their full understanding and the possibilities of auditing or future repetition.

2.2.4 Threats to Validity

Terminology issues have been constantly reported in published systematic reviews, and this review is not an exception for that. In order to mitigate the effects of not having a well-defined terminology, we previously performed an *ad-hoc* review to organize an initial set of keywords. Then, the research protocol including these keywords was submitted to the review of two experts about simulation in SE, aiming at identifying possible new terms. Besides, as we were targeting a wide-range context for SE, we choose not to adopt specific terms from each SE domain, such as Software Architecture or Design, Inspections, Testing, and others. However, it may have hindered simulation studies not using general terms.

Although the selected digital libraries (Scopus, EI Compendex and Web of Science) encompass the main publications regarding computer simulation and software engineering, it is not possible to assure 100% coverage. However, we consider to have used a comprehensive sample for characterization purposes.

Furthermore, there are threats to validity concerning the extensive work on the studies' selection and information extraction that, as any other human-intensive work, is error-prone. For that, we involved three researchers in the selection procedure in order to reduce the bias. Moreover, our results are exposed to publication bias, since the vast majority of papers published in SE reports only successful cases, rarely presenting negative aspects of the work.

2.2.5 Conclusions

Based on the review's findings, it is possible to observe that, in the SE context, even recognizing simulation approaches and model purpose, it seems there is no concern with the matching between the approach and the study goal. In order to verify whether the phenomena domain has influenced the simulation approach selection, we performed analyses among model characteristics and the SE domain, but we could not identify any trend in this sense.

There is a strong trend in developing simulation models. However, the simulation studies performed using such models face several threats to validity, compromising their results. Mainly, these threats are related to the model validity and the lack of a proper experimental design, as well as the presence of *ad-hoc* output analysis leading to untimely conclusions.

The findings are encouraging for future research in simulation-based studies, in the context of SE, with respect to methodological support. This way, we believe that proposing or adapting methodological support for conducting SBS, including model validity issues and statistical output analysis are practical challenges. Currently, it is possible to identify simulation-modeling approaches for Software Process Simulation with SD, which is the case of IMMoS (PFAHL and RUHE, 2002). However, such methodology does not embrace, at least in detail, activities concerned with the planning and execution of simulation experiments.

Methodological issues mentioned above are associated with many challenges regarding SBS, in the context of Software Engineering, according to our findings, for instance:

- Definition of research protocols for SBS. For the majority of study strategies applied to SE, researchers adopt plans containing contextual descriptions, goals, interest variables, procedures and instruments, subjects' selection and assignment. However, such concerns are usually overlooked when conducting SBS;
- Proposals and use of verification and validation procedures for simulation models. This challenge is also mentioned in the survey conducted by AHMED *et al* (2008), with practitioners using simulation as an alternative approach to study software processes;
- Mitigation actions for threat to simulation studies validity. Such challenge is strongly related to simulation model validity;
- Evaluation and assessment of quality and strength of evidence obtained from SBS. As this sort of study is verifiable, we need established criteria for judging

how close the outcomes are from reality, as well as analyzing how threats to validity influence on the quality of observed evidence;

- Repetition of SBS, concerning the difficulties on capturing relevant information from the reports, such as: full description of simulation models and environment, and raw data publication regarding the calibration and simulation data.

Besides the methodological issues, there are also issues on quality of reports found in the SE technical literature. Specifically, the lack of relevant information for the understanding of the SBS in the context of SE. We decided to handle reporting issues prior to methodological ones as we judge them more feasible considering the knowledge acquired at that time of the research. Reporting guidelines can handle such issue regarding which information is essential for the reports. Contextual information, research goals and questions clearly defined, simulation model specification, validation procedures for the simulation model, and the experimental design are examples of relevant information to compose the report. For that, we organized a set of reporting guidelines for SBS in the context of SE (Section 3.3), aiming at filling the gaps observed in the reports found through the qSLR.

Finally, it is worth to mention that we could not observe any improvement in the quality of reports for more recent studies. Lack of information seems to be independent of SE knowledge on simulation studies.

3 Reporting Guidelines for Simulation Studies

In this chapter, we discuss reporting guidelines for simulation-based studies in the context of Software Engineering. For that, we start with reporting guidelines for other research strategies in the context of Software Engineering. Next, we discuss existent simulation guidelines related to our goal from other research areas. Finally, we present the overview of the proposed reporting guidelines.

3.1 Reporting guidelines for Software Engineering studies

In the SE research community, there are initiatives concerning with the orientation on planning, execution and reporting of experimental studies. Generally, such initiatives present a set of aspects relevant to experimental studies and, for each aspect, they present an associated discussion, as well as examples showing what and how to handle it. For instance, these guidelines include aspects such as the determination of research context, the experimental design, data collection and presentation of results.

KITCHENHAM *et al* (2002) proposed a preliminary set of guidelines to support researchers, reviewers and meta-analysts in the design, conduction and evaluation of SE studies. It presents general guidelines, embracing any type of primary study. Besides, they highlight the need for developing and evaluating specific guidelines for each type of study.

Therefore, JEDLITSCHKA, CIOLKOWSKI and PFAHL (2008) proposed guidelines for conducting and reporting controlled experiments in SE. These guidelines discuss issues on redundant information on reports, textual elements (title, structured abstract) and aspects particular to controlled experiments, such as experimental unit, instruments, procedure, hypotheses, dependent and independent variables, the experimental design and the separation of subjects in groups for the treatments application. Furthermore, they discuss general aspects such as research goals, data collection and analysis, all under the perspective of controlled experiments.

In the sequence, RUNESON and HÖST (2009) proposed a similar set of guidelines, but for case studies in SE. Besides general aspects, these guidelines present specific issues regarding case studies definition and conduction, aiming to distinguish such strategy from the others, considering this term is often misused in the technical literature.

Other initiatives can be mentioned, for instance, for reporting replicated studies (CARVER, 2010). However, it is important to highlight the need for such guidelines considering the heterogeneity and lack of standardization of reports negatively affect their aggregation or synthesis, hindering the understanding of results, among other problems.

Recently, ALI and PETERSEN (2012) 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 is possible to identify some overlapping between their initial planning concerns and the proposed guidelines in this thesis, such as using GQM for goals and questions definition, assessing the feasibility of using simulation, and model validation. However, the remaining guidelines focus on model development rather than model experimentation.

3.2 Reporting guidelines for simulation

Considering the results from the qSLR (Chapter 2), we searched for similar guidelines to the ones previously mentioned in section 3.1, but specifically for SBS in SE. However, we could not find any work targeting both areas. As a result, we searched for similar guidelines in different research areas, which are known by the success on using simulation studies to support their researches and professional activities.

We identified orientations and relevant information on reporting simulation concerned with Computer Simulation (KLEIJNEN, 1975) (BALCI, 1990), Statistics (ÖREN, 1981), Medicine (BURTON *et al.*, 2006) and Social Sciences (RAHMANDAD and STERMAN, 2012)(RAHMANDAD and STERMAN, 2012).

ÖREN (1981) presents a series of concepts and criteria to evaluate credibility and acceptance of SBS. The mentioned concepts concern with input data, model (conceptual and execution), experimental design, and the chosen methodology to conduct the study. BALCI (1990) presents guidelines for the success of SBS organized according to the SBS lifecycle and the so called "credibility assessment", which is a set of V&V activities concerning each lifecycle stage. KLEIJNEN (1975) focuses on different techniques for the elaboration and statistical analysis of the experimental design, w.r.t. simulation experiments. In medicine, BURTON *et al* (2006) present a checklist emphasizing relevant issues for elaborating SBS protocols. RAHMANDAD and STERMAN (2012) published a set of reporting guidelines for SBS in social science research. In their set, they concern with three main aspects: model visualization for diagrams, model description for equations and algorithms, and simulation experiments design, including random numbers and optimization heuristics. Overall, there is a concern with the model validity and adequate experimental design.

3.3 Reporting guidelines for simulation studies in Software Engineering

According to RUNESON and HÖST (2009), ideally the reader should not distinguish the experimental study from its report. For that, the study planning and execution as well as the involved decisions should be made explicit in the report. In this section, we present an overview of the proposed reporting guidelines as in Table 3-1. The full description of such reporting guidelines is available at (DE FRANÇA and TRAVASSOS, 2014a) encompassing only the reporting perspective after all the evaluations on this set had been made (presented in Chapter 4).

Study reports are a way of communicating research findings. Hence, authors should consider the target audience, as well as theoretical foundation and terminology should be selected accordingly, enabling the full understanding of contributions. In addition, email or other contact data should be provided to allow the readers to possibly ask for further information or details regarding the study. Finally, this set of guidelines is organized in chained sections and this organization implicitly suggests a possible report organization structure.

Furthermore, it is important to highlight that each guideline should not be taken only by the recommendation statement, but also by the associated discussion and examples. Both discussion and examples often bring the perspective of Simulation-Based Studies and the SE research area. The next subsections briefly discuss the main aspects involved in the proposed guidelines.

The foundation for such reporting guidelines is a combination of evidence obtained in the technical literature (the outcomes from the literature review presented in Chapter 2 and related guidelines from Sections 3.1 and 3.2) and reasoning regarding simulation in SE. In other words, it is not only a compilation of current state of the art, but also comprehends analysis of concepts and practices inside and outside SE engineering to consolidate common and useful knowledge to be reported in SBS.

The sources come, in most cases (SG3-6, SG9-17, and SG19), from the qSLR analysis (DE FRANÇA and TRAVASSOS, 2013b). In this opportunity, we identified lack of information in the outcomes from the review by observing and analyzing empty fields in the information extraction sheet. The lack of such information hampered the understanding and analysis of how SBS have been conducted as well as their quality. Therefore, recurrent unreported information affecting both understanding and quality of the studies derived guidelines on what should be reported.

Table 3-1. Simulation Reporting Guidelines Overview (DE FRANÇA and TRAVASSOS, 2015).

ID	Guideline Statement	Ref
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	The context where the simulation study is taking place should be described in full.	ABCD
SG3	Explicitly state the problem that motivates 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 the 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 (HÖST and RUNESON, 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)		

The evaluation based on the approach proposed by KITCHENHAM *et al* (2008), as it will be presented in Section 4.2, encouraged the adding of general reporting guidelines (SG1-2, SG8, SG18 and SG20-22). The simulation feasibility guideline (SG7) was extended from (BALCI, 1990) however including technical aspects, since it originally discusses costs, schedule and resources for a simulation project. We also identify on Table 3-1 examples of guidelines sharing similar concerns, but not covering specifics from both simulation and SE.

3.3.1 Study Definition

As SBS, we mean studies performed in both *in virtuo* and *in silico* environments (TRAVASSOS and BARROS, 2003). *In virtuo* studies stand for human subjects interacting with a computerized (simulation) environment, while *in silico* studies stand for both subjects and environment being represented by simulation models. In both environments, the object of study always relates to the simulation model. It may relate as the phenomenon/system/process, which the model abstracts or as a model under validation. SE contextual factors may rely on supporting data used to calibrate the simulation model, in which the human nature of SE activities and the amount of unknown variables can affect the studies' results.

When implementing simulation results in real contexts, the assumed environment and pre-requisites should be guaranteed or handled in the real context. Consequently, some adjustments are required on target assets.

This contextual matching is possible only if the context description is available (SG2). DYBÅ *et al* (2012) propose the use of a broad perspective approach, describing the context in such a way the study report allows answering research questions such as "*What* technology is most effective for *whom*, performing *that* specific activity, on *that* kind of system, under *which* set of circumstances?"

The problem (SG3) should be stated and described in the same way it was observed in the described context. Problems may arise from repeated situations where the solution has a complex implementation or requires an expensive alternative, or from a specific critical situation.

Taken an adequate problem definition, goals' clear definition (SG4) is the next step, leaving no doubt as to what is to be achieved, similar to other SE studies, in which these definitions adopt the GQM approach (BASILI, 1992). Common purposes for SBS consist of developing a basic understanding (characterization) of a particular phenomenon (simulation model), finding robust or optimum decisions, or comparing the merits of several decisions. Besides, the simulation model should be able to answer research questions through its output data, and its input parameters (variables or constants) should allow the desired scenario configuration.

DAVIS, EISENHARDT and BINGHAM (2007) argue on the need for research questions by stating that without an intriguing research question, the simulation research relies on a 'fishing expedition', in which the researcher lacks focus and theoretical relevance and risks becoming overwhelmed by computational complexity. This way, once following the GQM approach to drive the goal definition, and deriving research questions (SG5), the next step is to define metrics from which the questions should be answered, since SBS are naturally quantitative or semi-quantitative. The metrics definition allows one to 'ask' the research questions as hypotheses (SG6), which should be submitted to statistical tests. The hypotheses definition reveals the assumptions regarding the relationships among the dependent and independent variables under investigation (PERRY, SIM and EASTERBROOK, 2005).

After the study definition, the feasibility of simulation as an alternative investigation approach should be evaluated (SG7). BALCI (1990) supports this kind of analysis suggesting some indicators like cost, time and benefits. Additionally, we decided to use the following issues to support this decision-making, focusing on criteria that are more technical: capacity of observing the system or phenomenon under investigation; available resources for data collection; and risk analysis of the real phenomenon.

3.3.2 Simulation Environment and Model Validity

In order to report simulation based studies, it is important to have a detailed description of the model (SG9). It comprehends the required 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. Experimental design also benefits from such information on determining values for input parameters.

Diagrams are useful for presenting the whole idea and the conceptual simulation model. Equations allow replicating the model in other simulation tools. Finally, a text description supplements and clears any doubt about the conceptual model.

Model validity should also be addressed (SG10), due to SBS validity being highly affected by the validity of the simulation model. For that reason, guideline SG10 states that the experimenter should be aware of the initiatives (previous reports and research papers) to submit the simulation model to V&V procedures. In the case where such validation evidence is absent, these procedures should be performed to ensure model validity, exposing the results as well as the decisions guiding the validation process.

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

The use of performance measures such as bias, accuracy, coverage, and confidence intervals is unusual in SE simulation studies. Such measures enable benchmarks to compare simulation models and to analyze risks assigned to SBS results. Burton *et al* present how to calculate such measures (BURTON *et al.*, 2006).

The study environment (SG17) should also be made clear when planning and reporting SBS. It comprises the simulation model itself, datasets, data analysis tools, and simulation tools/packages. Besides, the characterization of human subjects (SG11) is important as it may influence the interpretation of *in virtuo* results. Hence, the level of expertise, number of subjects per group (treatment and control, when applicable) and any relevant characteristic should be addressed in the subjects' assignment process to the experimental units.

Additionally, the training needs and its costs should be planned as well. With computerized subjects, their behavior model, configuration parameters, and process of assignment should also be considered when preparing the experimental design, if such behavior can be clearly identified in the simulation model. In addition, it is possible to have an implicit behavior for the subjects, embedded in the simulation model for *in silico* environments.

3.3.3 Experimental Design and Analysis

Experimental design issues should be considered for reporting and planning purposes (SG12) as soon as having reached a state of model understanding and validity checks. It comprises the definition of a causal model, establishing a relationship between independent (or factors) and dependent variables. In the course of the experiment execution, the design factors may be held constant or allowed to vary.

Research questions and hypotheses are the basis for the causal model definition, and it should reflect part or the whole simulation model. It is often represented by a design matrix, with interest factors and treatments for each factor. A design point or scenario is represented by every row in this matrix, defining a combination of different levels for each factor (KLEIJNEN *et al.*, 2005). Furthermore, control (baseline scenarios) and treatment groups should be defined for controlled experiments using simulation models as instruments.

The number of simulation runs (SG14) should be defined on the basis of selected scenarios and on the simulation model's deterministic or stochastic nature. 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, MARTIN and RAFFO, 2004).

The use of random variables should also be considered when using stochastic models, estimating a confidence interval from the sample size to determine the number of simulation runs or replications (LAW and KELTON, 2000). Replication is achieved by using different pseudo-random numbers (PRNs) to simulate the same scenario. In this case, the output is a time series with auto-correlated observations as they share common seeds (KLEIJNEN *et al.*, 2005).

Supporting data plays also an important role in these experiments (SG16). It supports model calibration for the generation of equations, parameters, and determination of random variables distributions. This way, procedures should be considered for data collection and quality assurance to also avoid measurement errors. Finally, one last important aspect relies on raw data publication when possible, since it may impose confidentiality issues and enough space to fit a conference or journal paper.

3.3.4 Output Analysis and Simulation Results

Simulation runs often produce large volumes of data, distributed to different output variables. In the context of SE, we observed output analysis concentrating on the use of charts, rather than statistical (hypothesis) tests or descriptive statistics (DE FRANÇA and TRAVASSOS, 2013b).

Protocols for simulation studies should contain procedures and instruments to perform the output analysis, which should be properly selected, as statistical instruments and methods have many assumptions and restrictions that should be guaranteed. In addition, evidence supporting 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 results' statistical correctness. However, simulation based studies need additional analysis, such as threats to validity, including the model and experimental design validity (SG19).

Common types of experimental validity relate to simulation model validity (DE FRANÇA and TRAVASSOS, 2014b). The SE community has discussed the threats to validity concerned with *in vitro* and *in vivo* experimentation (WÖHLIN *et al.*, 2012). However, different conditions emerge for *in silico* experiments, in which recognized threats appear in a different way, or particular threats concerning such environments affect the study's validity.

Simulation may improve *construct* and internal validity, by accurately specifying and measuring constructs and enforcing the theoretical logic through algorithms, respectively (DAVIS, EISENHARDT and BINGHAM, 2007). However, it does not avoid the occurrence of this type of threat. Moreover, external and conclusion validity should be accomplished by replicating empirical observations and applying adequate statistical tests over the model outputs.

At the end of the report, the 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. Moreover, the final discussion should include implications on the applicability (SG21) of the solution in real scenarios or practical use. Finally, the future directions (SG22) should be mentioned in the report, pointing to research challenges, maybe including hot topics and possible roadmaps for future research.

3.4 Conclusions

Apart from the qSLR results, we also analyzed existent guidelines as inputs and concluded that the aspects involving SBS needed to be discussed under a SE perspective, different from the ones already presented by the other ESE and simulation guidelines. It is justified by the issues identified in the SE studies adopting simulation as a research strategy (Chapter 2).

Some of the observed concerns from other research areas seem to be specific for issues identified in those fields. For instance, the guidelines from (BURTON *et al.*, 2006), which is actually a sort of planning guideline, have no specific concern with experimental designs, except for the number of required simulation runs and analysis with performance measures. Actually, we are not sure whether experimental design is not an issue for medicine anymore or, maybe, it has not matured enough to discuss it for now. Besides, the reporting guidelines for social sciences, which we consider to be the closest research area to SE in terms of simulation models as it often uses the SD approach for modeling (RAHMANDAD and STERMAN, 2012), focus on model description rather than its use and validity.

As the first goal is to provide a complete report for the SBS, we are primarily concerned with covering all relevant aspects from the study. As it can be seen in later chapters, such set of guidelines will be evolved with more ambitious goals and scope, requiring more specific discussions.

Finally, the proposed reporting guidelines are not closed. They still need further evaluation and improvement, as well as discussion by the SE community and feedback from the researchers regarding their application.

4 Reporting Guidelines Evaluation and Analysis

In this chapter, we present the set of evaluation and analyses performed in order to observe different characteristics of the reporting guidelines. For that matter, we selected different approaches to perform such evaluations, and each approach has contributed with insights to improve and evolve the set of guidelines.

4.1 Introduction

After the definition of a preliminary set of reporting guidelines for SBS in the context of SE (DE FRANÇA and TRAVASSOS, 2012), we investigated its completeness and correctness. For that, we decided to follow three assessment strategies, combining multiple perspectives, in the sequence they appear in the next sections.

These assessments the proposed set of reporting guidelines concern their capability to guide authors and readers on providing or identifying all the relevant information expected in simulation-based studies reports. This is what we call completeness. In other words, we are interested on assessing whether every relevant aspect is covered with at least one reporting guideline. Furthermore, these evaluations are also concerned with the reporting guidelines content in terms of theoretical background and accurately using concepts to discuss and exemplify each aspect. This is what we refer as correctness.

The following sections detail the three assessments conducted to evolve the reporting guidelines to their actual stage, as presented in Section 3.3. The planning perspective was not considered for these evaluations. Such assessments were sequentially executed and the output of each one was used as input for the subsequent one, as presented in Figure 4-1.

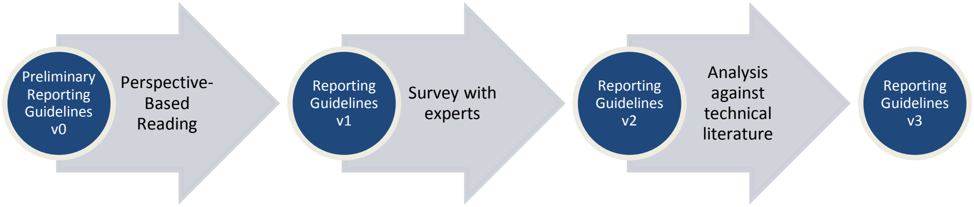

Preliminary Reporting Guidelines v0
Perspective-Based Reading
Reporting Guidelines v1
Survey with experts
Reporting Guidelines v2
Analysis against technical literature
Reporting Guidelines v3

Figure 4-1.Reporting Guidelines Evolution Assessments Pipeline (DE FRANÇA and TRAVASSOS, 2015).

4.2 Perspective-Based Reading

The first set contained 13 reporting guidelines concerned with the main issues observed in the results of the qSLR (Section 2.2), as presented in Table 4-1. Such version contains only statements and short discussions and no examples were available.

Table 4-1. Preliminary version of the reporting guidelines

Item	Description
Goals and Scope	Goals are the primary consideration to be addressed in experimental studies. Study boundaries and scope along with hypotheses should also be reported.
Model Description	It is important for the complete study understanding. It must encompass the model structure and behavior.
Model Validation	It allows reducing the threats to the study validity. All procedures performed to reach the model validity should be reported or indicated.
Simulation Scenarios	Suitable simulation scenarios selected for the study.
Subjects	Assignment of human (*in virtuo*) or computerized (*in silico*) subjects.
Experimental Design	Descriptions about the arrangement of dependent and independent variables, cause-effect relationships and combination techniques.
Number of simulation runs and criteria	Indicate the number of simulation trials and runs, including the decision criteria to reach that number.
Data support	Indication about the type of used data: Real-system or artificial. It also should describe how data was collected and evaluated. Descriptions about how calibration was accomplished using these data is recommended.
Tool Support or Simulation Package	It should report the decision criteria and important considerations considered any simulation tool used in the study.
Storage and Summary of Simulation Trials	It should explain how the storage of data produced by the simulation trials is performed, including the measures used to summarize them.
Performance evaluation	The bias, accuracy and confidence of the used simulation models should be addressed to improve the study acceptability.
Threats to validity	It should report identified limitation of the studies regarding the influence of simulation models.

From the overview presented in Table 4-1, we performed a preliminary evaluation based on the approach proposed by KITCHENHAM *et al* (2008). The main goal of this evaluation is described in Table 4-2.

Table 4-2. GQM goal definition for the review

Purpose	
Analyze	Reporting guidelines for simulation-based studies
For the purpose of	Characterize
Perspective	
With respect to	Completeness and Correctness
From the point of view of	Software Engineering Researchers
Environment	
In the following context	The authors of the reporting guidelines will perform the review, using an external approach, in the Experimental Software Engineering Lab at COPPE-UFRJ.

This approach is organized as an inspection technique using perspective based reading, comprehended by several checklists, one for each perspective. The proposed perspectives are:

- Researcher: those who aim at discovering whether the report presents relevant information for a given research area;
- Practitioner/Consultant: those who are interested on information for application at Industry and concerns with the possibility of results giving earns to organizations;
- Meta-analyst: those who extract quantitative information to be aggregated or synthesized with equivalent experimental results;
- Replicator: those who aim at replicating an experiment;
- Reviewer: those who evaluate the report for in journals or conference publications;
- Author: those who expect to use the guidelines to report their study.

Among the presented perspectives, KITCHENHAM *et al* (2008) provide checklists for the following perspectives: researcher, practitioner/consultant, meta-analyst, replicator and reviewer. However, we consider the meta-analyst perspective useful for SBS as not applicable, since this sort of study enables (by definition) multiple experimental trials. Therefore, aggregation is not relevant to increase statistical power in this context. This way, we adopted the remaining checklists.

On the original approach (KITCHENHAM *et al.*, 2008), the checklists are used by a group of researchers to review the reporting guidelines. In our review, we have no simulation experts available to perform this review, nor reviewers familiarized with the approach proposed by KITCHENHAM *et al* (2008), so the reviewer is also the author who proposed the reporting guidelines. This clearly introduces the reviewer bias. However, we opted for keeping the author to perform this first evaluation using the checklists proposed in the followed approach, with no adaptation, attempting to reduce the associated bias, since the approach has a systematic technique to conduct the evaluation through the questions in the checklists. For instance, Table 4-3 present the checklist for the researcher perspective.

Table 4-3. Checklist for the Researcher perspective [from (KITCHENHAM *et al.*, 2008)]

Number	Question	Rationale
P-1	Is the paper easy to find?	Consultants need to be able to find relevant research results
P-2	Is it a relevant paper?	Consultants should be able to identify quickly whether or not an article is relevant to their requirements
P-3	What does the paper claim?	Consultants need to identify exactly what claims the paper makes about the technology of interest
P-4	Are the conclusions/results useful?	Consultants need to know whether the conclusions/results have practical relevance
P-5	Is the claim supported by believable evidence?	Consultants need to be sure that any claims are supported by evidence
P-6	Is it clear how the current research relates to existing research topics and trends?	Consultants need to know how the current work relates to existing research trends
P-7	How can the results be used in practice?	Consultants need guidance on how the results would be used in industry
P-8	In what context is the result/claim useful/relevant?	Consultants needs to know the context in which the results are expected to be useful
P-9	Is the application type specified?	Consultants need to know what type of applications the results apply to. In particular whether they are specific to particular types of application (e.g. finance, or command and control etc.)
P-10	Is the availability of required support environment clear?	Consultants need to know whether any required tool support is available and under what conditions
P-11	Are any technology pre-requisites specified?	Consultants need to know whether there are any technological prerequisites that might limit the applicability of the results
P-12	Are the experience or training costs required by development staff defined?	Consultants need to know the training/experience requirements implicit in the approach
P-13	Is the expense involved in adopting the approach defined?	Consultants need some idea of the cost of adopting the approach, in order to perform return on investment (ROI) analyses
P-14	Are any risks associated with adoption defined?	Consultants need to know whether there are any risks associated with adoption of the technique
P-15	Do the results scale to real life?	Consultants need to be sure that the results scale to real life
P-16	Is the experiment based on concrete examples of use/application or only theoretical models?	Consultants need to be sure that the results have a clear practical application
P-17	Does the paper discuss existing technologies, in particular the technologies it supersedes and the technologies it builds on?	Consultants need to be sure that the experiment involves comparisons of appropriate technologies. They need to know that a new approach is better than other equivalent approaches not a “straw man”
P-18	Is the new approach, technique, or technology well described?	Consultants must be sure that they understand the new approach/technique/technology well enough to be able to adopt it
P-19	Does the paper make it clear who is funding the experiment and whether they have any vested interests?	Consultants need to be sure that the experiment is as objective as possible
P-20	Does the paper make it clear what commitment is required to adopt the technology?	A consultant needs to know whether adoption of an approach/technology requires a complete and radical process change or can be introduced incrementally
P-21	Are Technology Transfer issues discussed?	Consultants need to know what the objections to a new technology are likely to be, and whether there are any clear motivators or de-motivators
P-22	Is there any discussion of required further research?	Consultants need to know whether the research is complete or the approach needs further development

For each question in the checklist, it is possible to assign one of the three following values:

- Attended: there is at least one guideline answering the questions;
- Improvement Opportunity: issues related to unclear sections in the document, lack of details, lack of theoretical background, over standardization or organization for the report;
- Defect: issues related to the lack of essential content or information, non-relevant recommendations for particular situations, ambiguous sentences and incorrect concepts or reasoning.

Table 4-4 presents the results from this review, according to the number of defects and improvement opportunities, for each perspective. The preliminary set had only 13 reporting guidelines, organized in 11 sections. After this review, we proposed a new version containing 20 reporting guidelines organized in 14 sections (Table 4-5). Thus, it is possible to observe significant increase in the number of guidelines.

Table 4-4. Perspective-Based Reading Results (DE FRANÇA and TRAVASSOS, 2015)

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 for two reasons: (1) the preliminary version did not take into account general aspects often concerned in other types of studies; and (2) textual elements such as title, structured abstract and conclusions were also overlooked. However, the improvement opportunities consist in lack of details for some guidelines or even lacking discussions for some relevant aspects.

As a result of this review, we also evolved the presentation style in order to improve the understanding and clarity of their contents, as shown in the example:

Preliminary Version:
“Indication about the type of used data: Real-system or artificial. It also should describe how data was collected and evaluated. Descriptions about how calibration was accomplished using these data is recommended.”

Reviewed Version:
“The data used to support the simulation model development or SBS should be reported, whenever possible.”

It is possible to observe shorter and more objective statements for the new guidelines version. The details of the reporting guidelines are developed in a complementary text, which existed in the previous version and that evolved together, containing discussions regarding the concepts and implications associated to the reporting guidelines along with examples on how it has been observed in different situations in the technical literature.

As this review was performed after elaborating a first and stable version of the guidelines, we consider it as an evaluation. However, it could also be understood as an integrant part of the proposed guidelines development.

Table 4-5. Overview of the Reporting Guidelines v1

ID	Guideline Statement
Report Identification	
SG1	Proper title and keywords should objectively identify the study report, as well as have a structured abstract summarizing the report contents.
From Context to Research Questions	
SG2	The context where the research is taking place should be described in full.
SG3	Explicitly state the problem that motivates the study, so that research questions can be derived.
SG4	Clearly state the research goals and scope.
SG5	Present the research questions derived from established goals.
SG6	Clearly state the null and alternative hypotheses from research questions.
Background and related work	
SG7	Present only essential background knowledge and also the related works.
Simulation Model and Validation	
SG8	Describe the simulation model used in the study through its main variables, constants and the underlying simulation approach.
SG9	Present all possible evidence regarding the validity of the simulation model (conceptual and execution).
Subjects	
SG10	Characterize the subjects involved in the simulation study and report training needs.
Simulation Scenarios	
SG11	Describe the selected simulation scenarios and the procedure used to identify them as relevant.
Experimental Design	
SG12	Experimental design, including independent and dependent variables and how treatments are assigned to each factor should be reported.
SG13	The number of runs together with the rationale to determine it should be reported.
Intermediate Experimental Trial	
SG14	Describe which and how intermediate measures are stored among simulation trials to be used in the final analysis.
Supporting Data	
SG15	The data used to support the simulation model development or SBS should be reported, whenever possible.
Simulation Supporting Environment	
SG16	Describe the simulation environment, including the supporting tools, associated costs, and decision for using a specific simulation package.
Output Analysis	
SG17	Procedures and instruments for output analysis should be reported, as well as the underlying rationale.
Threats to Validity	
SG18	Always report the threats to study validity and limitations.
Conclusions and Future Works	
SG19	Main results/findings should be identified and summarized, as well as the conclusions arising from the results.
SG20	Applicability issues should be addressed in the report, considering organizational changes and associated risks.

4.3 Survey with simulation experts

The outcome from the review described in the previous section (overview in Table 4-5) had been released as a technical report (DE FRANÇA and TRAVASSOS, 2013a). Considering it a more comprehensive and self-contained set of reporting guidelines, including discussions and examples from the technical literature, we structured a collaborative review as a survey aiming at obtaining experts' opinion regarding completeness and correctness of the new set of proposed reporting guidelines as the research goal presented in Table 4-6.

Table 4-6. GQM goal definition for the survey

Purpose	
Analyze	Reporting guidelines for simulation-based studies
For the purpose of	Characterize
Perspective	
With respect to	Completeness and Correctness
From the point of view of	Researchers experienced with Simulation-Based Studies
Environment	
In the following context	External researchers from different institutions will receive the reporting guidelines to perform a collaborative review.

In this evaluation, we extended the former goal by embracing the perspective of researchers and practitioners knowledgeable in SBS, both at the industry and academia, in order to improve the reporting guidelines. The following sections present the survey definition and results.

4.3.1 Survey Definition

The survey structure follows a conference or journal review form. In order to accomplish the proposed review, we invited simulation experts, not only in the context of SE, but also on using simulation studies in different research areas. The survey was released using the web LimeSurvey tool (www.limesurvey.org), and it has been composed of five sections:

- **Presentation**: it presents the study context, research goals, and instructions to join the survey, as well as the contact information.
- **Subject's characterization**: it requires filling a four-question form concerning subject's experience in both SE and simulation.
- **Guidelines' review**: it presents the link to download the technical report containing the reporting guidelines, followed by a six-question form regarding originality and novelty, technical soundness and contribution, presentation and

readability, references to previous and related works, report strengths and weaknesses. This form is closely related to the guidelines correctness.

- **Feedback questions**: it presents a six-question form concerning the need for formalized reporting guidelines for SBS, the possible recommendation of standard content-only or including an outline, future usage, adoption by publication venues in SE, and the possibility of either missing or extra (superfluous) information.
- **Acknowledgment**: message recognizing the participation and registering the end of the study.

We adopted two approaches for the recruitment: by convenience and systematic. For the convenience approach, we used Lattes (lattes.cnpq.br) database to look for researchers in SE with background in computer simulation. Besides, we performed a similar search in the ISERN (International Software Engineering Research Network) members list. Then, we sent e-mails inviting them to participate in the study. In the systematic approach, we adopted the framework defined by DE MELLO *et al* (2014). This framework consists in a systematic approach to define adequate population and samples for SE surveys.

For the systematic approach, we adopted the ResearchGate (RG, www.researchgate.net) professional network as source of recruitment (SoR). This SoR has a meaningful constraint to send the invitations, allowing an account to send at most 20 invitations per day. The selection criteria should include researchers with background on SE and simulation. We used three accounts to enable the execution, and it took five days to invite 300 members (assumed as researchers).

We ran separate instances of the survey for each approach, using the same structure but using distinct timeframes. In both settings, the instances were available for one month long, due to the need of full technical report reading (23 pages). After each deadline, we resent the invitations and extended the deadlines to one month more. During this extra period, we received two more answers from the sample by convenience and 32 more from the systematic approach, including incomplete participations.

4.3.2 Results

The response summary (Table 4-7) for the first sample (Lattes and ISERN) comprises 10 responses, but only two completed the survey. For the second sample (systematic sample), we received 54 responses, with 13 complete answers.

Table 4-7. Response summary for Lattes and ISERN

	Lattes and ISERN		**ResearchGate**	
Invited	23	100%	300	100%
Total responses	10	43.5%	54	18%
Full responses	**2**	**8.7%**	**13**	**4.3%**
Incomplete responses	8	34.8%	41	13.7%

From a quantitative perspective, the amount of 15 completed responses for the analysis is not promising. The complexity and required effort of the involved task (read and review a technical report discussing guidelines for simulation in software engineering, containing 23 pages) may have influenced the participation rate negatively. Additionally, it is important to highlight that, except for one subject, all of them have PhD and experience on developing and using simulation models.

Having responses with similar quality from both approaches, we analyzed them as one single source. One subject also sent the reviewed technical report (pdf file) with comments by e-mail. Table 4-8 presents the quantitative results (number or responses) from both samples, Lattes/ISERN and RG. The column Total consolidates the results from both samples. Questions 1 to 6 concern the reporting guidelines review and questions 7 to 12 concern the expert’s opinion regarding the guidelines usefulness and application. Most of questions are mandatory, except from questions 5 and 6. Thus, the total number of responses for mandatory questions should be 15.

Generally, the results show this version of the reporting guidelines as comprehensive (questions 1, 2, 4, 5 and 12), understandable (question 3), useful (questions 7, 8, 9 and 10), but with improvements opportunities (questions 6 and 11).

The contributions from these responses are important to reinforce the relevance of the proposed guidelines. In other words, subjects supported the proposed guidelines mentioning the need for more experience and systematization on simulation studies in SE, as well as expressed agreement with the guidelines for different domains (e.g., e-commerce, physics). Specifically, they mentioned the guidelines are useful for young researchers, but not “out of scope” for more experienced ones, they pointed out aspects usually missing in reported studies (such as cost data, context information, underlying rationale for tool selection, and others), and they commented about the guidelines bringing principles from Experimental Statistics to the SE domain. Furthermore, the reporting guidelines’ presentation is considered concise, well written and includes well-chosen examples.

Table 4-8. Quantitative results for the survey (DE FRANÇA and TRAVASSOS, 2015).

#	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 exist?**			
	Yes	2	11	**13**
	No	0	2	**2**
10	**Empirical publication venues adoption**			
	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**

Subjects mentioned improvements opportunities like the need to emphasize the importance of supporting data, since it is critical to ensure the "health" of data. Actually, the section "Supporting Data" refers to this issue, in a reporting perspective. According to SARGENT (1999), data validity concerns appropriateness, accuracy, the amount of available data, and if all data transformations are made correctly. What can be done to ensure data validity is to develop good procedures for (1) collecting and maintaining data, (2) testing the collected data using internal consistency techniques, and (3) screening the data for outliers and determining if they are correct. Additionally, they discuss that model description is mainly influenced by the underlying simulation approach, which is already mentioned in the guidelines. One mentioned the existence of particular standards for reporting simulation models under specific approaches, for instance, System Dynamics (STERMAN, 2000) and Agent-Based Simulation (GRIMM *et al.*, 2010).

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

Concerning presentation, participants suggested a more concrete table to resume the guidelines and more examples would help their understanding, although this would unnecessarily increase the reporting guidelines' length.

From the negative aspects, some participants considered the list of references could be longer. However, the guidelines do not intend to comprehend a whole body of knowledge, but recommendations. The references include all the outcomes from the systematic review and many additional sources outside SE.

Another issue regards the reporting guidelines to resemble a reformulation of previously stated ideas or, perhaps, whether established guidelines (or even standards) outside SE, which could have simply been re-used (after re-wording), do not really exist yet. In this sense, we are aware that the proposed guidelines share common concerns with other SE (Section 3.1) and simulation (Section 3.2) reporting guidelines. These shared concerns were mainly added after the perspective-based reading (section 4.2). Nevertheless, we understand the whole set of reporting guidelines as an original perspective, discussing simulation-related aspects and their issues faced in SE studies.

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

Finally, we could not observe any theoretical or conceptual defect, or even extraneous information. The possible lack of information as mentioned before concerns mainly with the importance of valid data. It reinforces the positive research direction and the proposed guidelines soundness. However, one still may wonder if their content is obvious and, for this reason, the responses are dominantly positive. For this reason, we conducted an additional analysis (section 4.4), changing the perspective from the simulation experts to existent SBS reports obtained in the technical literature.

4.3.3 Threats to Validity

In this collaborative review, we adopted a survey strategy to release the evaluation. Besides, we presented the tasks as usual activities for researchers: reviewing of conference or journal papers. This brings internal validity to the study as the instrument and involved concepts are familiar for the audience. Still on internal validity, we faced difficulties on performing the recruitments through the RG social network. The constraints imposed by this platform caused an accidental recruitment of one same unit (subject) from both RG and convenience samples. However, this unit answered in just one survey, with no harms to the analysis.

From the conclusion validity perspective, we obtained small sample sizes, having no room for applying statistical tests or determining confidence intervals. Even though, from the qualitative perspective, the comments and contributions are worthy feedback, pointing out specific aspects that could be improved on the proposed guidelines. Furthermore, most comments reveals interest and expertise regarding the topic, which give us confidence regarding the subject's opinions.

Regarding our constructs, we captured correctness on items 2 (technical soundness) and 3 (presentation and readability), and completeness on items 4 (previous and related work), 11, and 12 from Table 4-8. Items 5 (strengths) and 6 (weakness) contribute for both constructs. The remaining items captured perception of usefulness.

Subjects' characterization and comments revealed different backgrounds. Some shared experiences of their work regarding simulation. It allowed us to identify their comments regarding the application of the proposed guidelines on their research/engineering activities. It is relevant to embrace multiple perspectives of what we are assuming as SBS and as SE issues. However, we have no ambition of generalizing from these results, since 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 more interested in qualitative data, we analyzed this study also for descriptive and interpretive validity. These types of validity concern, respectively, the researchers are not making up or distorting the collected data, as well as their inferences and conclusions. This way, we compared answers from different questions and comments, to assure consistence among the review data. For instance, we crosschecked answers for presentation and readability with comments on paper strengths and weakness. Also, we compared these last aspects to comments regarding missing or extra information.

4.4 Analysis against the technical literature

After evolving the reporting guidelines presented in the previous sections, we updated the qSLR (DE FRANÇA and TRAVASSOS, 2013b) aiming at comparing the most recent reports against the evolved set of reporting guidelines under the research goal described in Table 4-9.

Table 4-9. GQM goal definition for the analysis

Purpose	
Analyze	Reporting guidelines for simulation-based studies
For the purpose of	Evaluate
Perspective	
With respect to	Completeness and Usefulness
From the point of view of	Software Engineering Researchers
Environment	
In the following context	The authors of the reporting guidelines will perform the analysis against the recent technical literature in the Experimental Software Engineering Lab at COPPE-UFRJ.

For the update, we concentrate on the use of a simulation model or experimentation, excluding papers discussing just model development. Then, we added a new inclusion criterion to the research protocol, in which every paper should contain at least one simulation experiment, excluding papers that have only the simulation model proposal. Besides, we excluded the EI Compendex database from the sources, as we could not apply the same string used before. It displayed unexpected behaviors and faults. Therefore, we included the IEEE Xplore digital library as a counterpart measure.

In face of these changes, the period used to apply the search strings into the digital libraries differs. For Scopus and Web of Science, we set the period from March 2011 (year from the first round of the review) to the date of the update (November 2013). On the other hand, for IEEE Xplore, we set it until November 2013, as we did not apply the search string in this library at the first round.

The results for the application of the search strings and after the selection procedure, based on the reading of title and abstracts, are presented in Table 4-10.

Table 4-10. Results from the updated review (DE FRANÇA and TRAVASSOS, 2015).

Digital Library	***Number of Records***	***Duplicated Entries***	***Included***
Scopus	261	1	10
Web of Science	19	2	4
IEEE Xplore	172	59	6
Total	*452*	*62*	*20*

Twenty papers included were read in full and we excluded most of them, as we understood there were no experiments in those papers, 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) and (UZZAFER, 2013). Additionally, we searched for simulation studies in the main conference (ICSSP[5]) and journal (SPIP[6] including its new title JSEP), applying the same criteria from the systematic review. We considered these venues as they publish the majority of simulation issues in the SE context. We found seven other simulation studies (AL-EMRAN *et al.*, 2010), (BIRKHÖLZER, PFAHL and SCHUSTER, 2010), (BAI *et al.*, 2012), (HOUSTON and LIEU, 2010), (PAIKARI, RUHE and SOUTHEKEL, 2012), (CONCAS *et al.*, 2013) and (HOUSTON and BUETTNER, 2013) in these venues.

These eleven research papers were analyzed 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 in searching for information that could satisfy each guideline. In other words, we consider as complied when the report present the information regarding the aspects discussed in the guideline.

It is important to note that the reporting guidelines SG6, SG11, and SG15 (from Table 3-1) are out of scope for this analysis. Although we understand that these guidelines are strongly related to simulation studies, this is not entirely applicable for the studies selected for this analysis.

The SG6 relates to the establishment of hypotheses, which we understand as not being essential or necessary for characterization studies. Besides, SG11 relates to subject description, which not always applies to *in silico* studies. SG15 focuses on intermediate trials and this is not common for the simulation approaches adopted in these studies, mainly deterministic simulation. Rather, it is applied often to stochastic simulation, where many replications are executed for the same inputs. Particularly, when the simulation environment (or simulator) does not offer this kind of support.

The overall coverage for the reports studied as related to the reporting guidelines is shown in Figure 4-2.

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

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

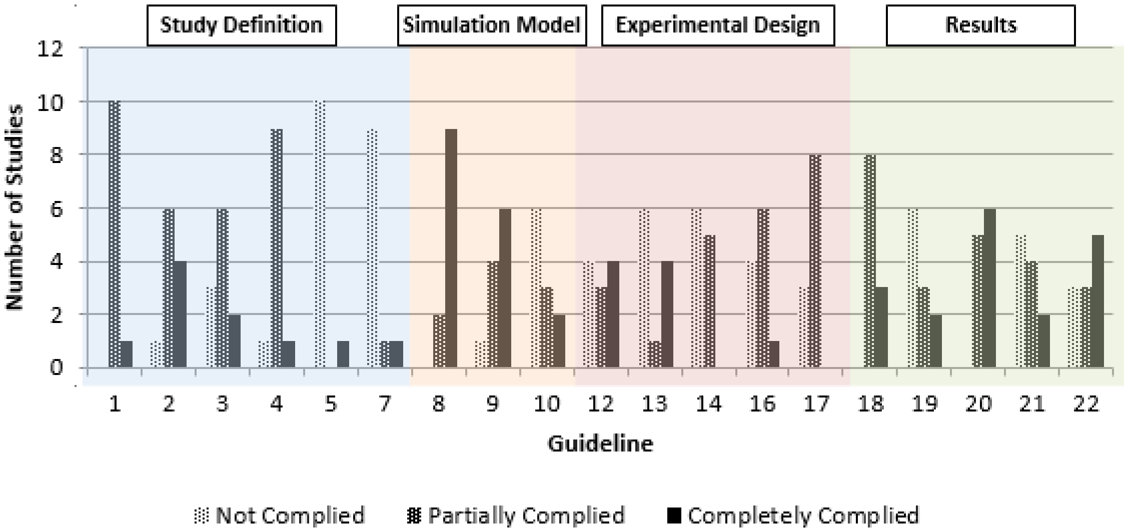


Figure 4-2. Guidelines Coverage for different research papers (DE FRANÇA and TRAVASSOS, 2015)

We roughly divided the set of guidelines into four sections for analysis purposes, in which the first one concerns the initial planning, i.e., issues related to context, problem, goals and research question definition. The second section concerns the simulation model description, as well as its foundations and validity evidence. For the third section, we concentrate on 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 Figure 4-2, it is possible to observe that all guidelines seem to be reasonable, as all of them are completely complied at least once (in one study), except for reporting guidelines SG14 and SG17 (from Table 3-1), which are respectively related to the number of runs and the simulation environment. For SG14, we could not identify the reasoning to determine the number of enough simulation runs. This aspect is important as establishing a loose number of runs may affect the output analysis effort and the confidence intervals. It is also an issue, in terms of replication, not having the complete supporting environment. Reports usually mention only the selected simulation tool, rather than statistical analysis tools, preparations for calibration, runtime environment and additional supporting technologies.

From an individual perspective, no report has mentioned the whole set of aspects covered by the reporting guidelines, as presented in Figure 4-3 and Figure 4-4, where the sequential numbers in the radars denote the simulation guidelines (SG) from the Table 3-1 and each axis in the radar can assume three possible values from the previous mentioned scale (Not Complied – 0; Partially Complied -1; and Complied - 2). Besides, there is no report presenting a homogeneous distribution of the relevant information. It means that every report focuses on one or two specific groups of reporting aspects.

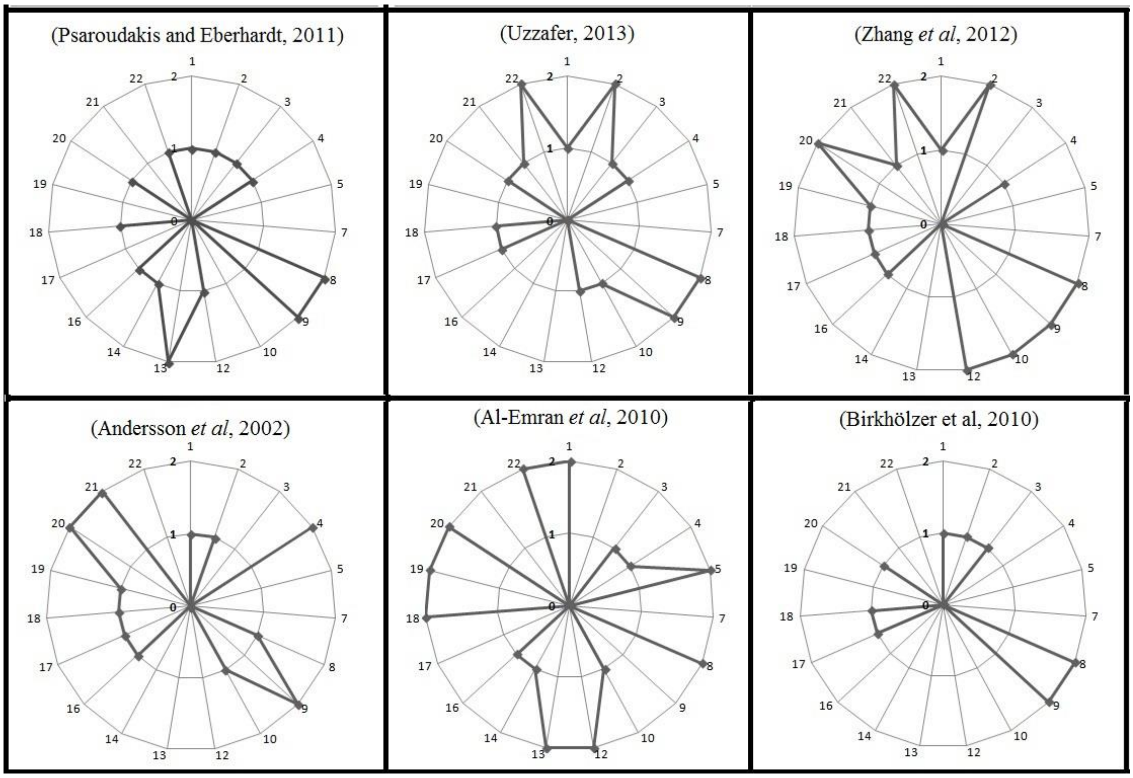
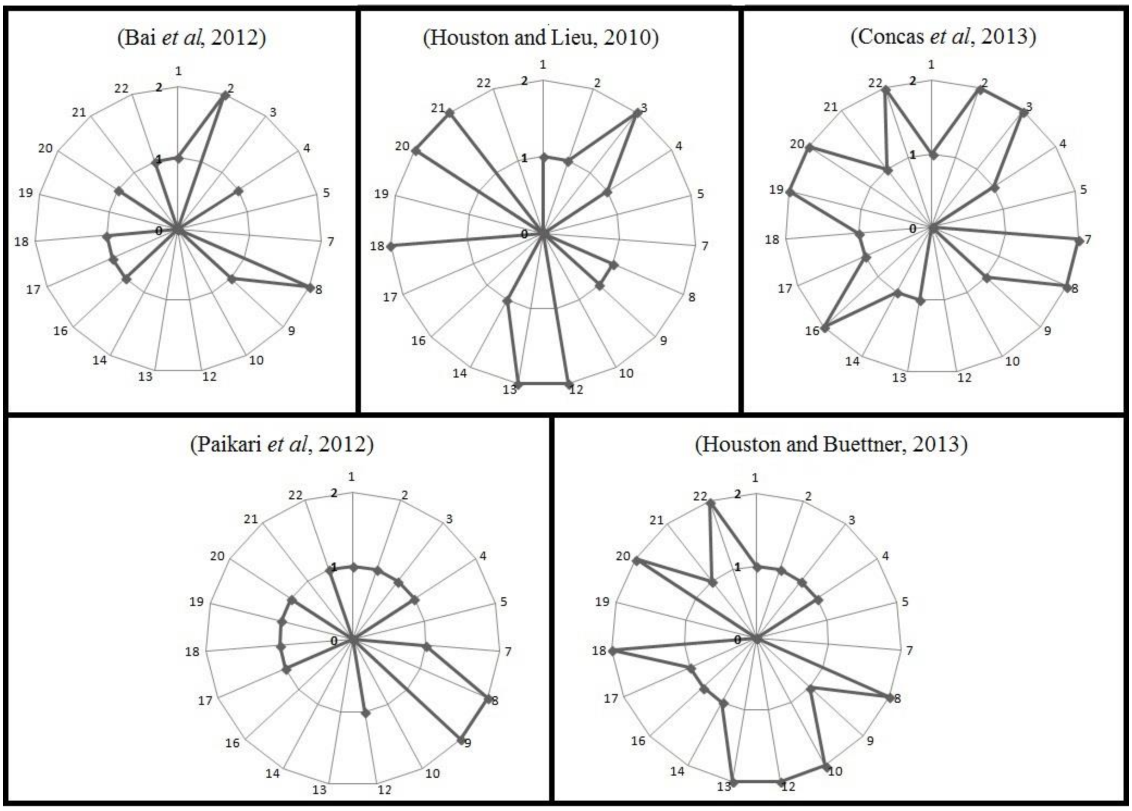


Figure 4-3. Individual profile according to reporting guidelines (DE FRANÇA and TRAVASSOS, 2015).

It is possible to observe a large variance on which kind of information the reports concentrate on, probably occurring due to the lack of a standard or recognized methodology guidelines. For instance, in Figure 4-4, the report by Houston and Buettner (2013) has a comprehensive simulation model description (SG8, SG9 and SG10) and 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 model to investigate sources of variation on deliveries and how to improve delivery quality of an agile software project, in which both customer and contractor had become concerned with the lack of predictability in deliveries, in the Aerospace Corporation.

One could suppose that this report followed an adequate simulation methodology, considering the number of guidelines that could be applied (at least partially) and the obtained results, which are discussed in terms of applicability in practice. However, there are other reports, such as (CONCAS *et al.*, 2013), where there is not enough information on the simulation model and its validity, as well as the experimental design, but it also presents an interesting analysis of the results. Indeed, for both reports, there is no explicit mention to the complete adopted methodology. Thus, it is not fair to judge the quality of the study based on the quality of the report.


(Bai *et al*, 2012)

(Houston and Lieu, 2010)

(Concas *et al*, 2013)

(Paikari *et al*, 2012)

(Houston and Buettner, 2013)

Figure 4-4. Individual profile according to reporting guidelines (DE FRANÇA and TRAVASSOS, 2015)

As general behavior, we can point out the perspectives regarding the model development and model experimentation. Usually, when a report focuses on the simulation model description, we can observe lack of information regarding the simulation experiment itself. Of course, this is a relevant aspect when considering the amount of pages available, mainly in conference papers. However, the model description alone cannot show real contributions, with no association to a simulation experiment indicating its validity and usefulness. The opposite is also possible, i.e., when the experimental design is emphasized, resulting on lack of data regarding model description.

The simulation model description, which is the most available information, should encompass at least the conceptual model, including main factors and response variables, as well as its equations. It also concerns the full understanding of the study being conducted. However, five of the eleven reports do not present one of these aspects. It not only reduces the understandability, but also compromises the possibility of replicating the study.

Regarding the model validity, V&V procedures are often mentioned, when applied, but not discussed. Authors should provide evidence regarding the execution of such procedures. For instance, the improvements regarding the issues found, the level of confidence in case of accuracy tests, or the assumptions that could be assessed and those that could not. Six out of eleven papers analyzed do not provide any information regarding model validity. Such lack of validity information compromises the study credibility and the confidence on the results.

Considering the initial planning, the lack of precision on communicating the problem under investigation is one example of issues we can identify. Motivations and what is to be solved through simulation are not sufficiently described or not described at all. It resembles works describing a solution, but with no clear problem that fits on 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 from the research questions, but this is not clear. Besides, the justifications for using simulation as an investigation approach are often neglected too. In some of the reports, it is possible to argue against the proper use of simulation for specific problems or goals, adopting other analytic methods. We generally argue that, without the information provided by these research questions, it is not possible to assess the feasibility of using simulation as an adequate strategy.

Simulation scenarios are mainly elicited *ad-hoc*, when the experimenters are not using DOE (Design of Experiments) to plan the experiment. In addition, it affects the determination of the required number of simulation runs to perform the study, which tends to be lower due to the bias in scenario selection.

The supporting data is another relevant criterion for the credibility and validity of simulation studies. Besides, as in other research strategies, the reporting of raw data is also an issue due to disclosure agreements. However, there are several ways for reporting it; for instance, using a multiplier factor to mask the data or contextual information can be given without naming organizations and people. The way such data is used to develop and validate the simulation model and how it is distributed is not presented too. Thus, without such details on model calibration and statistics applied, it is not feasible to judge whether it is adequate or not.

Another aspect overlooked is the simulation environment. We often identify indications of the selected simulation tool and rarely settings or statistical package for data analysis. However, it is not possible to reproduce studies without further information.

We could also identify some issues regarding simulation results in the selected papers. Simple comparisons of output variables are common place, but experimenters have to avoid using only this procedure. No determination of effect is provided for the input factors involved in the experimental design, when related to the output variables.

This way, there is only possibility of comparing scenarios without understanding which factors are more relevant. Apart from the outcomes report, often plotted in charts or tables, there are several discussions missing, such as threats to validity, conclusions, and the applicability of the results in the real world.

Threats to simulation studies validity are seldom discussed as such. Usually, authors refer to them as limitations and unverified assumptions, without discussing their consequences. AL-EMRAN *et al* (2010) and CONCAS *et al* (2013) present a discussion according to the types of threats proposed in (COOK, CAMPBELL and DAY, 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. Explanations should answer each research question and be grounded on the experimental design and model description. For instance, the conclusions should state how the 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, such an explanation should be reasonable as its attempts to model validation are successful and bring some confidence to the results. However, what we observed is that conclusions seem to be based on one-scenario design, without evaluating other possible interactions amongst the input factors.

4.5 Conclusions

The reporting guidelines for simulation-based studies in SE, proposed in this chapter, emerged from the analysis of the outcomes of a qSLR and evolved through three sequential evaluations.

In the first two attempts, the perspective-based and the collaborative review, the quality focus was on the guidelines' completeness and correctness. Therefore, the set of reporting guidelines evolved in this sense, by increasing the number of guidelines and discussion regarding each relevant information to be reported.

The results from the collaborative review (survey) did not change the set of guidelines meaningfully. Major modifications regard additional examples and rewording some phrases for better understanding. However, the contributions rest in the positive reviews and the fact that not having something new to be added or corrected, we assume the guidelines are complete and correct. Still, one may hypothesize whether the proposed guidelines are complete and correct just because they are obvious. Nevertheless, we showed in the analysis against the technical literature that many relevant information is still missing in recent reports. In other words, the reporting guidelines may not be obvious, but desirable.

5 From Threats to Validity to Planning Guidelines for Simulation Experiments

In this chapter, we present how the reporting guidelines were evolved to the planning perspective for simulation experiments in the context of Software Engineering. Additionally, we present threats to validity identified in the technical literature and how they could be mitigated through V&V procedures and proper experimental designs. This set of threats to validity drove the development of novel planning guidelines.

5.1 Introduction

Apart from the number of issues concerned with the reporting of SBS that led us to propose a set of reporting guidelines for such studies in SE (Section 3.3), we also observed issues regarding the methodological features. Such issues involve lack of (1) definition of research protocols for SBS, since features of the research planning are usually overlooked when performing SBS; (2) proposal and application of V&V procedures for assuring the validity of simulation models; (3) analysis and mitigation of threats to validity in SBS, which is strongly related to the validity of SE simulation models; (4) definition of criteria for quality assessment of SBS and the type of evidence we can acquire from them; (5) replication in SBS, given the absence of relevant information in the studies' reports.

These methodological issues and challenges motivated us to move forward the SBS planning needs to research protocols by starting with the study definition, i.e., the research context, problem, goals and questions. However, as we advanced, some issues on how to deal with the model validity and potential threats to validity in simulation experiments came up. Thus, with the clear needs of identifying the main and recurrent threats to validity in SBS and understanding how they can be mitigated, we primarily based our search in the previously obtained qSLR outcomes (Chapter 2). Nevertheless, as previously observed in (DE FRANÇA and TRAVASSOS, 2013b), there is no consensual terminology and authors in this field rarely discuss threats to validity using terms such as “threats to validity” or related ones. This way, we applied a systematic approach to handle the validity threats descriptions under the same perspective, adopting coding procedures, as described in the next section.

Besides the identification and classification of validity threats, we also performed an analysis regarding the model validity and experimental design to support the planning issues. The whole procedure is presented in Figure 5-1.

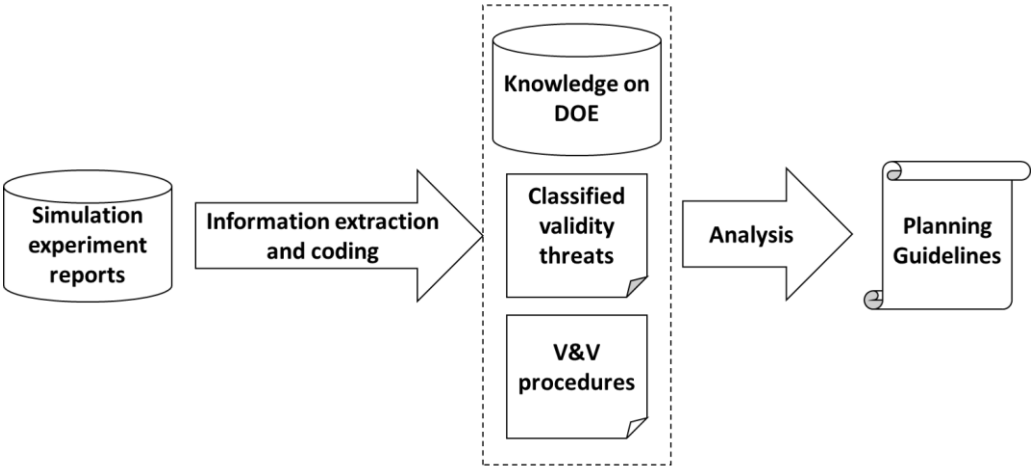

Knowledge on DOE
Simulation experiment reports
Information extraction and coding
Classified validity threats
Analysis
Planning Guidelines
V&V procedures

Figure 5-1. General procedure to achieve the planning guidelines (DE FRANÇA and TRAVASSOS, 2015).

As the classified validity threats concern both the model validity and experimental design, we performed analyses using the set of V&V procedures for simulation models acquired from the qSLR, by matching threats to validity and V&V procedures, and available knowledge on Design of Experiments (DOE), simulation (KLEIJNEN *et al.*, 2005) and classic (MONTGOMERY, 2008). The goal of both analyses is to identify whether V&V procedures and DOE can fully prevent simulation experiments from threats occurrences. As a result, the planning guidelines (Section 5.4) abstract DOE techniques and V&V procedures supporting the identification or mitigation of threats to simulation experiments validity in the context of SE. From the best of our knowledge, no work had presented threats to validity concerning SBS in the context of SE before. Consequently, it reinforces the novelty of the planning guidelines proposed in this thesis.

5.2 Qualitative Analysis

In this section, we present the secondary analysis performed over the 57 experiments captured in the qSLR by making use of the qualitative procedures borrowed from the Constant Comparison Method (CCM) (CORBIN and STRAUSS, 2008). These experiments are distributed over 43 of the 108 research papers, i.e., one paper may present more than one experiment. The aim of such analysis is to identify common threats to validity across the previously identified studies. Additionally, we performed an *ad-hoc* review to identify whether other science areas outside SE have already discussed this topic, since simulation studies in SE rarely refer to threats to validity using such terminology. In this opportunity, we identified and included in our analysis two research papers (DAVIS, EISENHARDT and BINGHAM, 2007) (ECK and LIU, 2008) discussing threats to simulation studies validity in the fields of Management Science and Criminology.

The CCM (CORBIN and STRAUSS, 2008) consists in several procedures intercalating both the data collection and analysis to generate an emerging theory from such collected data. It is important to mention we have no ambition, at this work, in generating theories, but to use the analysis procedures from CCM to support the identification of threats to simulation studies validity under the same perspective.

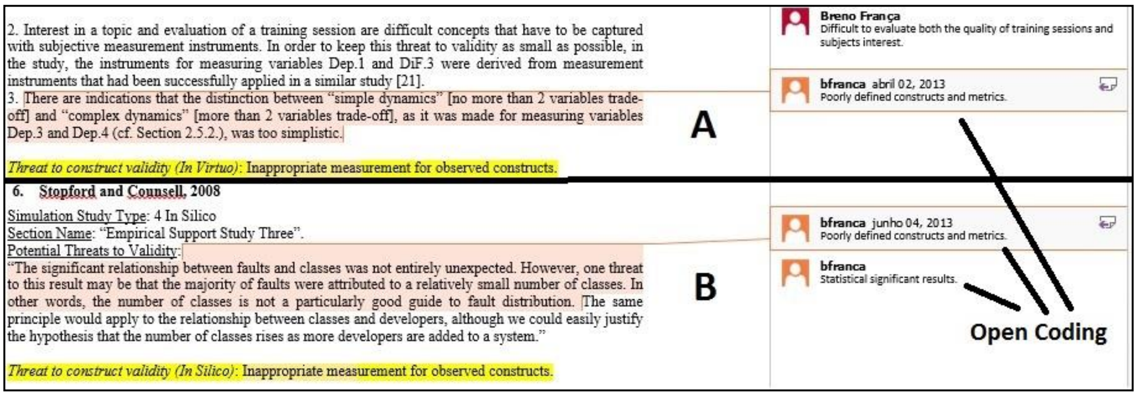
Concepts stand for the primary unit of analysis in CCM. In order to identify concepts, the researcher needs to break down the data and to assign labels to them. The researcher constantly revisits these labels to assure conceptualization consistency. Such analytic process is called *coding*, and it appears in the method in three different levels of abstraction and perspectives: *open*, *axial* and *selective coding*.

Open coding is the analytic process by which data is broken down and conceptually labeled into codes. The codes may represent actions, events, properties, and so on. It makes the researcher to rethink about the collected data under different interpretations or perspectives. In the *open coding*, the concepts are constantly compared with each other to find similarities and then grouped together to generate categories. On a higher level of abstraction, in *axial coding*, categories are associated to their subcategories and such relationships are tested against the collected data. This is also done constantly as new categories emerge. Finally, the *selective coding* consists in the unification of all categories around a central core one and other categories demanding further explanation are filled with descriptive details.

For the data collection, it was necessary to extract additional information from the studies: the study environment, whether *in virtuo* or *in silico* (TRAVASSOS and BARROS, 2003), and the validity threats description (identified as limitations, assumptions or threats to validity). The experimental environment is important since *in virtuo* contexts are supposed to be risky, due to the involvement of human subjects.

First, we extracted the threats to validity descriptions, grouping them by paper. Thirteen – out of 43 – research papers contain relevant information regarding threats to validity. For the two additional research papers, we intercalated the data collection with the analysis of the ones obtained through the qSLR. Different from the SE studies, we observed a shared consistency between the used terminology in these papers and the current terminology as presented in (WÖHLIN *et al.*, 2012), leading us to constantly review back the adopted SE terminology and search for discussions where it is possible to recognize threats to validity, limitations or assumptions.

Afterwards, we performed the initial (open) coding, assigning concepts to chunks of the extracted text, using comments in a Microsoft Word document. This way, for each new code, we compare to the other ones to understand whether it regards the same concept. In Figure 5-2, we present the example of two threats descriptions (A and B).



2. Interest in a topic and evaluation of a training session are difficult concepts that have to be captured with subjective measurement instruments. In order to keep this threat to validity as small as possible, in the study, the instruments for measuring variables Dep.1 and DiF.3 were derived from measurement instruments that had been successfully applied in a similar study [21].

3. There are indications that the distinction between "simple dynamics" [no more than 2 variables trade-off] and "complex dynamics" [more than 2 variables trade-off], as it was made for measuring variables Dep.3 and Dep.4 (cf. Section 2.5.2.), was too simplistic.

Threat to construct validity (In Virtuo): Inappropriate measurement for observed constructs.

6. Stopford and Counsell, 2008

Simulation Study Type: 4 In Silico
Section Name: "Empirical Support Study Three".
Potential Threats to Validity:
"The significant relationship between faults and classes was not entirely unexpected. However, one threat to this result may be that the majority of faults were attributed to a relatively small number of classes. In other words, the number of classes is not a particularly good guide to fault distribution. The same principle would apply to the relationship between classes and developers, although we could easily justify the hypothesis that the number of classes rises as more developers are added to a system."

Threat to construct validity (In Silico): Inappropriate measurement for observed constructs.

Breno França
Difficult to evaluate both the quality of training sessions and subjects interest.

bfranca abril 02, 2013
Poorly defined constructs and metrics.

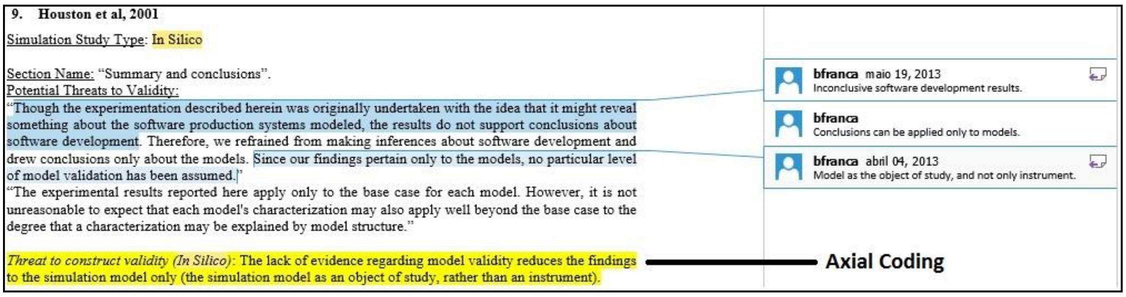
bfranca junho 04, 2013
Poorly defined constructs and metrics.

bfranca
Statistical significant results.

Figure 5-2. Open coding example, including repeated codes.

In the right side of Figure 5-2, the codes are assigned to chunks of text describing relevant threats aspects. Both descriptions share the code "Poorly defined constructs and metrics". This code lead to a threat defined in the axial code (highlighted text bellow the text description). This part of the analysis concerns with the surrogate measures defined for the interested constructs, which do not really represent the concepts under investigation.

Furthermore, we reviewed the codes and established relationships among them, through reasoning about the threat description, to generate the categories. The categories stand for the threats to validity. For instance, in Figure 5-3, we present an example of an emerged code from the interaction of three other codes.



9. Houston et al, 2001

Simulation Study Type: In Silico

Section Name: "Summary and conclusions".
Potential Threats to Validity:
"Though the experimentation described herein was originally undertaken with the idea that it might reveal something about the software production systems modeled, the results do not support conclusions about software development. Therefore, we refrained from making inferences about software development and drew conclusions only about the models. Since our findings pertain only to the models, no particular level of model validation has been assumed."
"The experimental results reported here apply only to the base case for each model. However, it is not unreasonable to expect that each model's characterization may also apply well beyond the base case to the degree that a characterization may be explained by model structure."

Threat to construct validity (In Silico): The lack of evidence regarding model validity reduces the findings to the simulation model only (the simulation model as an object of study, rather than an instrument).

bfranca maio 19, 2013
Inconclusive software development results.

bfranca
Conclusions can be applied only to models.

bfranca abril 04, 2013
Model as the object of study, and not only instrument.

Figure 5-3. Example of axial coding.

In Figure 5-3, the inconclusive results regarding software development and the use of the model as object of study limit the results to the model by itself, not allowing extrapolating behaviors from the model to explain the real phenomena. It shows one of the implications of not having information regarding the model validity.

Finally, we grouped the open codes into four major categories (axial coding), namely conclusion, internal, construct and external validity, based on the classification of threats to experimental validity presented by WÖHLIN *et al* (2012), but that could be extended in case of necessity. No selective coding was performed, since the main goal was to identify and categorize the threats to validity.

This way, the main result of this secondary analysis is a list containing 28 potential threats to SBS validity, labeled using the codes and organized according the classification proposed by Cook and Campbell, as presented in (WÖHLIN *et al.*, 2012).

5.3 Threats to Validity

The following subsections present the identified threats to validity according to the classification presented in (WÖHLIN *et al.*, 2012). The title (in bold) for each validity threat reflects the codes (categories) generated in the qualitative analysis. It is important to mention that we did not analyzed threats to validity for each one of the selected studies, even being possible to observe other potential threats to validity in these studies. However, we decided not to judge them based only in the report and therefore extracted those reported ones. For the sake of avoiding repeating threats already discussed in other ESE forums, since those threats in principle can be observed in any other sort of SE study, we concentrate on threats concerned with *in virtuo* and *in silico* experiments and not discussed in the SE technical literature yet.

It is possible to distribute the 28 identified threats to validity into the following subsets: conclusion validity (four), internal validity (ten), construct validity (ten) and external validity (four). The SE technical literature has already discussed most of the identified threats to validity regarding *in virtuo* studies, which strongly relates to the presence of uncontrolled factors of human behavior, typically addressed in internal validity issues. Conversely, threats to *in silico* experiments concentrate more on construct validity. This way, one may be tempted to point out this perspective as more critical. However, other threats can be more severe depending on the simulation goals.

5.3.1 Conclusion Validity

This type of experimental validity refers to the statistical confirmation (significance) of a relationship between the treatment and the outcome, in order to draw correct conclusions regarding such relations. Threats to conclusion validity involve the use of inappropriate instruments and assumptions to perform the simulation output analysis, such as wrong statistical tests, number of required scenarios and runs, independence between factors, among others. For instance, stochastic simulations always deal with pseudo-random components representing uncertainty of elements or behaviors of the real world. Therefore, experimenters need to verify whether the model is able to reproduce such behavior across and within simulation scenarios due to the actual model configuration or caused by internal and natural variation. The main threats to conclusion validity identified in SBS are:

- **Considering only one observation when dealing with stochastic simulation**, **rather than central tendency and dispersion measures** (ECK and LIU, 2008): we observed it into *in silico* context, where the whole experiment happens into the computer environment: the simulation model. It involves the use of a single run or measure to draw conclusions about a stochastic behavior. Given such nature, it has some intrinsic variation that may bias the results if not properly analyzed. We present an example of this threat, where ECK and LIU (2008) say, "*If the simulation contains a stochastic process, then the outcome of each run is a single realization of a distribution of outcomes for one set of parameter values. Consequently, a single outcome could reflect the stochastic process, rather than the theoretical processes under study. To be sure that the outcome observed is due to the process, descriptive statistics are used to show the central tendency and dispersion of many runs*".
- **Not using statistics when comparing simulated to empirical distributions** (ECK and LIU, 2008): also observed into the *in silico* context, this threat involves the use of inappropriate procedures for output analysis. It should be avoided comparing single simulated values to empirical outcomes. It is recommended to use proper statistical tests or measures to compare distributions with a certain level of confidence.

We also observed other threats to conclusion validity regarding *in virtuo* environments, for instance, small population sample hampering the application of statistical tests (PFAHL, KLEMM and RUHE, 2001), which is similar to the one mentioned by WÖHLIN *et al* (2012) as "Low statistical power". Besides, we identified the uneven outcome distribution (high variance) due to purely random subjects' assignment (PFAHL, KLEMM and RUHE, 2001) (PFAHL *et al.*, 2003), which is mentioned in (WÖHLIN *et al.*, 2012) as "Random heterogeneity of subjects".

5.3.2 Internal Validity

This type of experimental validity refers to the assurance that the treatment causes the outcome, rather than any uncontrolled external factor, i.e., avoid the indication of false relationships between treatment and outcome when there is none. As the experimental setting in SBS often relies on different input parameters configurations, the uncontrolled factors may be unreliable data, human subjects manipulating the model when performing *in virtuo* experiments or bias introduction by the simulation model itself. Events or situations that may impose threats in these inputs are to skip data collection procedures or to aggregate different context data, not giving an adequate training for subjects or lacking knowledge regarding the simulated phenomenon, and lack of explanation for the phenomenon occurrence, respectively. Thus, the main internal validity identified threats in SBS are:

- **Inappropriate experimental design (missing factors)** (PFAHL, KLEMM and RUHE, 2001) (PFAHL *et al.*, 2003) (PFAHL *et al.*, 2004) (RODRÍGUEZ *et al.*, 2006)**:** apart from disturbing factors, the experimental design plays an important role on the definition of which variables (both *in virtuo* and *in silico* experiments) are relevant to answer the research questions. We observed this threat occurring only into the *in virtuo* context, all of them from replications of the same research protocol, regarding to unexpected factors related to human subjects manipulating the simulation models, such as learning experience provided by manipulating simulation models and observing results. It is not common to miss factors at *in silico* environments, especially when simulation models are limited in number of input parameters. However, it is important to be cautious when dropping out factors to simplify the experimental design, as in fractional factorial designs.
- **Simulation model simplifications (assumptions) forcing the desired outcomes** (ABDEL-HAMID, 1988) (THELIN *et al.*, 2004) (MELIS *et al.*, 2006) (TURNU *et al.*, 2006) (ECK and LIU, 2008) (GAROUSI, KHOSROVIAN and PFAHL, 2009): this is the most recurrent threat reported in the analyzed papers. Always identified into the *in silico* context, it concerns with the simulation model itself. In this threat, the simulation model contains assumptions implemented in a way that they directly influence the response variables. Either establishing (coding) the intended behavior or hypotheses as truth directly from the input to output variables, or giving no chance to alternative results to occur. For instance, in one of the six studies we observed this threat (reported as an assumption) as the authors (TURNU *et al.*, 2006) say

"*In order to introduce the Test-First Development practice into the FLOSS simulation model, we make the following assumptions: (1) The average time needed to write a line of production code increases; (2) The number of defects injected during coding activities decreases; (3) The debugging time to fix a single defect decreases*".

In this case, it is possible to observe that the hypotheses (or beliefs) that Test-First Development productivity for coding decreases, the quality increases, and the maintenance time decreases are directly introduced in the model as assumptions. It goes in the opposite direction of SBS, where there is a theory with defined mechanism explaining a phenomenon, i.e., how the interactions among variables occur. In such case, there is no room for simulation, since the outcomes are predictable without running simulations. Such a black box (without mechanisms) approach is the typical configuration where *in vitro* experiments are more suitable.

- **Different datasets (context) for model calibration and experimentation** (ALVAREZ and CRISTIAN, 1997)**:** it is difficult to realize how external or disturbing factors may influence a controlled computer environment (*in silico*). Nevertheless, the supporting dataset, often required by the simulation models, may disturb the results whether data from different contexts have been compared. This is the case when calibrating the simulation model with a specific dataset, reflecting the context of a particular project, product, or organization and using the same calibration to run experiments for another (different) context. For instance, try to use cross-company data to simulate the behavior of a specific company.

We also observed other seven threats to internal validity, regarding *in virtuo* studies (PFAHL, KLEMM and RUHE, 2001) (PFAHL *et al.*, 2003) (PFAHL *et al.*, 2004) (RODRÍGUEZ *et al.*, 2006), similar to the ones already mentioned in (WÖHLIN *et al.*, 2012). It is the case of lack of SE knowledge hiding possible implications due to unknown disturbing factors, insufficient time to subjects' familiarization with the simulation tool and premature stage of the simulation tool (instrumentation effect). Also, non-random subjects' dropout after the treatment application (mortality), different number of simulation scenarios (instruments) for each treatment and available time to their performing, maturation effect by the application of same test both before and after treatments and different level of expertise required by the instruments for both control and treatments groups (instrumentation effect).

5.3.3 Construct Validity

This type of experimental validity refers to assuring the experimental setting (simulation model variables) correctly represents the theoretical concepts (constructs), mostly observed into the *in silico* context, where the simulation model plays the main role. Threats to construct validity may occur due to the lack of model variables exactness and relationships definition (and their respective equations), representing human properties, software products or processes, so the collected measures do not actually represent the desired characteristics. DAVIS, EISENHARDT and BINGHAM (2007) claim the nature of simulation models tends to improve the construct validity, since it requires formally defined constructs (and their measurement) and algorithmic logic representation for the theoretical mechanism, which explains the phenomenon under investigation. However, we could observe some threats to construct validity into the context of SBS, which are:

- **Naturally different treatments (unfair) comparison** (PFAHL, KLEMM and RUHE, 2001) (PFAHL *et al.*, 2003) (PFAHL *et al.*, 2004) (RODRÍGUEZ *et al.*, 2006): this happens when comparing simulation models to any other kind of model not only in terms of their output variables, but also in nature, like analytic models. We observed this threat occurring only in the *in virtuo* context, all of them from replications of the same research protocol.
- **Inappropriate application of simulation** (PFAHL, KLEMM and RUHE, 2001) (PFAHL *et al.*, 2003) (PFAHL *et al.*, 2004) (RODRÍGUEZ *et al.*, 2006): in the *in virtuo* context, it is possible to identify situations where to build the model can be more effective than to its use, considering that SBS involves both stages. It is the case when the learning level is the response variable and subjects have contact with model development issues and understand all details regarding the abstraction of the phenomenon or behavior. We observed this threat occurring only in the *in virtuo* context, all of them on replications of the same research protocol.
- **Inappropriate cause-effect relationships definition** (GAROUSI, KHOSROVIAN and PFAHL, 2009): this threat regards the proper implementation of the causal relationships between the simulation model constructs explaining the mechanisms under investigation.
- **Inappropriate real-world representation by the model parameters** (GAROUSI, KHOSROVIAN and PFAHL, 2009): the choice of values for the input parameters should reflect real-world scenarios, assuming suitable values that can be observed in practice and are worthy for the analysis.

- **Inappropriate model calibration data and procedure** (GAROUSI, KHOSROVIAN and PFAHL, 2009): it involves, as the previous one, data used to perform the study, mainly to instantiate the simulation model, i.e., to calibrate the model using data from the corresponding real world. It may cause unrealistic distributions or equations, scaling the effects up or down.
- **Hidden underlying model assumptions** (GAROUSI, KHOSROVIAN and PFAHL, 2009): if assumptions are not explicit in the model description, results may be misinterpreted or bias the conclusions, and may not be possible to judge at what extent they correspond to the actual phenomena.
- **Invalid assumptions regarding the model concepts** (STOPFORD and COUNSELL, 2008): it regards the validity of the assumptions made in the model development. Once they are invalid, the conclusions may also be corrupted. Every assumption made on a simulation model should be checked later. To make assumptions facilitate model development by reducing the model complexity and scope, but may also impose not observable conditions in the real application context.
- **The simulation model does not capture the corresponding real world building blocks and elements** (GAROUSI, KHOSROVIAN and PFAHL, 2009): it concerns with the model compliance with real world constructs and phenomenon representation. If there is no evidence of theoretical mechanism's face validity, it is possible that the simulation model has been producing right outcomes through wrong explanations.
- **The lack of evidence regarding model validity reduces the findings only to the simulation model** (HOUSTON *et al.*, 2001): it regards to SBS where a simulation model is chosen without proper information about its validity. Therefore, no conclusion can be draw about the abstracted phenomenon, but only about the model itself. Hence, the simulation model plays the role of an object of study, rather than an instrument. As an example, HOUSTON *et al* (2001) say: "*Though the experimentation described herein was originally undertaken with the idea that it might reveal something about the software production systems modeled, the results do not support conclusions about software development* [inconclusive results]. *Therefore, we refrained from making inferences about software development and drew conclusions only about the models. Since our findings pertain only to the models, no particular level of model validation has been assumed* [lack of validity evidence]."

We can also identify inappropriate measurements for observed constructs in SBS (STOPFORD and COUNSELL, 2008). WÖHLIN *et al* (2012) has already reported it as “inadequate preoperational explication of constructs”, and it was the only threat observed in both *in virtuo* and *in silico* contexts.

5.3.4 External Validity

This type of experimental validity involves the possibility of generalization of results outside the experimental settings’ scope. In simulation studies, it is particularly interesting to know if different simulation studies can reproduce similar results, called simulated external validity (ECK and LIU, 2008) or whether they can predict real-world results, called empirical external validity (ECK and LIU, 2008). For instance, a software process simulation model not being able to reproduce the results observed in one organization or not being able to obtain consistent results across different calibration datasets. Thus, the five identified (all concerned with *in silico* studies) threats to external validity are:

- **Simulation results are context-dependent, since there is a need for calibration** (GAROUSI, KHOSROVIAN and PFAHL, 2009): simulation modeling involves the definition of both conceptual and executable models. Therefore, to run simulations, the model needs to be calibrated using data representing the context in which the experimenter will draw conclusions. Results are as general as the supporting data. In other words, simulation results are only applicable to the specific organization, project, or product data.
- **Simulation may not be generalizable to other same phenomena simulations** (ECK and LIU, 2008): this threat refers to the emulation of a theoretical mechanism across different simulations. Such simulations may differ in terms of calibration and input parameters, but the results are only generalizable if they appear similar in different settings. In other words, the mechanism has to explain the phenomenon under different configurations to achieve such external validity.
- **Simulation results differ from the outcomes of empirical observations** (ECK and LIU, 2008) (GAROUSI, KHOSROVIAN and PFAHL, 2009): when simulation outcomes differ sufficiently from empirical outcomes, we may say that simulated results have no external validity. One example of such threat in (GAROUSI, KHOSROVIAN and PFAHL, 2009): “First, the results are only partly consistent with empirical evidence about the effects of performing V&V activities. While code quality can always be improved by adding V&V activities, it is not always true that adding V&V activities in earlier development is better than adding them in later phases”.

- **Simulation model not based on empirical evidence** (DAVIS, EISENHARDT and BINGHAM, 2007) (RAHMANDAD and WEISS, 2009): if the model constructs and propositions are all conjectural, i.e., with no ground in field studies or empirical experiments, integrally or partially, it is very important to invest effort on validation procedures, since the model itself cannot show any external validity (DAVIS, EISENHARDT and BINGHAM, 2007).

5.4 Planning Guidelines

Many of the reporting guidelines (section 3.3) are also useful on the planning perspective. In the sense, 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 the model description (SG9), reflecting the full grasp of the observation instrument. However, five guidelines (SG1, SG8, SG20, SG21, and SG22) exclusively focus on reporting aspects and do not contribute to the planning issues. The remaining aspects covered in the reporting guidelines are also important for planning activities, including the definition of valid scenarios and experimental design, aiming at avoiding potential threats to validity. Already knowing these potential threats in advance, we needed to identify approaches to handle them.

The approaches selected to handle the validity threats presented in the previous section include V&V procedures and DOE techniques. The main reason for selecting these approaches is that the identified threats concern with model validity and experimental design issues. The V&V procedures concern mainly with threats to construct and internal validities. The explanation for this assumption is that the successful application of V&V procedures to simulation models allows identifying and removing defects regarding model constructs, assumptions and theoretical logic, enabling the mitigation of potential threats in advance and, consequently, taking more confidence to the simulation experiment results. Additionally, DOE techniques can support the arranging of factors, as well as the investigation of effects under multiple runs.

The result of analyzing how the V&V procedures and DOE techniques could mitigate the validity threats presented in Section 5.3 are presented as suggestions or the planning guidelines for simulation experiments in SE, as summarized in Table 5-1. They are presented in the order we analyzed them, grouped according to their concern. Hence, their order has no relevant meaning.

Mainly, the planning of a simulation experiment concerns with the setting up of the experiment so one is capable of making correct and reliable inferences from the outcomes. For that matter, all decision-making should consider the potential validity threats on performing the experiment in one way or another. This is what the guideline SG32 states in some sense.

Table 5-1. Planning guidelines overview (DE FRANÇA and TRAVASSOS, 2015).

ID	Guideline Statement
Model Validity	
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.
SG24	Support model (causal) relationships, as much as possible, with empirical evidence to reinforce their validity and draw more reliable conclusions.
SG25	Always verify model assumptions, so that the results of simulated experiments can become more reliable.
Experimental Design	
SG26	Use results from Sensitivity Analysis to select valid parameter settings when running simulation experiments, rather than model 'fishing'.
SG27	Consider to use as factors (and levels) besides 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.
SG28	When dealing with simulation models containing stochastic components, determine the number of runs needed for each scenario, to capture phenomenon variance.
Data Collection and Use	
SG29	Keep track of qualitative data along with quantitative data. It is also important to record data contextual information.
SG30	Make sure that both calibration and experiment datasets came from the same population.
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.
Threats to Validity	
SG32	Consider checking for threats to the simulation study validity before running the experiment and analysing output data to avoid bias.
SG33	Be aware of data validity when comparing actual and simulated results: compared data should come from the same or similar measurement contexts.

Primarily, changing the arrangement of factors and levels, the number of simulation runs/trials or the output analysis instruments affect the results or how they can be observed. Moreover, changing these experimental properties after knowing the outcomes introduce bias, triggering a search for the desired results due to methodological problems.

Simulation results can be severely questioned if no evidence regarding the simulation model validity is given. Among several existent V&V procedures, Face Validity is one that can support the identification of threats to construct and internal validity. To expose the simulation model and its behavior to experts enables the identification of inappropriate constructs or relationships definition, as well as the possibility of reaching desired results by an unrealistic configuration of input parameters.

Model assumptions are also risky factors if not verified and may also affect internal and construct validity. As simulation models are abstractions over a system/process/phenomenon, the modeler assumes some constraints in the model validity. In other words, the modeled behavior is valid under certain conditions and scope. Those conditions need to be clear and explicit during all SBS, so they can be verified against the supporting data and results.

On the other side, external validity can be achieved by supporting the model's mechanism or causal relationships with empirical evidence. The existence of other experimental studies showing the pertinence of cause-effect relationships between model's variables reinforces that the outcomes are generated through a sound set of interactions among variables.

All these verification and validation may (and should) be performed before designing the simulation experiments that will answer the research questions. However, the validation process also involves running experiments (also called validation experiments), which should be properly designed as well. For that, the Statistics discipline called DOE offers several techniques that can support how to arrange factors and treatments in order to accomplish the research goals.

Techniques like Sensitivity Analysis can be useful to determine the input parameters, by performing small and large variations on the interest factors to understand how they influence the outputs. There are several designs for Sensitivity Analysis available in the simulation technical literature (KLEIJNEN, 2005). For some simulation models, the input parameters are not enough to design the simulation experiment and accomplish the research goals. In these cases, non-trivial factors like datasets for calibration and different versions of the same model may be considered.

Still on experimental design, stochastic models require several simulation runs/trials for the same input configuration given the internal variation caused by pseudo-random variables. Therefore, to determine the adequate number of runs that can reveal confident results requires the analysis of how close the results are from the expected variance. LAW and KELTON (2000) present a procedure on how to perform such analysis.

Apart from model validity and experimental design, the experimenter should also take care of the supporting data. If the model needs to be calibrated, the dataset supporting it and the scenario configurations defined in the experimental design need to represent the same context. In other words, the values should be meaningful for the scope under investigation.

Data collection for simulation should not only concern with the input and calibration data. It is important to have and use qualitative data explaining or clarifying assumptions and system behavior. Qualitative data can also support the output analysis, when dealing with unexpected patterns and results.

For the output analysis, the selection of proper instruments needs to address their assumptions on how the data is distributed and organized. For instance, parametric statistical tests and methods for calculating confidence intervals assume normally distributed and homoscedastic data. Besides, comparisons involving actual against simulated data should be performed under similar contexts, assuring fair comparisons and using proper instruments to perform cross-scenario analysis, and quantifying the amount of variation in multiple runs.

5.5 Conclusions

In this chapter, we presented a secondary analysis over the outcomes of the qSLR under the perspective of threats to validity. The main result from this qualitative analysis consists on the identification of threats to SBS validity. It is possible to observe that some threats are applicable only to *in virtuo* studies, due to the presence of human subjects. However, we could also identify threats to construct validity in the *in silico* context, contradicting the idea that simulation modeling improves this type of validity (DAVIS, EISENHARDT and BINGHAM, 2007). The main reason for that lies on the creative and human-intensive nature of modeling tasks, in which the researcher abstracts characteristics and behaviors of interest from his observations for the simulation model.

Additionally, we understood these identified threats should be mitigated in some sense. For that, we analyzed their main concerns and which kind of approach could be used to reduce the risks. Therefore, threats concerning with simulation model validity were associated to V&V procedures and the ones concerning with experimental design were associated to DOE techniques. This matching was performed by reasoning about the problems and the potential solutions, which derived the planning guidelines as suggestions for mitigation. However, we still need investigation on the effectiveness of these approaches to mitigate such validity threats.

6 The Evaluation of the Simulation Based Studies Planning Guidelines

In this chapter, we present a feasibility study concerned with the planning guidelines for simulation experiments in Software Engineering. This study followed a qualitative approach for the observation and analysis of how the proposed guidelines can be applied to the planning activities of a simulation experiment, in which an organization scenario and a pre-defined simulation model were given as input.

6.1 Study Protocol

The planning guidelines are based on findings from the qSLR (Section 2.2) and in the consolidated technical literature on Simulation and SE experimentation, as described in Chapter 5. They aim at guiding researchers in the earlier stages of simulation experiments, i.e., the study definition and planning, so that relevant information can be produced until the deployment of a complete report. For that, such guidelines are intended to drive the elaboration of simulation experiment plans, identifying *a priori* and eventually mitigating threats to the experiment validity, besides promoting a coherent plan, in which the planning information is logically organized by following a pre-defined structure.

As the previous evaluations (Chapter 4) covered only the reporting perspective, the planning perspective still needs external evaluation regarding their application to capture their effectiveness and perceived usefulness. The simulation guidelines under evaluation encompass both the ones sharing planning and reporting guidelines from Table 3-1 and the planning guidelines presented in Section 5.4.

6.1.1 Research Goals and Questions

The research protocol has been designed to evaluate the proposed guidelines and its main evaluation goal is presented in Table 6-1. As it is the first effort on the assessment of such planning guidelines, our purpose relies on the characterization of their use under five perspectives (effectiveness, coverage, coherence, perceive usefulness and ease of use), as described in Table 6-1.

Table 6-1. GQM goal definition

Purpose	
Analyze	Planning guidelines for simulation experiments
For the purpose of	Characterize
Perspective	
With respect to	- Effectiveness: identification of threats to validity[7] reported in the study plan; - Coverage: information contained in the plan; - Coherence: logical chain of plan sections; - Perceived usefulness: opinions whether the guidelines effectively support the plan elaboration; - Ease of use: explicitness, understanding and application.
From the point of view of	Software Engineering graduating students
Environment	
In the following context	Grad students engaged in the Experimental Software Engineering course at COPPE-UFRJ optionally using the proposed guidelines to support the planning of a simulation experiment.

The research questions, based on the goal definition from Table 6-1, are:

RQ1: Do the planning guidelines for simulation experiments enable the capacity of identifying threats to validity in the planning stage?

RQ2: Do the planning guidelines for simulation experiments promote the elaboration of a study plan containing the relevant aspects?

RQ3: Is it possible to observe logical chaining among the sections of the simulation experiment plan when using the set of guidelines for simulation planning?

RQ4: Do the planning guidelines for simulation experiments effectively support the elaboration of the study plan?

RQ5: Are the planning guidelines for simulation experiments clear, ease to understand and use?

6.1.2 Study Procedure and Instruments

This evaluation study follows a qualitative approach, in which researchers observe subjects in the elaboration of a simulation experiment plan with previously defined problem and general goal. The researchers observe the accomplishment of planning tasks through deliverables, namely the elaborated study plan for the simulation experiment, the reviews of these plans, and notes during a Focus Group (FG) session.

We selected the qualitative approach since we have a small sample of eight graduate students, hampering the use of statistical analyses, and due to the effort required

[7] Only threats to validity explicitly reported in the plans are considered, excluding any other.

to perform trials using control groups or additional treatments (for instance, other set of guidelines). Finally, we are interested in characterizing the planning guidelines application without assuming any initial hypothesis.

The execution procedure for the study is organized according to the stages sequentially described in Table 6-2. The execution starts with two classes on simulation in the context of ESE and SD approach. The first is a regular class in this course and the second one we added as part of training, on which all students are allowed to engage it. The main task in this study consists in the elaboration of a study plan for a simulation experiment in the domain of software project management.

It is important to mention that during stage 4 (step 2 in Table 6-2, plan elaboration) the subjects are free to use or not the proposed set of guidelines, since we did not assume they are all applicable and useful in the given context. Additionally, the subjects are free to use any other source of information.

The FG approach was selected instead of using feedback questionnaires for the evaluation of perceived usefulness and ease of use, such as (DAVIS, 1989). The justification concerns with the possibility of understanding the existing difficulties in the guidelines application and promoting a group discussion regarding the perceived usefulness and ease of use for the study context and possible improvement opportunities. Moreover, FG offers stimulating techniques for those feeling intimidated by interviews or feeling their opinion or experience are not relevant enough.

Besides the Consent Form (APPENDIX A) and the Subject's Characterization Form (APPENDIX B), we used as instruments:

- Slides presented in both classes, making no reference for the guidelines;
- Guidelines for simulation experiments: Technical Report containing the set of 33 guidelines involving both planning and reporting perspectives, each of them including discussions and examples from the technical literature. In this set, only 28 guidelines refer to planning issues.
- Brief specification of the simulation model proposed by (CHERNOGUZ, 2011);
- The executable simulation model, coded for the Vensim tool;
- The Vensim PLE tool, available at www.vensim.com;
- Template for the study plan: a document containing fill-in-the-blanks sections for a general purpose study plan (APPENDIX C);
- Discrepancies form for the reviews;

Table 6-2. Study execution procedure

Stage 1 – Recruitment and Subjects' Characterization
In this first stage, the students engaged to the ESE course are invited to volunteer themselves to participate in the study. Students interested in attending the study should read and sign the Informed Consent Form (APPENDIX A) and fill in the Subject's Characterization Form (APPENDIX B). This stage should happen after the Simulation Based Studies in Software Engineering classes.
Stage 2 – Training in the Simulation Environment
Each subject should receive the simulation model specification, proposed by (CHERNOGUZ, 2011), as well as one executable version of the model for the Vensim PLE. Furthermore, each subject receives the set of guidelines to support the planning of simulation experiments. Besides, each subject needs access to a computer with the Vensim PLE installed, along with the executable simulation model.
Stage 3 – Preparation
All tasks in the context of this study should be individually performed. The subjects need to read the following resources (available in the course's web portal): 1. The proposed scenario for the study: it is composed by the context description, problem definition and general goals; 2. Simulation model specification and the full paper (CHERNOGUZ, 2011); 3. Set of planning guidelines for simulation experiments in the context of SE.
Stage 4 – Execution
1. The subject should read the proposed scenario containing the problem description and the general goal for the simulation experiment to be planned, and begin the study plan elaboration (according to the available template for study plans in APPENDIX C), using both the simulation model and the set of guidelines as supporting instruments. 2. When elaborating the study plan, the subject should **identify and describe relevant aspects** for the study plan and record such information in the plan (MS Word template for the study plan). For each section of the plan, the subject needs to inform which guidelines were used to support the elaboration of that section. Thus, each subject should deliver the document containing the study plan and the references. 3. After delivering all study plans, each subject will peer review **one plan elaborated by another subject**, still using the planning guidelines to support the identification of possible issues. For this review, it is needed that the reviewer fills a discrepancy form (APPENDIX D). It is extremely important that subjects do not discuss the reviewed plans among themselves.
Stage 5 – Focus Group
After the plans elaboration and review, the subjects will engage a focus group to discuss some topics related to the study. • The groups will be organized based on subjects' characterization (level of instruction, both SE and simulation experience) and by their performance on the planning tasks. • Not having any drop out, two groups of four subjects will compose the focus group. • We defined two phases. In the first phase, we organized the groups in a role-play design using **lovers** and **haters** roles, in which lovers should argue in favor of the guidelines and the haters against them. • Ten planning guidelines were selected for the discussion. • The second phase consisted of the identification of improvements opportunities, but playing no role at this time.

6.1.3 Proposed Scenario

As input, we provided an organizational scenario (detailed scenario is described in APPENDIX D) under which the simulation experiment should be planned. In summary, it describes an environment of software development organization, engaged on its products and processes improvement, which aims at continuously investing on practices capable of returning positive results in terms of quality and cost.

The problem concerns repeatedly delays on product deliveries, causing losses on company relationship with clients and reputation, as well as financial losses. Delayed projects are characterized by the low number of experienced developers in the project beginning when compared to novices. Besides, an initial investigation, by the company, revealed that when a project reaches about 30% of its estimated progress and the project is considered to be late, managers add more work force to the project. However, such practice has not succeed in these projects. Besides, it increases the costs regarding additional workforce.

In this scenario, the subject is encouraged to give a diagnostic in a short period, explaining the reasons for the delays and additional costs and possibly presenting a feasible solution to reduce the losses in future projects.

The observed effect on the organization reminds the behavior described by Frederick Brooks (BROOKS, 1975): “adding manpower to a late software project makes it later”. Such behavior is called Brooks’ Law. Furthermore, understanding that the organization required an investigation in short time and there is some historical data available for the analysis, a feasible alternative is to conduct simulation-based experiments to understand the behavior influencing the delivery time and to test possible solutions.

The simulation model to be used as observation instrument for the phenomenon is proposed in (CHERNOGUZ, 2011). Figure 6-1 presents the causal model for the Brooks’ Law reference behavior. The description of input, intermediate and output variables are available in the APPENDIX D.

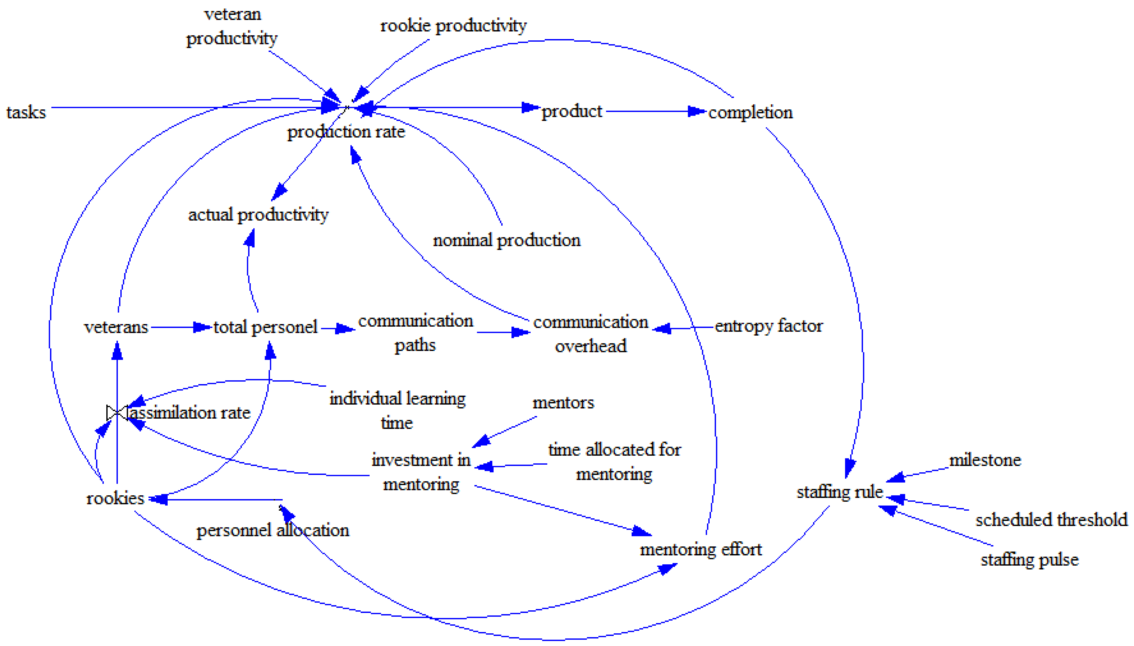
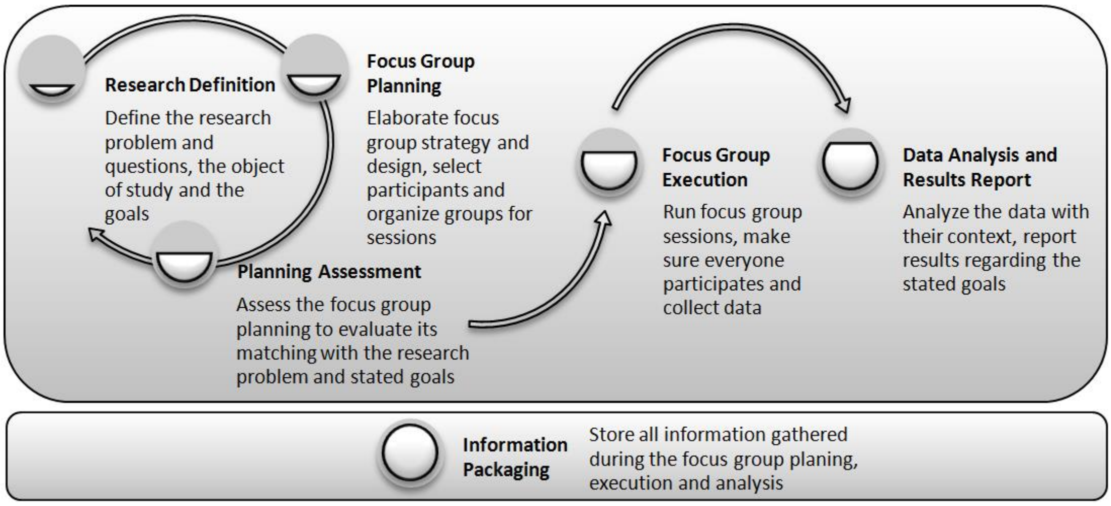


Figure 6-1. Cause-effect diagram for the Brooks' law [adapted from (CHERNOGUZ, 2011)]

6.1.4 Focus Group Methodology and Planning

Our FG methodology includes specific activities, as presented in (KONTIO, BRAGGE and LEHTOLA, 2008). Additional activities and steps were also included, since we understand the FG method not only as a group dynamics, but also as a primary study. Thus, general aspects regarding primary studies such as object of study and goals definition, as well as planning assessment and information packaging activities (MIAN *et al.*, 2004) were included in the FG process, as described in Figure 6-2.



Research Definition
Define the research problem and questions, the object of study and the goals

Focus Group Planning
Elaborate focus group strategy and design, select participants and organize groups for sessions

Planning Assessment
Assess the focus group planning to evaluate its matching with the research problem and stated goals

Focus Group Execution
Run focus group sessions, make sure everyone participates and collect data

Data Analysis and Results Report
Analyze the data with their context, report results regarding the stated goals

Information Packaging
Store all information gathered during the focus group planing, execution and analysis

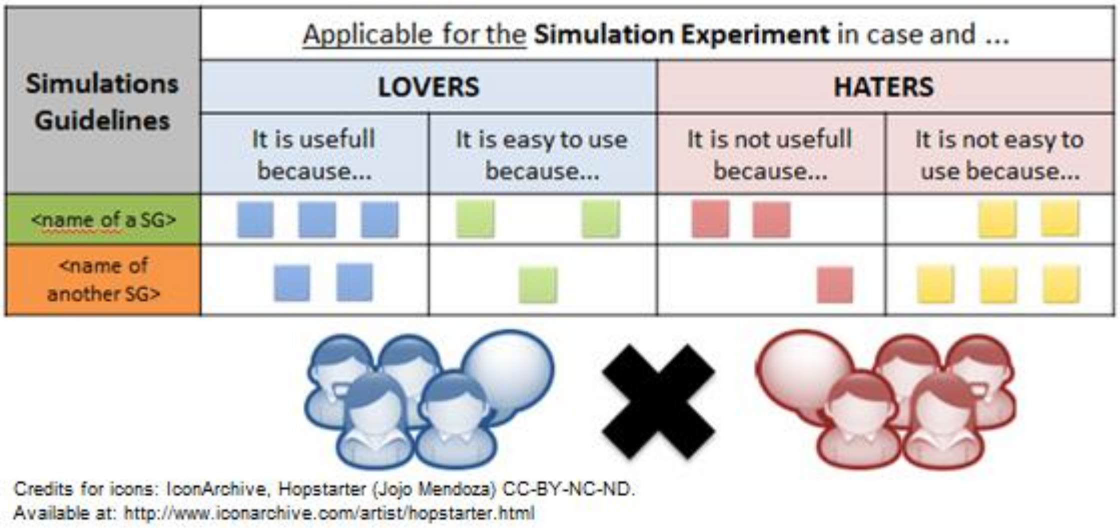
Figure 6-2. Adapted FG process (DE FRANÇA *et al.*, 2015)

The process consists in six major activities. During the research definition, the researcher identifies the research problem, questions and context, and the rationale for selecting the FG method, enabling to verify its suitability regarding the information needs and the environment for the study. After that, in planning, the researcher defines the FG strategy and design, i.e., participants and moderators, group settings and design of the participants' interactions. All planned information needs to be double-checked, so the strategy and design of the study comply with the research problem and goals. If required, these activities should be repeated until the planning is ready for the execution.

In the execution of the FG session, the moderator needs to ensure the involvement of all participants and to take notes of contributions that may help answering the research questions. Finally, the last activities include the analysis, reporting and packaging of all collected data, including the session context and potential treats to the study validity, so the researcher is able to triangulate data and make inferences about the object of study.

The FG participants' selection followed the sample from the study. Based on the subjects' characterization and experience in the planning tasks, we designed the FG in two subgroups of four subjects. Ten – out of 33 – guidelines were selected for discussion due to time constraints: the nine less used in the plans elaboration and review, and the most used one was selected as control (SG12) for analysis purpose. The control is used to start the discussions, since all subjects seem to be familiar with it. Later, we also used it to verify how the previous usage of the guideline influences the amount of discussion during the session.

The planned design organizes discussions in two stages. In the first stage, the groups are organized in a role-play design using *lovers and haters* roles, in which lovers should argue in favor of the guidelines and the haters against them. The second stage consists in identifying improvement opportunities without playing roles. Moreover, we adopted the label generation technique (COLUCCI, 2007), which uses small pieces of paper (post-its in our study) in which subjects write down their individual or consensual arguments and post them on the board. The board is divided into four sections where subjects have to use post-its of different colors (Figure 6-3) for their arguments. Two sections for subjects playing the lovers, each of them concerning one characteristic under evaluation: perceived usefulness and ease of use. Other two sections are reserved for subjects playing the haters. Each row concerns only one guideline under discussion.



Simulations Guidelines	Applicable for the **Simulation Experiment** in case and ...			
	LOVERS		**HATERS**	
	It is usefull because...	It is easy to use because...	It is not usefull because...	It is not easy to use because...
<name of a SG>	🟦 🟦 🟦	🟩 🟩	🟥 🟥	🟨 🟨
<name of another SG>	🟦 🟦	🟩	🟥	🟨 🟨 🟨

Credits for icons: IconArchive, Hopstarter (Jojo Mendoza) CC-BY-NC-ND.
Available at: http://www.iconarchive.com/artist/hopstarter.html

Figure 6-3. FG board for the study (DE FRANÇA *et al.*, 2015).

Following the design, the strategy consisted in presenting to the participants how the FG should work through slides containing the descriptions of each planning guideline under discussion at a time. Next, the subgroups (lovers and haters) discuss internally and post their arguments on the board, according to their roles. Later, the subgroups were encouraged to read their arguments and discuss them one against the other. After discussing each guideline, in the next stage, subjects posted their opinions regarding improvement opportunities w.r.t. perceived usefulness and ease of use.

In order to conduct the FG session, three researchers were involved in the session, playing different roles. One coordinator drove the discussions to keep the focus. In parallel, two additional researchers were taking notes: one responsible for collecting discussions and arguments regarding the usefulness and ease of use for each guideline, and the other researcher was responsible for capturing subjects' behaviors regarding the FG dynamics and the role-play design. Therefore, this last researcher captured behaviors like ironic arguments, laughs, change of mindset, consensus reaching, and other behaviors that could reveal how strong their arguments are in favor or not of the planning guidelines.

For data analysis, we need to take the answers and discussions in the FG according to the notes capturing the multiple perspective (post-its and notes from two observers) to analyze the accomplishment of the research goals.

Finally, an extra researcher helped on the planning assessment by reviewing the plan and checking for threats to validity.

6.1.5 Analysis

A systematic procedure needs to be defined to analyze the outputs, which involves the simulation experiment plan, the discrepancies from the review and the focus group discussions. First, such procedure should support the evaluation of the quality of data. Second, it supports the analysis regarding the outcomes to understand the effects from the guidelines application. This way, the following stages compose the analysis procedure:

1. **General evaluation of data quality**: by fully reading the simulation experiment plans, the researcher should be able to verify whether their content information are really related to the proposed scenario for the experiment and to the adopted simulation model (CHERNOGUZ, 2011).
2. **Effectiveness**: regards section related to results validity and the analysis of their correctness in terms of classification and expression of actual threats to validity in the proposed context.
3. **Coverage**: regards the subject's indications of guidelines supporting the elaboration of each section of the plan.
4. **Coherence**: regards the correct reviews in which the subjects point out the lack of logical chaining among different sections of the plan;
5. **Perceived usefulness**: regards the answers and discussions in the focus group, according to the notes capturing the multiple perspectives (post-its and notes from two observers).
6. **Ease of use**: regards the answers and discussions in the focus group, according to the notes capturing the multiple perspectives (post-its and notes from two observers).

6.2 Study Execution

After defining the study protocol, we scheduled the two simulation-related classes and carefully removed information regarding the planning guidelines and other information that could potentially bias the elaboration or review of the plans.

The first class was concerned with an overview of computer simulation concepts and how they have been applied to SE studies as an approach to support experimentation, including simulation studies taken from the SE technical literature. In the second class, we presented constructs and formalisms of System Dynamics, model examples, and the basics to use the Vensim PLE tool.

All eight students engaging the course volunteered to participate in the study, i.e., signed and filled the informed consent and the characterization form. An overview is given in Table 6-3. In summary, only one subject declared experience with discrete-event simulation. The other seven had no experience in any simulation approach. Regarding their experience with software development (in Industry and Academia), two of them were highly-experienced (participating in more than seven projects and above 5 years of experience, three with some experience (three projects and three years of experience) and three had low-experience. Regarding software process expertise, just one subject had high-experience.

Table 6-3. Subjects' Characterization.

Characteristic	Level	Number of subjects
Experience in Software Development	High	2
	Medium	3
	Low	3
Experience in Software Processes	High	1
	Medium	1
	Low	6
Experience in Simulation	High	0
	Medium	0
	Low	8

For the elaboration stage, we sent individual e-mails containing the instructions and all the instruments are available in the learning web environment (Moodle) usually used in the ESE courses. At this stage, the subjects had two weeks to work on the simulation experiment plan. After that, they sent the plans back and we distributed them to reviewers. The criteria adopted for the assignment are: (1) the reviewer could not review his own plan; (2) evenly assign subjects that used the proposed guidelines to elaborate the plan; and (3) evenly assign subjects that did not use the guidelines. This way, we sent the instructions individually by e-mail again containing also the plans to be reviewed and the discrepancies form. For that, the subjects had one week to perform the reviews. After the reviews, we sent back the discrepancies for the authors, so they could improve their plans for later execution and analysis.

Finally, we performed the FG three weeks later. The meeting took three hours and a half. Usually, FG are undertaken in a period of 3 to 4 hours to avoid participants being exhausted (KONTIO, BRAGGE and LEHTOLA, 2008). This time constraint forced us to reduce the scope of observations w.r.t. which planning guidelines should be discussed. Hence, as mentioned before, we estimated 10 guidelines to be discussed in the meeting. We firstly assumed the planning guidelines explicitly mentioned in the elaborated plans and review (Section 6.3.1) as having some indication of usefulness. This way, we selected the guidelines not used in, at least, one of the stages. In other words, we selected the less used planning guidelines from Figure 6-4, which presents an aggregated view of the guidelines (from SG2 to SG33, excluding the reporting ones) usage from both elaboration and review steps across eight (8) plans. The main reasoning is to understand why they were not used.

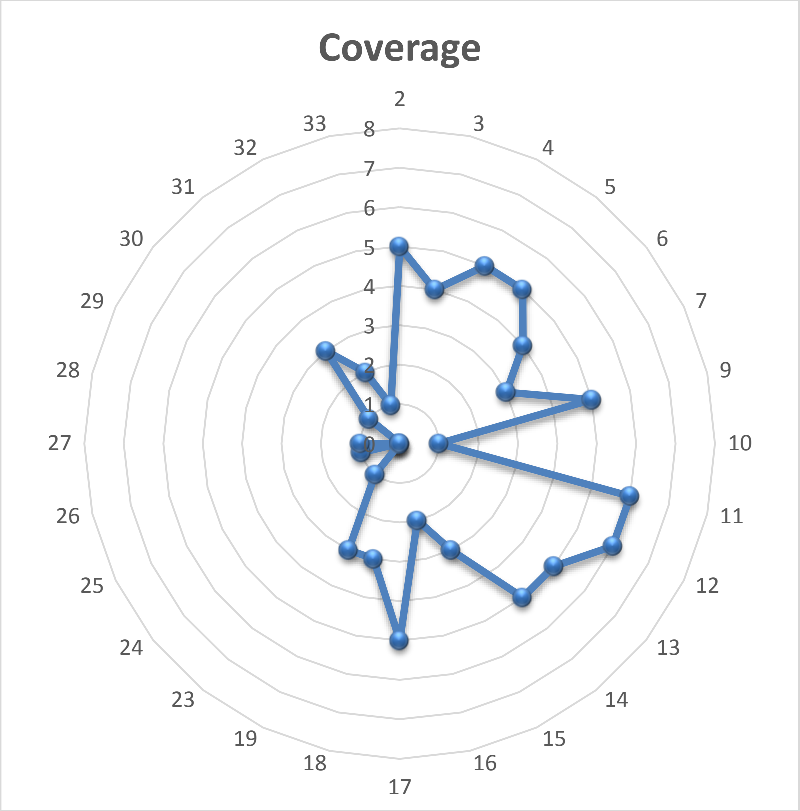


Figure 6-4. Overall usage including both elaboration and review stages

During the execution, we faced minor problems with respect to what we have planned: a few late deliveries (one day late), two subjects did not use the guidelines to support the reviews, and one subject did not inform which guidelines were used in each section of the plan, hampering the analysis of coverage for this case.

In the beginning, both subgroups had one subject not engaging the discussions. Therefore, the moderator asked them to contribute with their experience, whether they have similar experience when compared to other subjects. This sort of intervention was performed every time a subject was perceived to not contribute with the discussions.

As it was the first to be discussed, the guideline selected as control (regarding experimental design) worked also as an attempt to motivate the subjects to join the discussions, since it was the most used one. In terms of intensity of discussions, the previous contact with each guideline does not seem to influence that, as the first guideline discussed was the most used one with equivalent amount of arguments when discussing other initial guideline. However, the last guidelines had less discussion due to subjects' exhaustion to the long session.

6.3 Results

With the plans and reviews for the analysis, the first analysis is concerned with the guidelines' usage, i.e., which subject used the given planning guidelines and for what.

In general, we observed and confirmed by explicitly asking the subjects that 6 out of 8 adopted the proposed guidelines to support the elaboration and review of the simulation experiment plan. Two of them did not use the planning guidelines neither for elaboration nor for review, even knowing their use for reviewing was required. Besides, one of these six subjects did not report which guideline was used to support the elaboration of each section of the plan, but the subject detailed the guidelines used to support the review. Therefore, we can only make general assumptions for this case regarding the guideline application.

The following subsections present the results under different perspectives, as in the goals definition (Table 6-1), using both quantitative and qualitative data collected during the study execution and analysis.

6.3.1 Guidelines Coverage

As mentioned before, we consider coverage as the amount of planning information supported by the guidelines, according to the subject's indication. In this sense, we expected that the subjects could use the guidelines (from SG2-SG33) appearing as "1" (applicable) in Figure 6-5 to plan and review the simulation experiment. Conversely, they should not concern with guidelines appearing as "0" (not applicable).

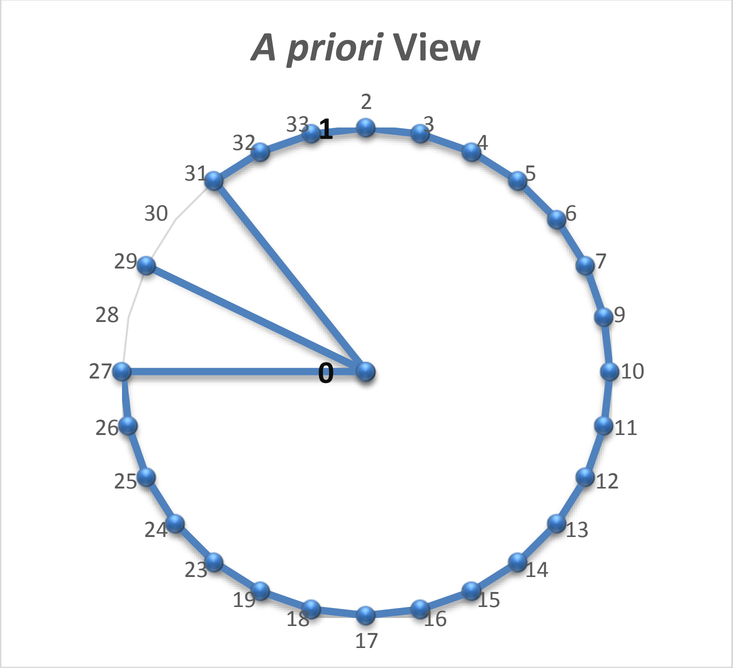


Figure 6-5. Usage expectation for the planning guidelines.

This way, apart from the exclusively reporting guidelines (SG1, SG8, SG20, SG21, and SG22) that are intentionally omitted in radars (Figure 6-5, Figure 6-6, Figure 6-7 and Figure 6-8), the subject may discard guidelines SG28 and SG30 in situations that no random variables are used and no calibration procedure was used for the simulation model, since it intends to be general, respectively.

For the elaboration stage, the frequency of use for each guideline (from SG2 to SG33) is presented in Figure 6-6. The frequency can reach up to five (5), since we have two subjects not using the planning guidelines for the elaboration and one not detailing which guideline is used in each section of the plan. The usage concentrates on the initial guidelines, concerned with the simulation experiment definition (context, problem, goals, questions, and others), simulation model description, and experimental design.

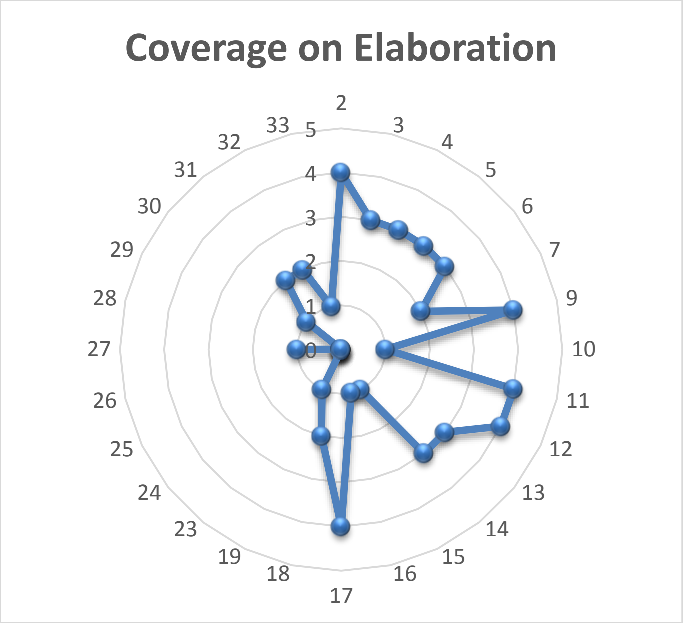


Figure 6-6. Coverage of Planning Guidelines in the Elaboration Stage

By fully reading the content of the plans, it is possible to observe that the planning guidelines were used twofold: (1) as conceptual reference to acquire better understanding of both simulation and planning issues; and (2) to effectively apply the guidelines to elaborate specific sections of the simulation experiment plan.

In addition, as also observed in the FG discussions (Section 6.3.4), all subjects assumed the adopted simulation model as being valid due to the existence of a journal publication. It partly explains the lack of usage regarding most of the guidelines from SG23 to SG33. In the same sense, the guideline SG10, also concerned with model validation issues, is mentioned only once. We understand these validity-related guidelines can be applied in two ways. First, by applying the V&V procedures mentioned in the guidelines. Second, by considering the unfeasibility of performing such sort of procedures as a potential threat to validity for the simulation experiment, besides addressing it in the analysis of results. Therefore, plans not taking guidelines model (SG10) and data (SG16) validity into account require more attention regarding the analysis of threats to validity.

At the review stage, the guidelines' coverage is very similar to the elaboration stage. Figure 6-7 clearly presents a recurrent pattern on the guidelines usage: the concentration in the initial guidelines. Even having six subjects using the planning guidelines as supporting instrument for the reviews, we separate the analysis for the subject that did not detailed it for the elaboration, as presented in Figure 6-8, where each guidelines (from SG2 to SG33) can be . In general, the use follows more or less the same patterns from Figure 6-7.

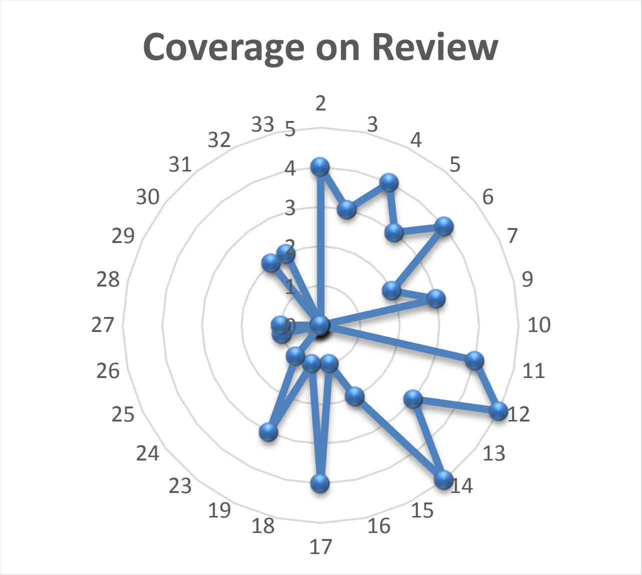


Figure 6-7. Coverage of Planning Guidelines in the Review Stage

Again, since the subjects assumed the simulation model as being valid, they had not much to point out in their reviews w.r.t. the lack of validation or threats to validity in other plans.

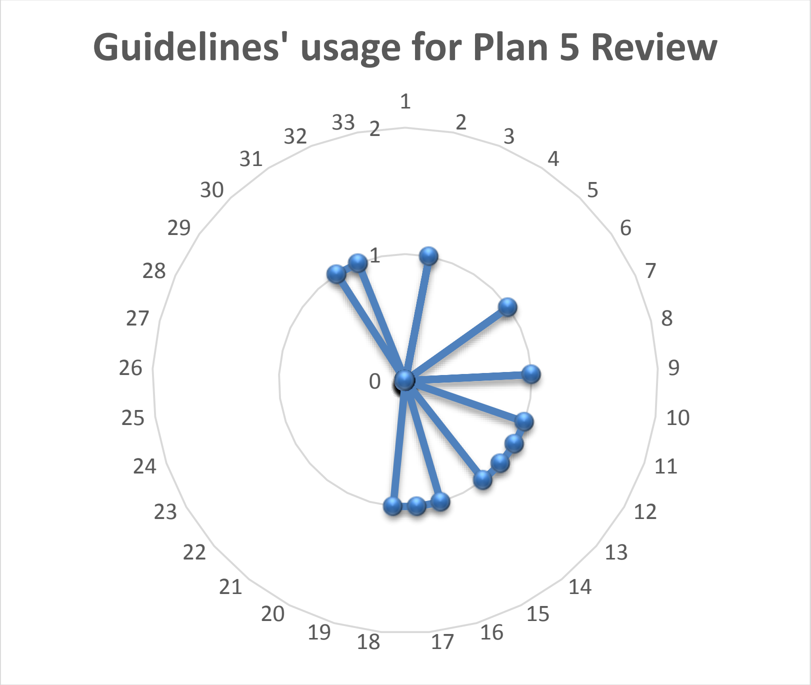


Figure 6-8. Planning Guidelines’ Usage for the review Plan 5.

In a general perspective, including both elaboration and review, we did not observed references to three - out of 26 - guidelines considered in the initial expectation (Figure 6-5). This may be an indication that these 23 planning guidelines can support (in some sense) the elaboration or review of the simulation experiment plan through the description of relevant aspects in the plan. The only aspect not considered by the subjects in the experiment plans regards the involved experiment costs. Considering the guidelines usage is not homogeneous and we cannot confirm their influence on the quality of the experiment plan, we are not able to affirm that the lack of cost-related information on the proposed guidelines caused the lack of such information on the elaborated plans.

6.3.2 Plans Coherence

The analysis procedure for coherence is based on the reviews performed by the subjects using the planning and reporting guidelines. Each discrepancy observed in the reviews forms is analyzed and the researcher checks whether it concerns any mismatch between plan’s sections. Besides, each discrepancy is associated to one guideline, which is driving the review at that topic. Examples of mismatches include problems and goals, goals and variables of interest, goals and experimental design, output analysis and experimental design, and so on.

Some general observation should be mentioned regarding six plans and reviews, since two reviewers did not follow the guidelines to perform the review and no relevant content was produced in their review report. For these two cases, reported issues include only presentation improvements, misunderstandings and empty sections in the reviewed plan.

In cases that participants followed no guideline to develop the experimental plan, it is possible to identify lack of coherence such as redundant or overlapping goals and mismatch among sections. However, we could also identify specific issues in this sense when authors declared to use the guidelines.

Ideally, the research plan should contain concise context descriptions, in which problem and goals should be grounded. Besides, research questions should be derived from the research goals. All this matching allows a coherent plan and makes easier the simulation output analysis. In general, the plans presenting coherent sections mostly mentioned the use of the initial planning guidelines, from SG2 to SG6. However, one reviewer reported a mismatch between the problem definition and the research goals and questions. Interestingly, in this case, the participant mentioned the use of SG3 (problem definition guideline), but did not mention the use of SG4 and SG5, respectively goals and questions definition guidelines. This may be the case in which the proposed guidelines promoted such effect, but we still need more control to affirm that.

We identify lack of understanding on hypotheses definition. Subjects do not separate null from alternative hypotheses. Furthermore, they rarely represent the research questions or the experimental design, in the sense that interest variables do not compose the hypotheses' statement.

The simulation model is usually presented in the simulation experiment plan, describing its variables and possible input parameters. However, plan 7 (elaborated without the guidelines' support) does not describe the simulation model as an instrument. Besides, as any other instrument, it should be characterized and assessed.

The experimental design should ideally accomplish the design matrix, determination of number of trials and output analysis. Actually, we observed interesting experimental designs, exploring a large number of input parameters and values. Two of them mentioned factorial designs, and others varied one factor (input parameter) at a time. For these cases, the reviews did not capture the most suitable type of design, w.r.t. research goals and questions, for instance. They mostly identified incoherence issues regarding research goals or questions mentioning specific variables not used in the experimental design, for example. Nevertheless, we could identify one reviewer pointing out an open question about reducing the number of factors for the simulation experiment. In this case, it does not seem to be applicable since it is a deterministic simulation not requiring a large amount of runs.

However, as the reviewers observed, not all the plans presented the design matrix, even when they mentioned the use of the proposed guidelines. In addition, we could also observe plans with empty sections regarding execution procedures. The number of trials seems to be a general concern, however there is only one rationale regarding how such number is determined, which regards the number of combinations given the number of factors and levels for the experiment. Such rationale is actually mentioned in the guidelines. Moreover, the output analysis, which is a consequence of the adopted design, is only mentioned regarding possible instruments for outliers' removal and statistical tests. One possible reason for the lack of planning regarding output analysis is due to classes/training being given after the elaboration of plans.

Participants presented few and short discussions regarding limitations and threats to validity in the plans. We will not discuss this issue here, but in the next section.

Finally, as the study followed a qualitative approach, we have no support to confirm causal relationships regarding coherence on the elaborated and reviewed plans. However, as mentioned before, it is possible to identify points of coherence on the elaborated plans and to link them to the aspects discussed in the guidelines, without statistical significance. Thus, we can interpret it as an indication of logical chaining among sections happening due to the use of planning guidelines.

6.3.3 Guidelines Effectiveness

We defined effectiveness in our study plan as the capacity of identifying threats to validity. Six plans containing threats to validity analysis: form 1 to 6. They are discussed in this section.

The mentioned threats to validity are unclear, leading to multiple interpretations. Besides, all plans present misclassification regarding the types of threats to validity: conclusion, internal, construct and external validity. For instance, in plan 4, the participant mentioned “the model calibration, by the input parameters setting, may not represent the best scenario for the observed project”. This statement is misclassified as a threat to conclusion validity, which clearly does not refer to the statistical confirmation (significance) of a relationship between the treatment and the outcome, in order to draw correct conclusions about such relations. Threats to conclusion validity involve the use of inappropriate instruments and assumptions to perform the simulation output analysis, such as wrong statistical tests, number of required scenarios and runs, independence between factors, among others.

Recurrent issues are threats to validity concerning the determination of parameters, which may not reflect real conditions, and model calibration. This threat is actually strongly related to the planning guidelines SG26, regarding valid parameters, and SG30 on calibration and experimentation data coming from the same population. However, just SG30 was mentioned in the elaboration stage and, since subjects had no information about calibration procedures and data collection, these aspects may actually impose threats to study validity and authors considered them correctly. On the other hand, the model specification presents the allowable range of input parameters. Subjects should reflect about them, adopt values close to the organization scenario, and explore additional scenarios when they realize feasible solutions for the reported problem. Valid values for input parameters are not only the observed ones, but also additional ones representing feasible alternatives to the current state of a system or process.

The input parameters validity is a concern of most experienced participants regarding their involvement with SE activities in the Industry. Only experienced subjects reported this kind of threat. It reinforces that skills for analyzing threats to validity not only depend on Experimental Software Engineering knowledge, but also on domain knowledge.

Seven subjects considered the model as valid. We have some indications of the beliefs towards the model validity is based on the existence of a published journal paper, which is explicitly mentioned by the author of plan 1. However, the only one highlighting model validity as a potential threat did not mention any mitigation action. Therefore, the participants ignored guidelines concerning planning activities to reduce threats to validity.

Some information regarding the model validity is available in (CHERNOGUZ, 2011). CHERNOGUZ (2011) presents several tests concerning with the model consistency in response to stepwise changes of input. They considered the model robust since it continues to operate despite abnormalities in input, at least within the reasonable scope of input assumptions. Furthermore, he explains the model assumptions and many perspectives about the Brook's Law. Such information could be explored in the threats to validity section.

No plan presented the perspective regarding the model validity against the organizational context. For this case, the conceptual model may be valid, but not representing the reality of the specific organization. The only concern related to this match (model - organization) refers to choosing adequate input parameters.

Maybe, the lack of instructions regarding specific techniques mentioned in the guidelines, such as Sensibility Analysis, Face Validity, Rationalism, as well as procedures for generating the design matrix for the mentioned types of design, discouraged the participants to use such guidelines accordingly. However, the proposed guidelines do not mean to be a tutorial on how to plan simulation experiments. Even though, we should cite direct resources on how to conduct each referenced procedure or technique.

Finally, the experimental plans do not mention threats to validity related to the experimental design. Missing factors, number of simulation runs, and correlation among factors may impose threats to simulation experiments. The only threat mentioning experimental design highlights the possibility of not finding an optimal solution (scenario), due to the use of fractional factorial design. However, the reduced number of scenarios by using fractional designs should not exclude robust or optimal scenarios.

6.3.4 Focus Group

As the FG design is organized to discuss each planning guideline at time, we present the results in a separate way. Besides, at the end, we present general contributions from the second stage, which apply to the whole set of guidelines.

Overall, eight planning guidelines were considered useful and two as out of scope for this study. Six of them were considered useful as checklist as some types of validation are often overlooked. This way, we have some indication regarding their usefulness even with low usage on the elaboration state, eventually biased by the existence of a refereed publication with the proposed model, as mentioned in section 6.3.1. None was considered easy to use, mainly due to lack of orientation on how to perform specific procedures, or by the lack of experts to support with domain knowledge. The first five evaluated guidelines had a lot of discussion regarding their usefulness and ease of use, explaining scenarios from the simulation experiment and possible alternative ways of applying them. Following, we present the results for each guideline in the order they were discussed during the session.

SG12. Experimental design (design matrix), including independent and dependent variables and how levels are assigned to each factor should be reported. In general, the guideline is claimed to be useful to remind the identification of factors, levels and their description using the design matrix, i.e., use it as a checklist. However, participants seem to face difficulties on how to perform what the guideline states. In part, such difficulty concerns domain knowledge. Besides, they mentioned the use of course's slides and textbooks on experimentation to overcome this. We discussed the possibility of such additional information to be already in the planning guideline SG27, since it discusses characteristics and relevant criteria to select a proper experimental design. The subjects expressed a concern regarding the scattered information across different guidelines, and it influences the way the planning guidelines are used. Furthermore, an example of how to build the design matrix can also be useful to improve the ease of use.

SG10. Gather as much evidence as possible regarding the simulation model (conceptual and execution) validity. Much has been discussed on how to perform the model validation procedures. In this case, the discussions lose focus from what the guideline actually states, that is, evidence regarding the model validity. The reason seems to be the list of V&V procedures presented in this guideline's description. However, such list is presented to illustrate which procedure is often performed on V&V attempts of SE simulation models. This way, the reasoning is that the use of this list could make easier the identification of attempts to verify or validate a simulation model in the report of simulation studies and, then, it enables to collect the successful attempts. At the end of the discussions, the subjects suggested that this guideline should be used as a more general one to reorganize the planning guidelines related to simulation model validity.

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. This guideline is a real example of what the FG classified as useful, but difficult to use. Its usefulness appears on the opportunity of an expert possibly recognizing what is being simulated and realizing reference behaviors that can be used as comparison baseline, as well as to identify eventual inconsistencies in model behavior. Among the hindrances, it is possible to highlight the tradeoff between the effort to perform Face Validity and the return w.r.t. the model validity. Furthermore, the experts' availability to perform such procedure can be difficult to dispose.

SG24. Support model (causal) relationships, as much as possible, with empirical evidence to reinforce their validity and draw more reliable conclusions. This is another guideline that is consensually claimed as useful. However, it was criticized regarding the ease of finding evidence for causal relationships in Software Engineering. However, the guideline mentions "as much as possible". It reinforces the idea that no procedure can assure validity alone.

SG25. Always verify model assumptions, so the results of simulated experiments can get more reliable. In this case, the arguments in favor of the guideline are weak and imposed by the role played by the group, for instance, "it is useful since it increases model confidence" and "it is easy to use because of the existence established procedures for that". It is an example of situations we used the third researcher perspective, by observing the focus group behaviors such as irony and jokes. The guideline description mentions V&V procedures to support the verification of model assumptions, since they are explicit. In general, this guideline was understood as just another to assure model validity, and during the discussions, some concepts needed some clarification. The subjects mentioned they would not think on verifying assumptions without the guidelines. It may be an indication that guidelines related to model validation need to be reorganized, making explicit which validity perspectives are discussed and that there is no unique procedure capable of assure general model validity.

SG26. Use results from Sensitivity Analysis to select valid parameters' settings when running simulation experiments, rather than model "fishing". The subjects playing the "haters" role showed distinct perspectives regarding how to perform the Sensitivity Analysis (SA). Even recognizing that SA may be effort and time-consuming on its execution and output analysis, it is the main approach for understanding and characterization of phenomena represented by the model.

SG28. When dealing with simulation model containing stochastic components, determine the number of runs needed for each scenario in order to capture the phenomenon variance. The subjects understand this guideline is useful, but it does not really gives guidance on how to determine the number of required simulation runs for stochastic models. We understand it hampers the application of this planning guideline. This way, we can include the procedure proposed by LAW and KELTON (2000) on the guidelines or a reference for it.

SG29. Keep track of qualitative data along with quantitative data. It is also important to record data contextual information. For this study, this guideline does not seem to be useful, either by the lack of relevant qualitative and contextual data or by the unfeasibility of collecting them during the simulation study execution.

SG30. Make sure that both calibration and experiment datasets came from the same population. This guideline does not apply to the study under planning, due to the simulation model intent to be general-purpose, using no calibration procedure. The project context for this model is always defined in terms of input parameters, i.e., the calibration data is always the same when compared to the simulation experiment input data. For this guideline, it is important to add model characteristics that should be satisfied.

SG33. Be aware about data validity when comparing actual and simulated results: compared data must come from the same or similar measurement contexts. During the discussion of this guideline, the group was exhausted so the arguments presented regarding the guidelines usefulness and ease of use did not advance so far.

Even though, both lovers and haters presented consistent arguments regarding its use, mainly concerning with external validity issues.

General contributions to facilitate the guidelines use (second stage). Two participants suggested making explicit whether the guidelines apply to the simulation model development or experimentation. Such discussion makes sense because simulation-based studies' lifecycle encompasses both stages. However, the planning guidelines meant to be specific for simulation experiments and not the whole lifecycle. Besides, we agree that guidelines related to the study definition, like context, problem, research goals and questions are applicable for both situations.

In some cases, they mentioned that explicit definitions or concepts (something such as a glossary) could help on understanding what the guidelines are stating. Furthermore, they claimed for examples on how to apply each guideline, due to the level of abstraction they are discussed. These are the cases where we explicit mention V&V procedures and experimental designs for simulation.

6.3.5 Quality of Elaborated Plans

Apart from the previous perspectives adopted in the analysis, we also analyzed the elaborated plans with respect to their content. This analysis considers the quality of each plan based on three criteria: (1) the matching between the study plan and the given scenario and simulation model description, (2) correctness of each plan section based on the defined goals, and (3) the capability of executing the plan.

In general, all the simulation plans establish a link with both the organizational scenario and model description. Considering the general directions given in the scenario, the plans establish at least one goal concerning the number of additional team members when the organization perceives the project is late. This is the main problem described in the given scenario, but other goals are investigated in the plans. Therefore, research questions derived from these goals raised the impact of the team configuration on the project schedule.

As discussed in Section 6.3.1, the produced plans share, in general, some issues regarding hypotheses definition. Null and alternative hypotheses, when defined, do not relate to each other, i.e., they capture distinct concerns, which clearly denotes two separate hypotheses with no null-alternative definition. Two study plans correctly present null and alternative hypotheses that actually match the research questions and base the experimental design.

Besides other research goals presented across the plans, the variables of interest selected as independent variables or factors for the simulation experiment vary a lot from one plan to another. In the simulation model description, we highlight eleven input parameters (independent variables). It is possible to observe one plan using only one variable as a factor and another one considering all variables. In addition, we observed one case selecting an intermediate variable as a factor. However, it is not an input parameter for the simulation model and the model user should not modify it. For instance, the variable **Total Personnel** (Figure 6-1) represents the sum of number of rookies and veterans in the project team. In this case, the number of rookies and veterans should be selected as factors instead.

We considered all plans having relevant information to investigate scenarios or configurations to keep the project behind the schedule. However, some of them miss important factors and it can bring incomplete or incorrect answers for the research questions. Even worse, for characterization or understanding purposes, they may have no relevant result to present at the end of the simulation experiment.

Another important issue is the selection of non-relevant variables as factors. Some variables are more distant from the research question and understanding their effects may not contribute to answer the research questions. In other words, the set of interest variables should encompass the scope of the research question in terms of scenarios, since it may enlarge the effort of analysis.

For this simulation model (CHERNOGUZ, 2011), many variables have a degree of interdependency. It may be the case that some factors (input parameters) do have implicit correlations. In other words, the effects of changing two or more factors simultaneously can reveal positive interactions, where they complement each other, or negative interactions, implying factors to be partial substitutes for each other. This way, experimental designs should test possible interactions among factors and analyze their sensitivity. Nevertheless, except from one plan, no other plan provides an experimental design capable of testing such interactions, for two reasons: (1) either plans miss potential correlated factors, overlooking one or more relevant factors or (2) the experimental designs test one factor at a time.

The lack of information and precision regarding the experimental design definition compromised the executability on half of the plans. Several results can be obtained by running the scenarios, since design matrices do not specify which values need to be assigned for each factor in each simulation run. The same can be observed for output analysis, there is no information regarding how instruments have to be applied to the outcomes in order to determine the effects of each factor and identify feasible solutions for the investigated problem. One example is the general claim (all plans mentioned that) to the use of box plots in order to identify outliers. However, there is no reflection regarding what could be an outlier in a simulation experiment, using a deterministic simulation model.

The instrumentation for the simulation experiment is not clear in all the plans. In general, the authors mention the instruments (mainly, the simulation model and the Vensim PLE tool), but there is no procedure defining their use. It is quite important since the Vensim PLE version has several limitations regarding multiple trials execution.

As all subjects assumed the model as being valid, there is no discussion of threats to validity regarding model validation issues. However, the discussed experimental design issues also represent threats to validity and they were rarely discussed, except for the parameters validity and calibration method. Both threats discussed in half of the plans.

Overall, both lack of information and correctness in the plans compromise their execution or make their results unreliable. The incomplete stage of some plans can hinder the capability of executing them, at least, not executing as planned. Additionally, some of these problems were captured in the reviews as mentioned in Section 6.3.1. Therefore, for the sections actually supported by the planning guidelines, some of these issues could be avoided. Such result cannot be generalized neither for all the guidelines nor for all the plans, since the use of planning guidelines was optional to elaborate the plans. Besides, the usage of one guideline may require the usage of another, but we could not observe this kind of dependency.

6.4 Threats to Validity

After the analysis of results, we understand the evaluation as successful in two aspects: (1) every focus presents positive aspects regarding the simulation guidelines, and (2) several improvement opportunities that can potentially contribute to the guidelines' evolution. These results are constrained by the threats to validity discussed in the following subsections.

6.4.1 Conclusion Validity

As an observational study, we have no ambition of reaching a quantitative analysis or showing statistical significance regarding the results. Mainly, our results are expressed in terms of reasoning. Therefore, our results are limited regarding the conclusion validity.

6.4.2 Internal Validity

Our study comprehends the observation through instruments of human subjects during tasks involving planning and execution of a simulation experiment. Thus, we need an understanding on how the use and application of the proposed guidelines could influence the quality of a simulation experiment plan, for this context. In general we perceived positive aspects regarding the application of the planning guidelines (see Sections 6.3.2, 6.3.4 and 6.3.5), even with no statistical significance. However, we observed particular situations where the use of the proposed guidelines has no effect, which is the case of not presenting a design matrix even mentioning the use of guidelines related to it.

There are possible factors contributing to such adverse cases, including the instruments. One of them is the adopted template for the study plan, since it intends to be general (no specific study strategy). In this case, the lack of specific sections to fill may have discouraged the subjects to add new sections to the simulation experiment plan. Actually, we identified no new section throughout the plans. Furthermore, training sections (classes) avoided, at most, presenting the planning guidelines. However, it was unfeasible to not present closely related aspects and concepts.

Two subjects did not use/read the planning guidelines before the FG meeting, which was possible for the plan elaboration, but required for the review. They were the ones with lowest participation in the discussions, since they have no previous experience on the application of the guidelines. This way, they only realized how the guidelines application would be for each case.

Finally, the existence of a published journal paper influenced the subjects' opinion regarding the simulation model validity. It limited our capacity of observing planning guidelines related to several threats to validity, since they assumed these threats are not applicable to the simulation model and study.

6.4.3 Construct Validity

From the construct point of view, our main threat is the possibility of surrogate measures not representing the interest variables, i.e., effectiveness, coverage, coherence, perceived usefulness and ease of use.

The effectiveness is the most challenging focus to evaluate. Actually, to assess the capability of avoiding or identifying threats to validity we need a more controlled context, in which we have previously identified threats (like a thesaurus) for the specific case and determine the if the application of guidelines is enough to identify all of them.

As the perceived usefulness and the ease of use have a subjective meaning, we adopted the strategy of triangulate the data collected during FG (using post-its and notes from two observers).

6.4.4 External Validity

For this study, we have no expectation of generalizing results from such small population and context, since we intend to understand how these guidelines could be applied and their effects.

The guidelines propose orientation for researchers not familiarized with simulation studies on how to plan simulation experiments. The sample population share common characteristic with the expected population, since they are young researchers with no previous experience in simulation studies.

6.5 Conclusions

The qualitative nature of the evaluation strategy defined for this evaluation reduced the chance of hypotheses testing, but it increased the observation capacity due to the variety of sources of information, including the elaborated plans, the reviews, and the information collected in the FG dynamics.

The details on the guidelines' application were analyzed in micro and macro perspectives. In other words, we could identify general and specific issues regarding their applicability for the proposed scenario. The study's results showed positive aspects regarding coverage, coherence and perceived usefulness. Besides, we have a limited experience on capturing their effectiveness, since the identification of threats to validity was hindered by the subjects' assumption regarding the model validity. Finally, the ease of use was mainly evaluated based on the guidelines application. As the specifics of V&V procedures and DOE techniques are not included in the guidelines discussion, the subjects did not follow the suggestions on their application.

Specifically about FG, this systematic approach adopted to conduct FG indicates its usefulness as an alternative tool to support data collection in SE research (DE FRANÇA *et al.*, 2015). Particularly, we observed advantages on exploring strategies to stimulate participants on performing the activities involved on the FG dynamics. The most significant ones are their commitment to the study and the detailed data regarding the participants' perceptions about how the planning guidelines can be applied.

We are aware of potential threats to the studies validity, including some we could not anticipate. However, we understand that most of these threats are applicable for other settings, also involving qualitative methods.

In the next chapter, we present the set of guidelines in a joint perspective including both planning and reporting as a result of this evaluation. Therefore, we performed changes on their organization and clarified some concepts and discussions.

7 Guidelines for Simulation Experiments in Software Engineering

In this chapter, we present the results of successive improvements as consequence of the performed studies aimed at evolving the guidelines for planning and reporting of simulation experiments in Software Engineering. We also provide examples from a proof of concept to observe the guidelines feasibility.

7.1 Introduction

As presented in Chapter 6, we identified many improvement opportunities in the proposed set of guidelines for reporting and planning SBS in SE. Possibly, the main one regards its organization and presentation. The former separation on reporting and planning perspectives caused some misunderstandings as to which and when to apply each guideline, as well as the observation of multiples guidelines covering the same aspects under different perspectives.

Essentially, this section presents the current version of the proposed guidelines as one single set (Table 7-1). Neither theoretical knowledge nor new aspects were changed in the guidelines. The modifications include the reorganization of sections and guidelines, as well as three unifications we performed with the guidelines SG14 and SG28, SG18 and SG31, and SG19 and SG32, since they were closely related. Besides, we improved and clarified the statements and discussions, added new examples from a proof-of-concept and reviewed the whole text. The proof-of-concept consists on planning and executing a simulation experiment focused on software evolution, using the SD model proposed by (ARAÚJO, MONTEIRO and TRAVASSOS, 2012). All changes aim at improving the readability and understandability of the guidelines.

Table 7-1. Overview of the simulation guidelines

ID	Guideline Statement
Identification	
SG1	Proper title and keywords should objectively identify the simulation study, and a structured abstract should summarize its contents
From Context to Research Questions	
SG2	The context where the simulation study is taking place should be captured in full
SG3	Explicitly state the problem motivating the simulation study, so that research questions can be derived
SG4	Clearly state the simulation study goals and scope
SG5	Derive the research questions from the established goals

ID	Guideline Statement
SG6	Clearly state the null and alternatives hypotheses from the research questions
Simulation Feasibility	
SG7	Present justifications for considering simulation studies as the ideal or feasible observation strategy
Background and related work	
SG8	Present only essential background knowledge and the related works
Simulation Model Specification	
SG9	Have a detailed description and understanding of both conceptual and executable simulation models, as well as its variables, equations, input parameters and the underlying simulation approach
Simulation Model Validation	
SG10	Gather all evidence regarding the simulation model (conceptual and execution) validity
SG11	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
SG12	Support model (causal) relationships, as much as possible, with empirical evidence to reinforce their validity and draw more reliable conclusions
SG13	Always verify the model assumptions, so the results of simulated experiments can get more reliable
Subjects	
SG14	Characterize the subjects involved in the simulation study as well as their training needs
Experimental Design	
SG15	Describe the experimental design (design matrix), including independent and dependent variables and how levels are assigned to each factor
SG16	Use Sensitivity Analysis to select valid parameters settings when running simulation experiments, rather than model “fishing”.
SG17	Consider as factors (and levels) not only the simulation model’s input parameters when designing the simulation experiment, but also internal parameters, different sample datasets and simulation model versions, implementing alternative strategies to be evaluated
SG18	When adopting ad-hoc design determine the selected simulation scenarios and explain the criteria used to identify them as relevant
SG19	When dealing with simulation model containing stochastic components, determine the number of runs required for each scenario, along with its rationale, in order to capture the phenomenon variance.
Supporting Data	
SG20	Assess, whenever possible, the data used to support the simulation model development or experimentation
SG21	Keep track of contextual information (including qualitative data) along with quantitative data
SG22	Make sure that both calibration and experiment datasets came from the same population
Simulation Supporting Environment	
SG23	Set up and describe the simulation environment, including the supporting tools, associated costs, and decision for using a specific simulation package
SG24	Determine which and how intermediate measures are stored among simulation trials to be used in the final analysis
Output Analysis	
SG25	Determine which statistical procedures and instruments support the output analysis, as well as the underlying rationale, quantifying the amount of internal variation embedded in the (stochastic) simulation model to augment the precision of results
SG26	Be aware about data validity when comparing actual and simulated results: compared data must come from the same or similar measurement contexts
Threats to Validity	
SG27	Consider to check for threats to the simulation study validity before running the experiment and analysing output data to avoid bias, as well as to report non-mitigated threats, limitations and non-verified assumptions
Conclusions and Future Works	
SG28	Main results/findings should be identified and summarized, as well as the conclusions arising from the results.
SG29	Applicability issues should be addressed in the report, considering organizational changes and associated risks.
SG30	Point out future research directions and challenges after current results.

The following subsections present the set of guidelines according to their new organization (Table 7-1), aiming at representing the logical chaining of the SBS lifecycle (BALCI, 1990). In each subsection, we present at least one guideline, as well as an associated discussion and applicable examples from the conducted proof-of-concept regarding software evolution.

7.2 Identification

At first, a study report should be accessible. In other words, it should be easy to find it in (digital) libraries or through search engines. For that, the report title, abstract and keyword should contain all relevant words regarding the main topic and findings.

SG1. Proper title and keywords should objectively identify the simulation study, as well as to have a structured abstract summarizing its contents.

The choice of a proper title has no straightforward rule, but it should address the main topic of the study and also the main research contributions. Keywords generally depend on a glossary of terms used by the publishers. For instance, the term “Computer Simulation” can be identified in many libraries as a general term.

We suggest the use of structured abstract, as this eases the identification of the research context, problem, goals, applied methods, main results, and conclusions. It helps readers to quickly identify whether the study is relevant for their research purposes. An example of structured abstract can be found in IST (Information and Software Technology) instructions for authors’ page[8].

7.3 Study Definition

As any research initiative, the context, problem, goals and scope are extremely important, even when talking about SBS. This kind of study strongly depends on the collected data supporting the simulation model development and calibration. It is also true in SE, where the context of software projects, the human nature of software development activities, and the amount of unknown variables may influence the results of the studies.

SG2. The context where the simulation study is taking place should be captured in full.

[8]http://www.elsevier.com/journals/information-and-software-technology/0950-5849/guide-for-authors#39001

Simulation models in SE often come from research initiatives. Both academic and industrial projects are potential environments for SBS taking place. In industrial contexts, the description should characterize the organization where the phenomenon has been observed and data has been collected. Information regarding involved technologies, personal profiles, types of projects performed in the organization, operational procedures, and non-technical issues (cultural, restrictions imposed by policies, laws, and standards, for instance) are relevant for correct interpretation of results. Such contextual information can clarify unexpected behaviours or explain why specific behaviours cannot be generalized. In academic contexts, the research background and the project goals should be addressed.

HÖST, WÖHLIN and THELIN (2005) propose a context classification scheme (Table 7-2), based on two orthogonal factors: incentives and the experiences of subjects. It is particularly applicable to *in virtuo* experiments, where human subjects affect the simulation progress.

Table 7-2. Context Classification Scheme [adapted form (HÖST, WÖHLIN and THELIN, 2005)]

Incentive	**Experience**
I1: Isolated artefact	E1: Undergraduate student with less than 3 months of recent industrial experience
I2: Artificial project	E2: Graduate student with less than 3 months of recent industrial experience
I3: Project with short-term commitment	E3: Academic with less than 3 months of recent industrial experience
I4: Project with long-term commitment	E4: Any person with industrial experience, between 3 months and 2 years
	E5: Any person with over 2 years of industrial experience

The incentive factor is more related to the study relevance and environment setting. Artefacts and projects are usually artificial in simulation, which may keep the incentive low. However, the supporting data for calibration can be real. The Experience factor is strongly related to subject characterization, which is a concern in Section 7.8. Anyway, the incentive and experience scales should be revisited for the application of such schema of SBS characteristics.

PETERSEN and WÖHLIN (2009) present a more general set of context information to be considered, where they propose a context description based on six facets related to the object of study, according to Figure 7-1. This proposal concerns industrial studies. However, some of these facets can also be used to contextualize SBS.

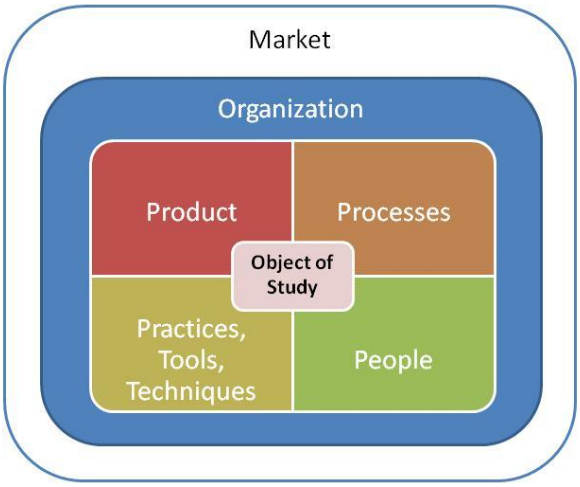


Figure 7-1. Context facets [adapted from (PETERSEN and WÖHLIN, 2009)]

In SBSs, the object of study regards the simulation model. Therefore, depending on the research goal and model validity, the object of study may be the simulation model itself or the phenomenon/system/process abstracted by the model. All the facets in Figure 7-1 interact with the object of study in some way. Moreover, these facets are represented, in SBS, by the calibration data or input parameters.

These facets also concern the practical implementation of the simulation results into a real context. This way, the entire context (environment and prerequisites) assumed by the simulation model should be guaranteed or handled in the real context. Consequently, target processes need to be modified, techniques or tools need to be incorporated, and teams need training sessions.

Both proposals for contextual descriptions aforementioned establish discrete variables (such as incentive, experience, processes, people) to describe the context information. However, DYBÅ, SJøBERG and CRUZES (2012) propose the use of a broad perspective approach for the so-called *omnibus* context. In summary, it describes the context in such a way that allows answering research questions such as "*What* technology is most effective for *whom*, performing *that* specific activity, on *that* kind of system, under *which* set of circumstances?" For the authors, the object of study and its context keep a 'mutually reflexive relationship', i.e., with both shaping each other in the same intensity. Thus, the context definition depends on the ecosystem in which the object under investigation takes place.

SG2. Example from the proof-of-concept.

The proof-of-concept motivation converges on two aspects: (1) a feasibility assessment of the proposed guidelines for simulation experiments; and (2) the understanding of how a project manager can breakdown long-term releases of a large-scale Web-based information system for business processes control (in financial and administrative domains) in a research supporting organization into different strategies for definition of the releases periodicity. The project team is geographically distributed in two sites, following an iterative and incremental software development process, in a CMMi Level 3-like maturity level, emphasizing V&V activities. Twelve developers compose the geographically distributed team, using Java and JSF as the development platform. Also, using CASE tools such as version control repositories, *bug tracking* and effort spreadsheets. This simulation (*in silico*) experiment also intends to show how the SD model for the observation of software evolution (ARAÚJO, MONTEIRO and TRAVASSOS, 2012) can be used as instrument to support the answering of research questions regarding software maintenance.

Once contextual information has been gathered and understood, the problem should then be stated and described as to how it was identified. Problems arise from a critical situation or from repeated situations where the solution is complex or expensive.

SG3. Explicitly state the problem motivating the simulation study, so that research questions can be derived.

Along with the problem statement, the reason why it happens and the impact it causes are important to highlight the implications of not solving such problem. For problem statement, we adopt a template proposal[9] that satisfies these needs, 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).*

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

SG3. Example from the proof-of-concept.

The problem investigated in our simulation experiment (proof-of-concept) regards the software life cycle at the time the information system changes from development to 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 new product releases observing the restrictions regarding the 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 of costs and possibly influencing the decay of product quality.

SG4. Clearly state the simulation study goals and scope.

The clear definition of research goals is the first step after establishing the problem. It is likely to find, in SE studies, the definition of the goals using the GQM approach (BASILI, 1992). It is completely useful for defining SBS goals, since current studies present non-structured goal definitions, as in (MELIS *et al.*, 2006) (CELIK *et al.*, 2010), and it may be difficult getting the right point. Besides, the scope should be explicitly stated, establishing boundaries for the research area, domain, and type of systems or processes under investigation.

Therefore, the common goals for SBS shall include developing a basic understanding (characterization) of a particular simulation model or phenomenon, finding robust or optimum alternatives, and comparing the merits of various alternatives.

In software process simulation with SD, IMMoS methodology (PFAHL and RUHE, 2002) provides a more specific template, similar to GQM goal definition, structured in five dimensions as presented in Table 7-3.

Table 7-3. Goal definition dimensions from GQM and IMMoS [adapted from (DE FRANÇA and TRAVASSOS, 2015)].

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

SG4. Example from the proof-of-concept.

The goal from the proof-of-concept in the GQM goal template (Table 7-4).

Table 7-4. Goal definition using GQM template for the proof-of-concept

Purpose	
Analyze	The evolution of a large-scale information system
For the purpose of	Characterization
Perspective	
With respect to	Duration of maintenance cycles (periodicity), as well as its effects on product quality
from the point of view of	SE Researcher
Environment	
In the following context	Simulating (*in silico* environment) quality decay for a large-scale Web information system for business process control (in financial and administrative aspects), with the use of a SD model as instrument. The supporting data reflects a software project running an iterative and incremental lifecycle, in a CMMi Level 3-like maturity level, including verification, validation and testing techniques. Twelve developers compose the geographically distributed team, using Java and JSF as the development platform. Also, using CASE tools such as version control repositories, *bug tracking* and effort spreadsheets.

SG5. Derive the research questions from the established goals.

The SBS goals should match the capabilities of the simulation model. This way, the model should be able to answer the research questions through the output data. Deriving research questions from the defined goals is part of the GQM approach, as well as defining metrics to answer such questions. Optionally, these metrics can support the definition of hypotheses representing assumptions under which the model has been developed, and that should be submitted to statistical tests.

SG5. Example from the proof-of-concept.

Based on the goal defined in Table 7-4, we derived two research questions for the simulation experiment:

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 the product quality?

SG6. Clearly state the null and alternatives hypotheses from the research questions.

Considering the controlled environment, there is always (at least) a hidden hypothesis. It is also useful to discuss how such hypotheses were raised, describing the rationale or theory from where they came.

7.4 Simulation feasibility

It is important to assess the feasibility of simulation as a candidate approach to solve or investigate the problem. To the best of our knowledge, BALCI (1990) is the only resource available in the technical literature supporting this kind of analysis. BALCI (1990) suggests some questions as indicators of simulation feasibility. These questions are driven by context variables such as cost, time, benefits and the relationships among them, what naturally limit the observation field. To overcome this limitation we have added questions to his proposal.

SG7. Present justifications for considering simulation studies as the ideal or feasible observation strategy.

The simulation goals should be more than getting values for output variables. This sort of goal resembles studies using analytical or regression models. Simulation outputs also comprehend a rationale, i.e., an explanation or chain of changes in the system that results in 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 on processes, products or people may give a suitable solution. In this sense, we recommend additional questions to support the decision-making about performing or not the SBS. Therefore, it is necessary to focus in more technical constraints regarding the 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 risks (including loss of money or time, reach an irreversible state of the system, safety) of the real phenomenon high? In addition, data should be available in order to accomplish statistical tests and calibration of variables and equations involved in common approaches such as SD and DES.

SG7. Example from the proof-of-concept.

Although we have presented some motivations to perform the simulation experiment, the use of simulation in this context can be justified 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 the loop involving Effort, Maintainability, and Reliability).

7.5 Background and Related Works

Theoretical foundations and background knowledge are essential parts of the study report. Without them, it could be a great barrier for a distant reader or junior researchers.

SG8. Present only essential background knowledge and the related works.

Presenting all the theoretical foundations may miss the focus in the study results. Essential knowledge should be presented and some important references should be pointed out for detailed understanding. Besides, the same would be applied to related works, presenting just the SBS closely related to the performed study, i.e., investigating the same or related phenomenon. Any other study can be just referenced. It also includes previous related works from the study author.

7.6 Simulation Model Specification

The guidelines focus is on simulation experiments. Model development issues are out of their scope, except those aspects in the frontier between model development and use as model specification and validation. For such purpose, it is important to know the model in detail. No matter if the model has been developed or not by the experimenter. 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. The lack of knowledge about any of these aspects may impose different types of threats to validity.

SG9. Have a detailed description and understanding of both conceptual and executable simulation models, as well as its variables, equations, input parameters and the underlying simulation approach.

The model description is useful to supplement the information regarding the experimental design and on how the values for input parameters in each simulation run are determined. Such description should include essential characteristics of the underlying simulation approach. The abstraction and execution mechanisms are understood immediately by identifying the simulation approach. For instance, when describing a SD model, it is possible to infer how simulations are executed and the stocks and flows modelling abstractions, as well as description of the causal relationships and feedback loops.

Usually, model specifications are expressed as diagrams, equations, and textual descriptions. Diagrams capture the overview and the conceptual simulation model. Equations detail the model, allowing coding the model in other simulation tools. Lastly, textual descriptions supplement and clarify any doubt regarding the previous specifications.

Moreover, the model boundaries should be specified. It is possible to identify in some reports simulation models labelled as 'requirements engineering simulation model', for example. However, such model rarely encompasses the whole requirements engineering process, including all possible activities and variables. Therefore, uncovered aspects, assumptions and limitations representing simplifications of the real system should also be explicitly described.

SG9. Example from the proof-of-concept.

ARAÚJO, MONTEIRO and TRAVASSOS (2012) proposed the model used as instrument in our simulation experiment. It presents an infrastructure based on the Laws of Software Evolution (LSE) to observe software quality decay throughout software development and maintenance processes. The main idea is to get a better understanding on how the software system may be affected by several changes occurring in its lifecycle. In order to support the evolving system's behavior 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 Figure 7-2.

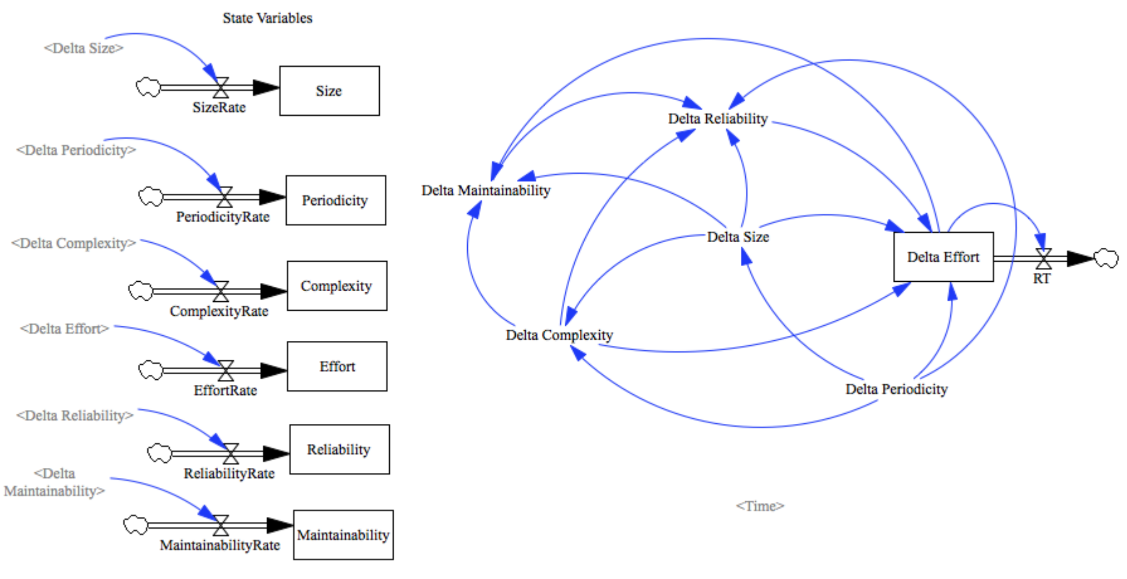


Figure 7-2. Software evolution model adapted from (ARAÚJO, MONTEIRO and TRAVASSOS, 2012)

The model was developed over six state variables, which represent the combined status for both project and product:

- **Periodicity**, the time interval between each release version of a produced artifact (e.g., software or documentation versions);
- **Size**, the magnitude of artifacts produced in each life cycle 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 artifact (e.g., cyclomatic complexity of methods, or number of classes in a class diagram);
- **Effort**, the amount of work done to produce a version of some artifact (e.g., measured in terms of man-hours or equivalent unit);
- **Reliability**, the number of defects corrected per artifact in each software version;
- **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.

The model equations for each relationship was derived from the project historical data, using least squares regression. This way, each model variable is defined in terms of its related ones in a linear shape. The equations for each variable are presented below:

$$\begin{aligned} Delta\ Reliability \\ &= (0.199239 \times Delta\ Periodicity) + (2.23944 \times Delta\ Complexity) \\ &+ (0.167732\ \times\ Delta\ Maintainability) - (0.92313 \times Delta\ Size) \end{aligned}$$

$$\begin{aligned} Delta\ Maintainability \\ &= (1.01131 \times Delta\ Effort) - (11.9523 \times Delta\ Complexity) \\ &- (4.37683 \times Delta\ Size) \end{aligned}$$

$$\begin{aligned} Delta\ Effort = (0.389594\ *\ Delta\ Periodicity)\ &- (8.17702\ *\ Delta\ Size) \\ &- (75.3565\ *\ Delta\ Complexity)\ + (4.54464\ *\ Delta\ Reliability) \\ &-\ RT \end{aligned}$$

$$Delta\ Complexity = (0.0706133\ *\ Delta\ Size)\ + (0.00251206\ *\ Delta\ Periodicity)$$

$$Delta\ Size = 0.0256675\ \ *\ Delta\ Periodicity$$

7.7 Simulation Model Validity

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

SG10. Gather all evidence regarding the simulation model (conceptual and execution) validity.

Evidence regarding model validity means the experimenter should be aware about the initiatives (previous reports and research papers) of submitting the simulation model to V&V procedures, understanding their results. Table 7-5 presents examples of V&V procedures for simulation models found through the undertaken qSLR (section 2.2).

Table 7-5. Verification and validation procedures for simulation models

Procedure	Description
Face Validity	Consists of getting feedback from individuals knowledgeable about the phenomenon of interest through reviews, interviews, or surveys, to evaluate whether the (conceptual) simulation model and its results (input-output relationships) are reasonable.
Comparison to Reference Behaviors	Compares the simulation output results against trends or expected results often reported in the technical literature. It is likely used when no comparable data is available.
Comparison to Other Models	Compares the results (outputs) of the simulation model being validated to results of other valid (simulation or analytic) model. Controlled experiments can be used to arrange such comparisons.
Event Validity	Compares the “events” of occurrences of the simulation model to those of the real phenomenon to determine if they are similar. This technique is applicable for event-driven models.
Historical Data Validation	If historical data exists, part of the data is used to build the model and the remaining data are used to compare the model behavior and the actual phenomenon. Such testing is conducted by driving the simulation model with either sample from distributions or traces, and it is likely used for measuring model accuracy.
Rationalism	Uses logic deductions from model assumptions to develop the correct (valid) model, by assuming that everyone knows whether the clearly stated underlying assumptions are true.
Predictive Validation	Uses the model to forecast the phenomenon’s behavior, and then compares the phenomenon’s behavior to the model’s forecast to determine if they are the same. The phenomenon data may come from the real phenomenon observation or be obtained by conducting experiments, e.g., field tests for provoking its occurrence. Also, data from the technical literature may be used, when there is no complete data in hands. It is likely used for measuring model accuracy.
Internal Validity	Several runs of a stochastic model are made to determine the amount of (internal) stochastic variability. A large amount of variability (lack of consistency) may cause the model’s results to be questionable, even if typical of the problem under investigation.
Sensitivity Analysis	Consists of changing the values of the input and internal parameters of a model to determine the effect upon the model’s output. The same relationships should occur in the model as in the real phenomenon. This technique can be used qualitatively—trends only — and quantitatively—both directions and (precise) magnitudes of outputs.
Testing model structure and behavior	Submits the simulation model to tests cases, evaluating its responses and traces. Both model structure and outputs should be reasonable for any combination of values of model inputs, including extreme and unlikely ones. Besides, the degeneracy of the model’s behavior can be tested by appropriate selection of values of parameters.
Based on empirical evidence	Collects evidence from the technical literature (experimental studies reports) to develop the model’s causal relationships (mechanisms).
Turing Tests	Individuals knowledgeable about the phenomenon are asked if they can distinguish between real and model outputs.

This list can be used to support the identification of attempts to verify or validate a simulation model in existing SBS reports, enabling the experimenter to collect the successful attempts. Moreover, having no evidence regarding the model validity, some of these procedures should be performed, exposing the results as well as the decisions that guided the validation process.

Such procedures have been extensively discussed in the technical literature on computer simulation. Besides, we identified nine V&V procedures applied to simulation models in the context of SE (DE FRANÇA and TRAVASSOS, 2013b) and merged this list with the one presented by SARGENT (1999), which are fifteen V&V procedures often performed for DES models in several domains, excluding three useful instruments to perform V&V activities, rather than procedures or techniques. This way, the merge of the remaining twelve with the procedures identified in the qSLR are presented in Table 7-5.

None of the V&V procedures from Table 7-5 can avoid all the potential threats to simulation study validity whether used alone. However, successfully applying some of these procedures together can help to increase the confidence on simulation's results.

SG10. Example from the proof-of-concept.

In order to improve the model validity, ARAÚJO, MONTEIRO and TRAVASSOS (2012) collected evidence for each relationship amongst model variables from the technical literature. For the complete set of 22 evidence, see (ARAÚJO, MONTEIRO and TRAVASSOS, 2012). Besides, the model was successfully assessed using the procedure of Historical Validation, 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). Therefore, the model was able to predict the trends for the output variables. This is considered enough for the understanding purposes of our study.

SG11. 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 simulation model and I/O matching. It enables the investigation of internal properties and behaviors of a simulation model, like model variables, equations and relationships, rather than dealing with it as black box, i.e., observing just the I/O matching. This way, domain experts may identify threats to construct validity in advance. The threats involve the mechanisms explaining the phenomenon captured by the simulation model. For instance, experts can find out both inappropriate definition of cause-effect relationships and failure on capturing the corresponding real world building blocks and elements.

As it relies on experts' opinion, it may not be the only perspective to take into account on the simulation model validity. However, it is a relevant indication that, in face of the model representation and its generated behavior, it gives the impression of being valid. Face validity needs to be complemented by other procedures already mentioned in Table 7-5, for example.

Face validity sessions may happen on workshops, group or private interviews. The main idea is to present the model by following a walkthrough approach to show how the input values generate outcomes, exemplifying with real scenarios so that experts can realize the model behavior and validate the simulation results for a given set of inputs.

SG11. Example from the proof-of-concept.

The authors did not perform the Face Validity procedure, which we consider an important procedure to capture internal problems in the simulation model. In the model equations, it is possible to identify inconsistencies regarding variables units, unless one consider them as dimensionless.

SG12. Support model (causal) relationships, as much as possible, with empirical evidence to reinforce their validity and draw more reliable conclusions.

From an external validity perspective, it is sound to have the simulation model's causal relationships supported by empirical evidence (DAVIS, EISENHARDT and BINGHAM, 2007). Empirical evidence can support the existence of properties in the simulation model, as well as model assumptions. Besides, it reduces the modeler's bias, since such evidence does not rely only on experts' opinions or *ad-hoc* observations of the phenomenon under study. This way, secondary studies may be performed to search for evidence, if it is not known. The search may include at least the core causal relationships, since bigger models may impose a great effort on performing secondary studies, such as a systematic literature review.

SG12. Example from the proof-of-concept.

As mentioned before, ARAÚJO, MONTEIRO and TRAVASSOS (2012) undertook a qSLR review encompassing 15 research questions, being one for each pair of possible relationship among the model variables. Each question regarded the existence of such relationship, as well as its direction and intensity. A set of quality assessment criteria was defined to prioritize evidence to be considered in the model. As a result, only relationships referred by some evidence were kept in the SD model.

SG13. Always verify the model assumptions, so the results of simulated experiments can get more reliable.

Face Validity can also be combined with procedures to compare empirical data/behavior (SARGENT, 1999) to assess model's assumptions regarding the underlying concepts. For instance, using Comparison to Reference Behaviors, Historical Validation or Predictive Validation to understand if the model (including its assumptions) is capable of reproducing an empirical behavior in terms of internal variables and outcomes. However, when the model assumptions are hidden or unclearly stated, Face Validity is not applicable. In these cases, using other procedures is more suitable. The expected behaviors can give insights about how the hidden model assumptions are affecting the results.

The verification of model assumptions also appies to simplifications imposing the expected behavior (ECK and LIU, 2008). When the simulation model is to be developed, the modeler makes, even implicitly, some assumptions regarding the phenomenon. For instance, the increasing of a response variable directly caused by the presence of a given treatment. If these assumptions are embedded in the model, it may represent a threat to internal validity, since it should not be coded directly in the model, rather it should be treated as an effect of a chain of actions, events and conditions generating such behavior in the response variable.

SG13. Example from the proof-of-concept.

ARAÚJO, MONTEIRO and TRAVASSOS (2012) clearly establish their assumptions as logical formulations representing the expected behaviors (trends) for each software characteristic in the presence of a particular item of the LSE. These software characteristics' trends (increasing - ↑, decreasing - ↓ or no changing - ↔) are tested as hypotheses when some LSE can be observed or not in the software project. It is important to highlight that such trends are not directly implemented in the SD model. Rather, the simulation results may present or not the expected behavior regarding the LSE depending on the project dataset and input parameters. On conducting the model assessment, they present which reference behavior (from historical dataset) could be reproduced, as well as the laws influencing the observations. The assessment results are sufficiently positive to assume the model as valid for our purposes.

From many simulation studies found in SE, just a few of them report performance measures. Measures such as bias, accuracy, coverage, and confidence intervals frequently go unreported. The importance of such measures relies in the possibility of using them as benchmark criteria to compare and choose more accurate simulation models. In addition, this will directly affect the risks assigned to SBS conclusions. For instance, outcomes are obtained in a SBS, but if the simulation model has a low accuracy or its results are in a wide confidence interval, these results may be far from reality. This information also brings credibility to the simulation study. Burton *et al* discuss how to calculate such measures (BURTON *et al.*, 2006).

As an example of performance measures, LAUER, GERMAN and POLLMER (2010) use the relative error in mean values and confidence intervals to compare different configurations from the perspective of timing problems in the context of an automotive embedded system.

7.8 Subjects

As both *in virtuo* or *in silico* studies are under the scope of the proposed guidelines, it is important to characterize subjects, no matter if they are human or not. Human subjects may influence the interpretation of *in virtuo* results. For this reason, characteristics such as the level of expertise, domain knowledge and background should be captured, as well as, the number of subjects per group (treatment and control, when applicable) and any other relevant characteristic affecting the results should be addressed in the subjects' assignment process to the experimental units whether made randomly or not, for example.

SG14. Characterize the subjects involved in the simulation study as well as their training needs.

Additionally, the training sessions for human subjects and their costs should be planned and reported as well. With computerized subjects, their behavior, configuration and parameters should also be considered when designing the experiment, in case that such behavior can be clearly identified in the simulation model. However, it is possible the behavior being implicitly embedded in the simulation model when dealing with *in silico* environments. PFAHL, KLEMM and RUHE (2001) present an example of subjects' description in the experiment on software project learning, involving twelve computer science graduate students, who were enrolled in the advanced software engineering class lasting one semester. Besides, they captured information about personal characteristics, education, background regarding experience in software development and project management, software project management literature background, and preferred learning style.

SG14. Example from the proof-of-concept.

As the proof-of-concept regards an *in silico* experiment in which the model focuses on software characteristics, 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.

7.9 Experimental Design

Experimental design issues involve the definition of a causal model establishing a relationship between independent (or factors) and dependent variables, in a cause-effect nature. Research goals and questions should drive the causal model definition, by taking the model part that reflects the concerns in the goal and the variables that can help on answering the research questions. Here, the importance of describing the model and its variables is clear (see section 7.6). Once they are described, the experimental design can be more easily understood.

As seen in (BALCI, 1990), different parameters (input variables), behavioral relationships, and auxiliary variables may represent model variants, since they constitute the statistical design factors. During the experiment execution, the design factors may be held constant or allowed to vary. Therefore, interest factors may be: controllable, which are possible to measure and vary; uncontrollable, possible just to measure; and noise factors, the ones we cannot measure and they naturally vary.

According to Montgomery (MONTGOMERY, 2008), levels (or treatments) correspond to the range of interest over which the factors will be varied. This way, the experimenter should have practical experience and theoretical understanding on the domain.

SG15. Describe the experimental design (design matrix), including independent and dependent variables and how levels are assigned to each factor.

The experimental design is often fully described by a **design matrix**. In this matrix, every row is called a design point or a scenario, which is a combination of different levels of each factor. However, there are several different designs that can be generated for the same set of factors. In Statistics, there is a mature discipline called Design of Experiments (DOE). We have no ambition to contribute with DOE, but to bring such knowledge and apply it to SE simulation experiments, considering the particular 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. Although 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. Besides, KLEIJNEN *et al* (2005) claim that DOE for simulation experiments is different since in simulation we are not limited by real world constraints.

Factorial designs are the most recurring 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 2^k design, meaning the number of scenarios needed to determine effects from *k* factors and their interactions.

There are also variants proposed for large number of scenarios in which the simulation runs are time consuming, since time grows exponentially with the number of factors and levels. Therefore, it is possible to reduce the number of scenarios, but still having 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 determined also 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 an exhaustive list:

- Simulation goals, since designs for understanding or characterization are not the same for comparisons or optimizations;
- Experimental frame, whether the area of interest is local or global, and it impacts in the range of levels;
- Number of factors and levels, since they exponentially increase the number of scenarios in full factorial designs;
- Domain of admissible scenarios, it is important since full factorial designs may generate inadmissible scenarios;
- Simulation model's deterministic and stochastic components, since 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.

SG15. Example from the proof-of-concept.

In the proof-of-concept, the variables 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 adopt low, medium and high values, to understand how the response variables behave by increasing the periodicity. The level differences intend 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 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. Conversely, variable-duration means that each cycle may have a different periodicity.

In the causal diagram on the right side of Figure 7-2, it is possible to identify first-order and other higher-order possible effects of Periodicity on both Reliability and Maintainability. Therefore, in this experiment we will explore full factorial designs for questions Q1 and Q2 as shown in the design matrix (Table 7-6). Each scenario in this matrix corresponds to a possible combination of factors and their levels.

Table 7-6. Design matrix for the simulation experiment (DE FRANÇA and TRAVASSOS, 2015)

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

SG16. Use Sensitivity Analysis to select valid parameters' settings when running simulation experiments, rather than model "fishing".

Techniques such as Sensitivity Analysis are useful when selecting the groups of interest factors and levels range. Once the more sensible factors are determined, the number of levels for each factor and the values they assume can be properly defined. Furthermore, a systematic way of defining the levels reduces the bias and avoids the fishing for positive results. For characterization studies, it is recommended to keep a low number of levels per factor, but covering a high region of interest (MONTGOMERY, 2008).

SG16. Example from the proof-of-concept.

The experimental design presented in Table 7-6 includes the principle of Sensitivity Analysis, particularly for the factor Periodicity, in which we established scenarios using low and high values, as well as small (from 2 to 10) and large (from 10 to 40) variations, in both deterministic and stochastic scenarios. It enables the observation of the model behaviour in a large space of possibilities, and characterizing the effects of each factor on the output variables.

In addition, it is important to identify control and treatment groups when performing 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.

SG17. Consider as factors (and levels) not only the simulation model's input parameters when designing the simulation experiment, but also internal parameters, different sample datasets and simulation model versions, implementing alternative strategies to be evaluated.

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.

In our simulation experiment, this guideline is not applicable since the required factors are mainly input parameters. However, GAROUSI, KHOSROVIAN and PFAHL (2009) present a simulation experiment, using a SD model for software processes, aiming at comparing scenarios representing combinations (and intensity levels) of development, verification, and validation techniques that should be applied in a given context to achieve defined time, quality, or cost targets. In their experiment, the authors use two calibrations based on data from the technical literature to derive the scenarios. These datasets differ on intensity of rework effort for faults detected in integration and system testing.

Clearly, the use of scenarios in simulation experiments can be viewed as consequence of selecting a proper experimental design. However, it can also be a cause of it, since it is common to make use of scenarios even when an *ad-hoc* experimental design is adopted. In this case, the experimenter plans the scenarios of interest (BARROS, WERNER and TRAVASSOS, 2000) and then derivate the design. By adopting the last strategy, the relevance and adequacy of each chosen scenario should be explained and tied to the study goals.

SG18. When adopting ad-hoc designs determine the selected simulation scenarios and explain the criteria used to identify them as relevant.

Representative scenarios, including those that both check best and worst cases, can help foreseeing behaviours in regular and exceptional circumstances. The scenarios description needs to be as precise as possible, clarifying all the relevant contextual information, as well as input parameters values for the scenarios.

The main drawback of achieving *ad-hoc* the experimental design is the potential embedding of some bias, especially for non-experienced experimenters, 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 providing successful results (KLEIJNEN *et al.*, 2005) such as Central Composite Designs, Sequential Bifurcation and Latin Hypercube Sampling.

As we adopted DOE for defining our design matrix, this guideline is not applicable. However, AMBROSIO, BRAGA and RESENDE-FILHO (2011) use three scenarios (optimistic, baseline, and pessimistic) in two sets of simulations by changing the value of model components related to risk factors in a model concerned with requirements activities: requirements errors and volatility, and workforce turnover. These scenarios are described as three different model input parameters settings.

The number of simulation runs depends 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 levels are assigned to specific factors. The more simulation scenarios involved in the study, the more simulation runs are needed. For instance, factorial designs usually require one simulation run for each combination of factors and treatments, in case of dealing with deterministic simulation models. So, if three factors and two treatments are considered, we have a design 2^3 with 8 simulation runs required. A detailed discussion on how to determine the number of simulation runs, based on factorial designs for deterministic models, can be found in (HOUSTON *et al.*, 2001) and (WAKELAND, MARTIN and RAFFO, 2004).

SG19. When dealing with simulation model containing stochastic components, determine the number of runs required for each scenario, along with its rationale, in order to capture the phenomenon variance.

Simulation models containing stochastic components naturally produce an intrinsic noise in the output, due to the pseudo-random number generators. Thus, one single run of each scenario using those stochastic components cannot reveal the amount of variance in this noise. On the other extreme, the greater the number of runs (replications), the greater the approximation of a desired accuracy level. 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). Thus, given the required accuracy level and a sample estimate from few model runs, it is possible to determine the number of required runs and avoid this threat to conclusion validity. Such procedure for calculation can be found in (LAW and KELTON, 2000).

SG19. Example from the proof-of-concept.

For the scenarios concerning with fixed-duration strategies, the model behaves deterministically, and therefore we need just three runs, one for each periodicity level. Conversely, 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, we use 100 runs for each one of the 3 scenarios, being a total of 300 runs for the variable-duration scenarios. For each simulation scenario, we defined an output dataset, resulting in six datasets. The simulation runs were executed in the Vensim PLE environment, by explicitly setting the input parameters for each scenario.

7.10 Supporting data

When conducting SBS it is important to check the availability of supporting data. Simulation models need to be calibrated, requiring data for the generation of equations and parameters, and to determine the distribution of random variables. Therefore, it is important to determine the type of data: real or synthetic ones (ÖREN, 1981). If synthetic data has been used, some evidence should be presented to guarantee data's validity, i.e., the report should answer questions such as 'How far the simulated data is from real-system data?' and show indicators of this gap. Here, statistical tests can be applied to verify how close both real and synthetic samples could be.

SG20. Assess, whenever possible, the data used to support the simulation model development or experimentation.

SG20. Example from the proof-of-concept.

In the performed simulation experiment, the data and procedure used for model calibration came from (ARAÚJO, MONTEIRO and TRAVASSOS, 2012). This historical dataset encompasses 13 different system releases. The data was collected from three different sources: version control system logs, bug tracking services, and effort registration spreadsheets. Such measurements are relevant for the observation of system evolution, and for each release, they comprehend measures for the six variables mentioned in section 7.6.

Planning the data collection also avoids measurement mistakes, promoting the collection of data as soon as they are made available for the target model variables and to capture the contextual information associated to the quantitative data are relevant for SBS.

SG21. Keep track of contextual information (including qualitative data) along with quantitative data.

The contextual data is important to provide better and accurate reasoning when performing output analysis and interpretation, supporting the explanations based on it.

SG21. Example from the proof-of-concept.

For instance, in the performed experiment, it is important to mention some contextual information since they explain trends in the simulation results. The system releases resulted from corrective, adaptive and perfective maintenance activities. The perfective maintenance mainly regards, in this dataset, the enhancements concerning security, performance, maintainability, and graphical user interface. No new functionality was considered during these releases. Therefore, our simulation results are limited to these types of maintenance. Furthermore, the system's users reported the corrected defects for each release, during the system's operating lifecycle. This way, data regarding defects does not consider remaining defects.

SG22. 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 need 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 generalize behaviors to different contexts inappropriately. The use of cross-company data is an example of how it can impose a threat to internal validity on the simulation results.

After the collection, quality assurance procedures ought to take place in order to verify their consistence 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 has been calibrated or not, including the procedure used to accomplish the task and its results.

Simulation models often require time-sensitive data. Hence, in order to avoid biased observations and an exposure to risk (i.e. undetected seasonal data); the data collection time period should represent both transient and steady state behaviours.

SG22. Example from the proof-of-concept.

In the performed simulation experiment, we determined the input parameters from the empirical data used for the model calibration. In other words, these values and distributions are in the same scope of the observed ones, representing feasible scenarios for the observed phenomenon. Besides, the data collected for the simulation of software evolution requires time-sensitive and real-system data. In this model, the time when the data is collected is important since it is desirable to observe how the successive maintenance cycles influence the software quality. The study presents the observations made over a 2-year software project executed in the industry. Therefore, real data was treated and analysed accordingly.

Another important aspect related to the collected data (or used datasets) relies on the raw data publication. However, it is rarely reported, for two reasons: (1) most papers report that it was not possible to present the raw data as it is confidential and (2) since simulation studies usually involve a large amount of data and it may not fit in conference or journal papers. Even so, the raw data should be reported when possible or make available by consulting the authors or publishing it at a downloadable source.

7.11 Simulation supporting environment

The simulation environment consists of all instruments needed to perform the study. It encompasses the simulation model itself, datasets, data analysis tools (including statistical packages), and simulation tools/packages. As the simulation model and datasets have already been discussed, here the supporting tools are the focus as an important feature to be considered.

SG23. Set up and describe the simulation environment, including the supporting tools, associated costs, and decision for using a specific simulation package.

Ideally, the simulation package should support not only the underlying simulation approach, but also the experimental design and output data analysis. Simulation packages often differ on how they implement the simulation engine mechanism. Therefore, it is possible to get different results depending on the engine’s implementation. Moreover, the process used to translate the conceptual simulation model description to the simulation language offered by the package should be considered. Information should also be provided on how such translation was performed and if any model characteristic could not be implemented due to technological constraints. In stochastic models, the use of random number generators and on how the starting seeds were selected is fundamental.

The choice of a simulation package should depend on the fit of the research questions, assumptions, and the theoretical logic of the conceptual model with those of the simulation approach (HOUSTON *et al.*, 2001). It is an important decision as the simulation approach may impose a theoretical logic, type of research questions, or related assumptions.

Raw input data always requires an extra effort to understand its properties (such as data distribution and shape, trends, and descriptive stats) and perform the transformations (such as scale transformations and derived metrics) needed to fit the model parameters and variables. Similarly, the simulation output data needs specific analysis techniques such as statistical tests and accuracy analysis. For both input and output data there is a need for other supporting tools like statistical packages or even specific ones. These tools compose the whole simulation environment in case.

Another important perspective is related to the computational infrastructure. The settings to run the simulations need to be settled up and reported so that one can understand the requirements for replicating the study. Processor capacity, operating system, amount of data, and execution time interval are relevant characteristics to estimate schedule and costs for the study.

Finally, SBS involving multiple trials and runs often needs to summarize information from each intermediate trial for the final output analysis. Mean and standard deviation are common measures for this purpose and determine confidence intervals, for instance. This way, the individual information (measures) is stored in a database or external files. In addition, the experimenter has to concern with how such data can support the analysis, whether on charts or used as threshold values to support decision-making, for instance.

SG23. Example from the proof-of-concept.

The simulations in the proof-of-concept are executed in the Vensim environment (www.vensim.com), 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. We also adopted spreadsheets to support output analysis.

SG24. Determine which and how intermediate measures are stored among simulation trials to be used in the final analysis.

Specific or customized simulation environments should be concerned with these capabilities, since commercial tools already support it. For instance, the Vensim PLE version does not support multiple runs automatically. Therefore, we needed to store the results from each simulation run as an output file and then imported altogether in a spreadsheet to consolidate the analysis.

7.12 Output Analysis

As simulation runs generate a considerable amount of data, involve complex relationships among variables and possibly spread over different output variables, it is possible to identify prior to the execution what are the statistical procedures and instruments to support the output analysis, as it regards the understanding and quantification of simulation results.

SG25. Determine which statistical procedures and instruments support the output analysis, as well as the underlying rationale, quantifying the amount of internal variation embedded in the (stochastic) simulation model to augment the precision of results.

The simulation study protocol includes definitions concerning with the procedures and instruments to be used in the analysis of simulation results. Common output analyses for simulation experiments include main and interaction effects among factors, simulation confidence and accuracy, quantifications of variance (in case of stochastic simulation) and comparisons with reference behaviors or alternative system configurations. For that, statistical charts and tests, along with descriptive stats can help, but for every instrument, there are assumptions and restrictions that have to be assessed in the output data like normally distributed data, independent samples, and homogeneous variance. Moreover, simulation experiments use such statistical measures for accuracy indicators, for instance. Mean Magnitude of Relative Error and Balanced Relative Error are examples of such measures (FOSS *et al.*, 2003). Charts often assume 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. Additionally, evidence supporting these properties should be given.

It is also important to take care of the perspective of the analysis, whether it is 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 measures such as mean, standard deviation, and confidence intervals across replications, but not within a replication.

SG25. Example from the proof-of-concept.

For output analysis, statistical charts are used, namely histograms and sequence run charts, to characterize response variables behaviors. 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 analyze 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 *Periodicity* factor constant in the base value, as it is not a variable of interest for this research question. Similar analyses are done with other factors or interactions concerning with 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.

SG26. Be aware about data validity when comparing actual and simulated results: compared data must 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] or it can predict real-world results [empirical external validity] (ECK and LIU, 2008). Threats to external validity can also appear as context-dependent results, since there is a need for calibration and simulation model not based on empirical evidence.

7.13 Threats to Validity

SBS protocols need, as any other empirical studies, to mitigate and discuss possible threats to the study validity. Common types of experimental validities are closely related to the simulation model validity (DE FRANÇA and TRAVASSOS, 2014b). Therefore, such model should be valid to assure 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 by WÖHLIN *et al* (2012). Most of them have to be considered when conducting simulation studies, especially considering *in vitro* experiments, in which the human nature may impose risks to the study. Still, new situations emerge from *in silico* experiments. Either recognized threats appear in a different outlook, or specific threats of such environment affect the results validity. Here, we concentrate our perspective on these new situations.

RAFFO (2005) and GAROUSI, KHOSROVIAN and PFAHL (2009) consider model validity in a similar way. They contemplate several perspectives, 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 (DE FRANÇA and TRAVASSOS, 2014b). It is also important to consider the simulation experiment design, for example.

For the discussion about threats to validity, we categorize them as in (WÖHLIN *et al.*, 2012): conclusion, internal, construct and external validity. Several potential threats that need to be verified before the simulation experiment can be found in section 5.3.

SG27. Consider to check for threats to the simulation study validity before running the experiment and analysing output data to avoid bias, as well as to report non-mitigated threats, limitations and non-verified assumptions.

According to DAVIS, EISENHARDT and BINGHAM (2007), simulation tends to improve both the *construct* and internal validity, by accurately specifying and measuring constructs (and the relationship among them) and the theoretical logic that is enforced through the discipline of algorithmic representation in software, respectively. However, it is possible to observe threats to construct validity into the context of SBS, such as inappropriate cause-effect relationships definition, real-world representation by model parameters and model calibration data and procedure, hidden or invalid underlying model assumptions regarding model concepts, and the simulation model not capturing the corresponding real world building blocks and elements.

Regarding external and conclusion validity, they can be accomplished by reproducing empirical behaviour, and applying adequate statistical tests over the model outputs, respectively. However, the conclusion validity also relates to design issues like sample size, number of simulation runs, model coverage, and the degree of representation of scenarios for all possible situations.

SG27. Example from the proof-of-concept.

As a general limitation, the model adopted in the proof-of-concept has a perspective abstracting the process-level details and presents only the behavior of continuous variables involved in its causal model. Hence, 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 behavior. However, we can also see both short enhancement cycles and long correction cycles in the initial dataset. This kind of behavior is suppressed by the trends, which are obtained by linear regression to generate the model equations, and not generated by the model. Thus, it represents an external validity threat.

7.14 Conclusions and future work

In order to draw solid conclusions, the experimenter should be aware about the model and data validity, as well as non-mitigated threats to validity, to limit the findings. The reasoning should establish a link from the goals, using the output analysis procedures, to the findings. In other words, the results need to be explained through a chain of decisions and performed steps that generate the outcomes, including the reason for these results reflecting the simulated phenomenon.

SG28. Main results/findings should be identified and summarized, as well as the conclusions arising from results.

Furthermore, it is important to discuss the implications about the applicability of the solution in real scenarios, e.g., use in practice. The experimenter should realize how to implement the solution, as well as the required knowledge, capabilities and training needed. In addition, the associated risks in adopting the solution should be explicitly stated. The risks relate to contextual description (facets in SG2), so it means that changes do occur not only in processes and methods, but also with personnel, IT infrastructure, financial costs, need for consultancy, and so on.

SG29. Applicability issues should be addressed in the report, considering organizational changes and associated risks.

Finally, the way ahead should be mentioned in the report, pointing out further work and research challenges. It may also include hot topics and possible roadmaps for future research.

SG30. Point out future research directions and challenges after current results.

Mainly, future directions consist in new or refined goals, detailing a specific phenomenon by including new variables and relationships. The investigation of particular cases or multiple cases can be addressed too. Additional validation can be required as well, concerning with different types of validity not covered before.

SG28 and SG29. Example from the proof-of-concept.

As mentioned in Section 7.9, the experiment requires 303 simulation runs to evaluate all the planned scenarios. After running these simulations, we could observe exclusively at the context of this project dataset, that:

Shorter maintenance cycles lead to greater reliability. As Figure 7-3 shows, the shorter periodicity scenario (1) has a higher number of corrected defects over six months. This result is explained by two main reasons: shorter cycles are mostly related to corrective and adaptive 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.

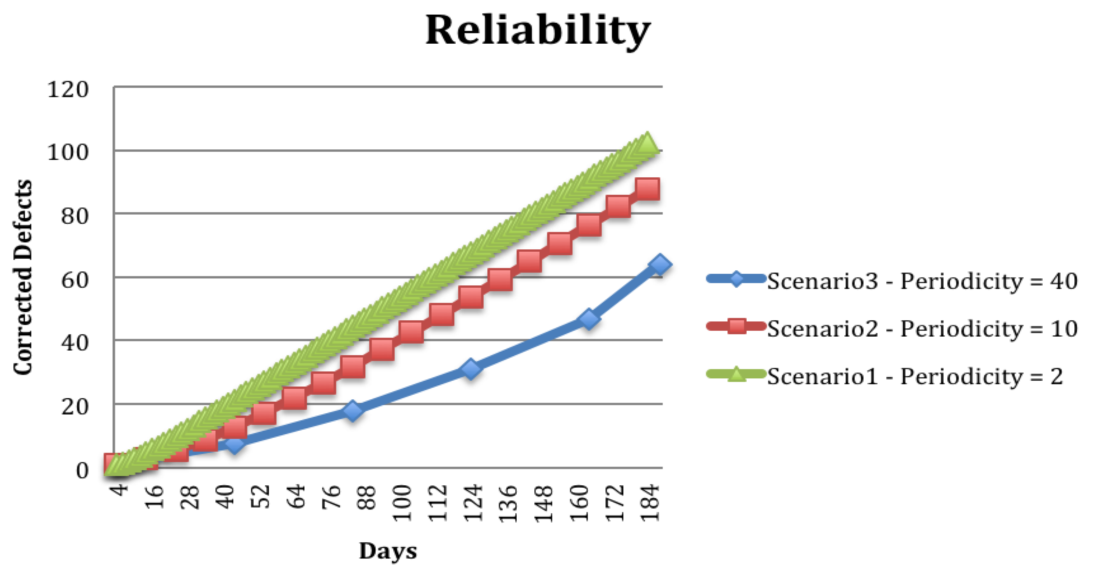


Figure 7-3. Reliability output for Fixed-Durations

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.

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. Thus, again, if there were 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. Instead, it should be clear 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 behavior for the variable-duration strategy. This behavior 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.

7.15 Conclusions of this Chapter

The proposed set of simulation guidelines presented in this chapter embraces different stages of the SBS lifecycle. Intentionally, the scope share common aspects with other research strategies, such as controlled experiments, case studies or action research. However, in these guidelines we discuss and present examples for these aspects under the simulation and SE perspectives. Additionally, it is possible to identify similar concerns in other simulation-related works already mentioned in this Thesis (Section 3.2). Nevertheless, the simulation guidelines originated in the former planning perspective (presented in Chapter 5) add a new perspective on the mitigation of validity threats that has not been presented in the technical literature before.

We recognize the need for more evaluation on this set, including both experimental studies and application on simulation experiments. Furthermore, our expectation is to spread this set of simulation guidelines over the SE community and to get feedback from its application on actual SBS as well as discuss possible improvements to evolve the knowledge and benefit more from the experimentation with dynamic simulation models in SE, as it has been started with (DE FRANÇA and TRAVASSOS, 2015).

8 Conclusions and Future Work

In this chapter, we present the conclusions, emphasizing the main contributions of this research. Additionally, we present some limitations and open questions not addressed in this thesis. Finally, the way ahead is outlined in order to show possible future work to come as result of the current achievements.

8.1 Final remarks

In this Thesis, we presented the organization, evaluation and evolution of a set of guidelines concerned with the reporting and planning of simulation experiments in Software Engineering. Such guidelines were organized based on evidence acquired through a secondary study (*quasi*-Systematic Literature review), evolved with the results of the conducted primary studies and information from other research areas. These guidelines concentrate on how conventional aspects of empirical studies should be considered when conducting simulation experiments in SE. Moreover, the concerns regarding the simulation model and study validity are justified by the importance that such model assumes (main observational instrument) and the bias promoted by the experimental design over the interpretation of results.

The motivation for simulation guidelines emerged from the opportunity to promote the quality on reported simulation studies in SE, since it is one of the issues identified in the qSLR (Chapter 2). Additionally, we reinforce that the issues revealed in the previous characterization of SBS in the context of SE are still present in the studies so far reported in the technical literature (Section 4.4). These simulation guidelines can help authors, researchers interested in simulation results, practitioners, and reviewers, on which information should be presented when reporting SBS in the context of SE. As far as we are aware, this is the first set of simulation experiments guidelines in the context of SE.

The contextual and planning information suggested by the guidelines motivate the software engineers to observe specific features when planning simulation studies in SE. Researchers and practitioners can recognize core information concerning the SBS results that may be applicable to their interests. Reviewers, members of conference programs and editorial boards of journals need to identify the relevant contributions, as well as the evidence confirming the contributions and the possible limitations of the SBS.

Besides the possible overlap to some extent with other disciplines and research strategies, the guidelines suggest how they should appear in SE studies. Some particularities can be observed since SE, at least as a science field, is not mature yet. Examples of these particularities include lack of knowledge about relevant factors and variables for a given phenomenon, both quantitative and qualitative nature of SE phenomena, and the relevance of social and technical aspects involved.

In general, the current set of guidelines organization intends to provide a logical sequence, by specifying the next step in a straightforward way, which supports the organization of research protocols and reports for SBS in the context of SE. Such sequence allows a reasonable reasoning flow from goals to output analysis, through discussions involving experimental validity, which can help according to the experience in the decision-making. The different evaluations performed and evidence used to evolve the set of simulation guidelines enhanced its quality under different perspectives and enabled its application on situations that indicate the feasibility.

8.2 Contributions of this research

The contributions of this research include aspects regarding computer simulation and experimental software engineering. Mainly, these contributions are listed below:

- Organization of a body of knowledge regarding SBS in the context of SE: the characterization using the SLR methodology enabled the identification of aspects hampering the understanding of SBS in the context of SE (DE FRANÇA and TRAVASSOS, 2012) (DE FRANÇA and TRAVASSOS, 2013b);
    - Identification of methodological issues and challenges that need to be addressed in future SE simulation research regarding experimental design and simulation output analysis;
    - Identification of V&V procedures applied to SE simulation models: it allowed us to understand which initiatives support the verification and validation of SE simulation models, and how much these attempts contribute for the validity of studies results.
    - Identification and synthesis of potential threats to simulation studies validity, which are somewhat described in simulation reports, however as general limitations, and not discussing their consequences for the study (DE FRANÇA and TRAVASSOS, 2014b) (DE FRANÇA and TRAVASSOS, 2015);
- Organization of a set of guidelines to support reporting and planning of simulation experiments in the context of SE (DE FRANÇA and TRAVASSOS, 2015), involving:

    - Identification of relevant information to compose a SBS report in the context of SE (DE FRANÇA and TRAVASSOS, 2012) (DE FRANÇA and TRAVASSOS, 2014a);
    - Consolidation of the list of V&V procedures by extending the ones presented in (SARGENT, 1999) with the procedures identified through the qSLR (Table 7-5);b
    - Extension of the indicators of simulation feasibility presented in (BALCI, 1990);
    - Analysis on how existent DOE techniques and V&V procedures can support the mitigation of potential threats to simulation experiments validity (DE FRANÇA and TRAVASSOS, 2014b) (DE FRANÇA and TRAVASSOS, 2015).
- All versions of the simulation guidelines include evaluations to enhance them and to support its validity (DE FRANÇA and TRAVASSOS, 2015), from which we highlight:
    - The survey (Section 4.3) involving simulation and software engineering researchers regarding guidelines completeness and correctness. The assessment performed by experts with heterogeneous background reinforced the relevant aspects for SBS;
    - The observational study (Chapter 6) as a different approach to evaluate this sort of technology, as the ones previously presented in the technical literature are mostly based on expert opinion (KITCHENHAM *et al.*, 2008) and application (RUNESON and HÖST, 2009). Such evaluation indicates the usefulness of the proposed set of simulation guidelines (DE FRANÇA *et al.*, 2015).
    - The proof of concept (presented as examples in Chapter 7), which allowed us to think over the proposed guidelines application and represented an actual simulation experiment concerning the planning of software maintenance cycles.

8.3 Limitations

In spite of the contributions and interesting results, there are still some limitations that are important to be mentioned:

- By design, the scope for the simulation guidelines does not cover modeling issues, as we understand there are enough contributions in this direction, even for the SE context. However, we understand that existing orientation may not be easy to apply for large and complex models, requiring more investigations.

- The simulation guidelines did not discuss details on how to apply or execute established procedures such as Sensitivity Analysis, generation of specific design matrices and calculation for number of required runs. For that, their original references describe how to perform them. The interesting point is how they contribute for the quality of SBS protocols and when they should be applied.
- The first evaluation (Section 4.2) performed for the preliminary version of the reporting guidelines has a threat to validity that regards to the reviewers being the authors of the guidelines. It occurred due to the unavailability of human resources in position to perform the review with the required expertise. However, the aspects were evaluated according to the original checklists delivered by the evaluation approach proposed by (KITCHENHAM *et al.*, 2008), which established the perspective with unbiased questions. Moreover, we evaluated the same aspects in the survey, in which experts reviewed the improved version;
- The survey's scope encompassed only reporting guidelines and part of the planning ones (Section 4.3), since the rest of the planning guidelines did not exist at that time;
- The existence of a journal publication (CHERNOGUZ, 2011) for the adopted simulation model in the observational study clearly influenced the perception of validity by the subjects. Therefore, the study protocol needs adjustments for future trials in order to mitigate such bias.

8.4 Open Questions and Future Work

In the context of this research, some questions remain open and, consequently, they are candidates for future research. Some regard our initial research questions, which rely on the need for more evaluation studies, and others arise as consequences of the knowledge evolution. Thus, it is important to present them clearly.

How general or specific are the proposed guidelines? All the effort regarding investigation and analysis was concentrated on issues identified in the context of SE. It is possible that other science areas share common concerns as we adapted some knowledge outside SE. However, this work has no ambition to propose general guidelines for simulation regarding phenomena from other areas. Additionally, we understand that most of the issues discussed in the guidelines are related to the immature stage of

SE research regarding experimental aspects and that affects the way studies are conducted. Therefore, the proposed guidelines are meant to give a broad orientation on relevant aspects to concern with regarding experimentation with simulation models without getting into the specifics of each technique, method or procedure that may be adopted.

How can we evaluate the quality of evidence from simulation studies? Regarding our research questions *Q₂*, concerned with study validity, we identified potential threats to SBS validity, as well as analyzed how V&V procedures for simulation models, experimental designs techniques and output analysis instruments applicable for simulation experiments can mitigate these threats. However, we have no criteria to evaluate the quality of the evidence resulting from SBS. Initially, the proposed guidelines could be used as input to propose a small set of criteria, but it certainly requires more investigation;

Which process to follow when conducting simulation-based studies? The proposed set of guidelines for simulation does not meant to be a process or methodology to perform SBS. The idea is to cover specific topics not covered by existing approaches, focusing on model experimentation and validity. In fact, there are some lifecycles already presented on the technical literature (Section 1.1) concerning the whole process in an abstract way. Besides, methodologies for software process simulation and modeling are also available in the technical literature (PFAHL and RUHE, 2002) (ALI and PETERSEN, 2012), although they focus on model development.

How to deal with specific steps on the simulation lifecycle? Processes for selecting the suitable simulation approach, V&V procedure, experimental designs or analysis instruments are beyond the purpose of the guidelines. However, they configure areas of interest for many research areas and have been investigated over the years. The specificities of SE domains (such as software process) or simulation approaches (such as System Dynamics) are not covered either, as this work has a general purpose in this sense.

How could simulation support software engineering research? A part from performing regular *in virtuo* and *in silico* experiments, simulation have been used to support experimentation in other forms. It is the case of theory development, as discussed in (DAVIS, EISENHARDT and BINGHAM, 2007), in which simple theories are initially described as simulation models. Later, such theory evolves based on the results of experiments encompassing simulation scenarios to test hypotheses regarding the theory under development or establishing scenarios to falsify the proposed theory. Additionally, simulation can support feasibility studies for SE technologies and attempts to observe phenomena under large-scale perspective. For instance, the investigation of how many events should occur before a certain behavior become observable, or the limits for the application of SE technologies in terms of scale (organization size, project duration, number of development sites, functional size or complexity, among other scale variables). In order to analyze these phenomena, there are several limitations regarding observation capacity, which is unfeasible in many *in vivo* or *in virtuo* settings.

Finally, simulation in the SE context is not restricted to software process simulation. Product simulation is also interesting but scarce. For instance, simulation of architectural issues regarding security attacks and performance limits are required for large and complex systems that demand time to be ready for testing such attributes. In addition, peopleware behavior in the software development is also interesting and relevant. The possibilities of simulating the impact of motivational and exhaustion factors, development team dynamics and human resource allocation policies illustrate this point.

9 REFERÊNCIAS BIBLIOGRÁFICAS

ABDEL-HAMID, T. Understanding the “90% syndrome” in software project management: A simulation-based case study. Journal of Systems and Software, 8, 1988. 319-330.

ABDEL-HAMID, T. K.; MADNICK, S. E. Lessons learned from modeling the dynamics of software development. Communications of the ACM, 32, n. 12, 1989. 1426-1438.

ABDEL-HAMID, T.; MADNICK, S. E. Software project dynamics: an integrated approach. Upper Saddle River, NJ, USA: Prentice-Hall, Inc, 1991.

AHMED, R.; HALL, T.; WERNICK, P.; ROBINSON, S.; SHAH, M. Software process simulation modelling: A survey of practice. Journal of simulation, 2, n. 2, 2008. 91-102.

AL-EMRAN, A.; JADALLAH, A.; PAIKARI, E.; PFAHL, D.; RUHE, G. Application of Re-estimation in Re-planning of Software Product Releases. Paderborn, Germany: Springer Berlin Heidelberg, 2010. p. 260-272.

ALEXOPOULOS, C.; SEILA, A. F. Output data analysis. In: BANKS, J. Handbook of Simulation. Atlanta, Georgia, USA: Wiley-Interscience, 1998. p. 225-272.

ALI, N. B.; PETERSEN, K. A consolidated process for software process simulation: State of the art and industry experience. 38th EUROMICRO Conference on Software Engineering and Advanced Applications (SEAA). Cesme, Izmir, Turkey: IEEE. 2012. p. 327-336.

ALI, N. B.; PETERSEN, K.; WÖHLIN, C. A systematic literature review on the industrial use of software process simulation. Journal of Systems and Software, 97, 2014. 65-85.

ALVAREZ, F.; CRISTIAN, G. Applying simulation to the design and performance evaluation of fault-tolerant systems. Proceedings of the IEEE Symposium on Reliable Distributed Systems. Durham: IEEE. 1997. p. 35–42.

AMBROSIO, B.; BRAGA, J.; RESENDE-FILHO, M. Modeling and scenario simulation for decision support in management of requirements activities in software projects. Journal of Software Maintenance and Evolution, 23, n. 1, 2011. 35–50.

ANDERSSON, C.; KARLSSON, L.; NEDSTAM, J.; HOST, M.; NILSSON, B. Understanding software processes through system dynamics simulation: a case study. Proceedings of the Ninth Annual IEEE International Conference and Workshop on the Engineering of Computer-Based Systems. Lund, Sweden: IEEE. 2002. p. 41-48.

ARAÚJO, M.; MONTEIRO, V.; TRAVASSOS, G. Towards a Model to Support in silico Studies regarding Software Evolution. Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement. Lund, Sweden: ACM. 2012. p. 281-290.

BAI, X.; HUANG, L.; ZHANG, H.; KOOLMANOJWONG, S. Hybrid modeling and simulation for trustworthy software process management: a stakeholder-oriented approach. Journal of Software: Evolution and Process, 24, n. 7, 2012. 721-740.

BALCI, O. Guidelines for successful simulation studies. Proceeding of the Winter Simulation Conference. New Orleans, LA: IEEE. 1990. p. 25-32.

BANKS, J. Introduction to Simulation. Proceedings of the Winter Simulation Conference. Phoenix, AZ, USA: ACM. 1999.

BARNEY, S.; PETERSEN, K.; SVAHNBERG, M.; AURUM, A.; BARNEY, H. Software quality trade-offs: A systematic map. Information and Software Technology, 54, n. 7, 2012. 651-662.

BARROS, M.; WERNER, C.; TRAVASSOS, G. Applying system dynamics to scenario based software risk management. International System Dynamics Conference. Bergen, Norway: The System Dynamic Society. 2000.

BASILI, V. R. Software Modeling and Measurement: The Goal/Question/Metric Paradigm. University of Maryland at College Park. College Park, MD, USA. 1992.

BIOLCHINI, J.; MIAN, P. G.; NATALI, A. C.; TRAVASSOS, G. H. Systematic Review in Software Engineering: Relevance and Utility. PESC-COPPE/UFRJ. Rio de Janeiro, Brazil. 2005.

BIRKHÖLZER, T.; PFAHL, D.; SCHUSTER, M. Applications of a Generic Work-Test-Rework Component for Software Process Simulation. Proceedings of International Conference on Software Process. Paderborn, Germany: Springer. 2010.

BIRTA, L. G.; ARBEZ, G. Modelling and Simulation: Exploring Dynamic System Behaviour. London, UK: Springer-Verlag London Limited, 2007.

BORGES, A.; FERREIRA, W.; BARREIROS, E.; ALMEIDA, A.; FONSECA, L.; TEIXEIRA, E.; SILVA, D.; ALENCAR, A.; SOARES, S. Support mechanisms to conduct empirical studies in software engineering. Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '14). New York, NY, USA: ACM. 2014. p. 1-4.

BROOKS, F. P. The mythical man-month. Reading, MA: Addison-Wesley, 1975.

BURTON, A.; ALTMAN, D.; ROYSTON, P.; HOLDER, R. The design of simulation studies in medical statistics. Statistics in medicine, 25, n. 24, 2006. 4279-4292.

CARVER, J. Towards Reporting Guidelines for Experimental Replications: A Proposal. RESER. Cape Town, South Africa: ACM/IEEE. 2010.

CELIK, N.; XI, H.; XU, D.; SON, Y. Simulation-based workforce assignment considering position in a social network. Proceedings of Winter Simulation Conference. Baltimore, MD, USA: IEEE. 2010. p. 3228–3240.

CHERNOGUZ, D. G. The system dynamics of Brooks' Law. Simulation: Transactions of the Society for Modeling and Simulation International, 2011. 1-29.

COLUCCI, E. "Focus Group Can Be Fun": The Use of Activity-Oriented Questions in Focus Group Discussions. Qualitative Health Research, 17, n. 10, 2007. 1422-1433.

CONCAS, G.; LUNESU, M.; MARCHESI, M.; ZHANG, H. Simulation of software maintenance process, with and without a work-in-process limit. Journal of Software: Evolution and Process, 25, n. 12, 2013. 1225-1248.

COOK, T. D.; CAMPBELL, D. T.; DAY, A. Quasi-experimentation: Design & analysis issues for field settings. Boston: Houghton Mifflin, v. 351, 1979.

CORBIN, J.; STRAUSS, A. Basics of qualitative research: Techniques and procedures for developing grounded theory. California, USA: Sage, 2008.

DAVIS, F. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS quarterly, 1989. 319-340.

DAVIS, J.; EISENHARDT, K.; BINGHAM, C. Developing Theory Through Simulation Methods. Academy of Management Review, 32, n. 2, 2007. 480-499.

DE FRANÇA, B. B. N.; RIBEIRO, T. V.; DOS SANTOS, P. S. M.; TRAVASSOS, G. H. Using Focus Group in Software Engineering: lessons learned on characterizing software technologies in academia and industry. Iberoamerican Conference on Software Engineering (CIbSE). Lima, Peru: Curran Associates, Inc. 2015.

DE FRANÇA, B. B. N.; TRAVASSOS, G. H. Are We Prepared for Simulation Based Studies in Software Engineering yet? 15th Iberoamerican Conference on Software Engineering (CIbSE/ESELAW 2012). Buenos Aires, Argentina: Curran Associates, Inc. 2012.

DE FRANÇA, B. B. N.; TRAVASSOS, G. H. Reporting guidelines for simulation-based studies in software engineering. Proceedings of the 16th International Conference on Evaluation & Assessment in Software Engineering (EASE 2012). Ciudad Real, Spain: IET Digital Library. 2012. p. 156-160.

DE FRANÇA, B. B. N.; TRAVASSOS, G. H. Reporting Guidelines for Simulation-Based Studies in Software Engineering. PESC/COPPE/UFRJ. Rio de Janeiro, p. 18. 2013a. (RT – ES 746 / 13).

DE FRANÇA, B. B. N.; TRAVASSOS, G. H. Reporting Guidelines for Simulation-Based Studies in Software Engineering. COPPE / Universidade Federal do Rio de Janeiro. Rio de Janeiro. 2014a. (ES-747/14).

DE FRANÇA, B. B. N.; TRAVASSOS, G. H. Experimenting with dynamic simulation models in Software Engineering. Empirical Software Engineering, 2015.

DE FRANÇA, B. B. N.; TRAVASSOS, G. H. Simulation Based Studies in Software Engineering: A Matter of Validity. CLEI electronic journal, 18, n. 1, 2015. Paper 4.

DE FRANÇA, B. B.; TRAVASSOS, G. H. Are We Prepared for Simulation Based Studies in Software Engineering Yet? CLEI Electronic Journal, 16, n. 1, April 2013b. Paper 8.

DE FRANÇA, B.; TRAVASSOS, G. Simulation based studies in software engineering: A matter of validity. Proceedings of the 17th Ibero-American Conference Software Engineering (CIbSE 2014). Pucón, Chile: Curran Associates, Inc. 2014b. p. 308-321.

DE MELLO, R.; DA SILVA, P.; RUNESON, P.; TRAVASSOS, G. Towards a framework to support large-scale sampling in software engineering surveys. Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '14). New York, NY, USA: ACM. 2014. p. Article 48, 4 pages.

DRAPPA, A.; LUDEWIG, J. Simulation in software engineering training. Proceedings of the 22nd international conference on Software engineering. Limerick, Ireland: ACM. 2000.

DYBÅ, T.; DINGSøYR, T.; HANSSEN, G. Applying Systematic Reviews to Diverse Study Types: An Experience Report. First International Symposium on Empirical Software Engineering and Measurement. Madrid, Spain: IEEE. 2007.

DYBÅ, T.; SJøBERG, D. I.; CRUZES, D. S. What works for whom, where, when, and why?: on the role of context in empirical software engineering. Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement. Lund, Sweden: ACM. 2012.

ECK, J. E.; LIU, L. Contrasting simulated and empirical experiments in crime prevention. Journal of Experimental Criminology, 4, 2008. 195-213.

FLORAC, W.; CARLETON, A. Measuring the software process: statistical process control for software process improvement. Indianapolis, IN: Addison-Wesley Professional, 1999.

FOSS, T.; STENSRUD, E.; KITCHENHAM, B.; MYRTVEIT, I. A Simulation Study of the Model Evaluation Criterion MMRE. IEEE Transactions on Software Engineering, 29, n. 11, November 2003. 985-995.

GAROUSI, V.; KHOSROVIAN, K.; PFAHL, D. A customizable pattern-based software process simulation model: Design, calibration and application. Software Process Improvement and Practice, 14, n. 3, 2009. 165–180.

GRIMM, V.; BERGER, U.; DEANGELIS, D.; POLHILL, J.; GISKE, J.; RAILSBACK, S. The ODD protocol: a review and first update. Ecological Modelling, 221, n. 23, 2010. 2760-2768.

HÖST, M.; WÖHLIN, C.; THELIN, T. Experimental Context Classification: Incentives and Experience of Subjects. IEEE Conference Proceedings International Conference on Software Engineering. St. Louis, USA: IEEE. 2005. p. 470-478.

HOUSTON, D.; BUETTNER, D. Modeling user story completion of an agile software process. Proceedings of the 2013 International Conference on Software and System Process. San Francisco, CA, USA: ACM. 2013. p. 88-97.

HOUSTON, D.; FERREIRA, S.; COLLOFELLO, J.; MONTGOMERY, D.; MACKULAK, G.; SHUNK, D. Behavioral characterization: finding and using the influential factors in software process simulation models. Journal of Systems and Software, 59, n. 3, 2001. 259-270.

HOUSTON, D.; LIEU, M. Modeling a Resource-Constrained Test-and-Fix Cycleand Test Stage Duration. Proceedings of International Conference on Software Process. Paderborn, Germany: Springer. 2010.

IVARSSON, M.; GORSCHEK, T. A method for evaluating rigor and industrial relevance of technology evaluation. Empirical Software Engineering, 16, n. 3, 2011. 365-395.

JEDLITSCHKA, A.; CIOLKOWSKI, M.; PFAHL, D. Reporting Experiments in Software Engineering. In: SHULL, F.; SINGER, J.; SJøBERG, D. I. K. Guide to Advanced Empirical Software Engineering. London, UK: Springer, 2008.

KITCHENHAM, B. Procedures for performing systematic reviews. Keele University. Keele, UK, p. 1-26. 2004. (33).

KITCHENHAM, B. A.; DYBA, T.; JORGENSEN, M. Evidence-based software engineering. Proceedings of the 26th international conference on software engineering. Scotland, UK: IEEE Computer Society. 2004.

KITCHENHAM, B.; AL-KHILIDAR, H.; ALI BABAR, M.; BERRY, M.; COX, K.; KEUNG, J.; KURNIAWATI, F.; STAPLES, M.; ZHANG, H.; ZHU, L. Evaluating guidelines for reporting empirical software engineering studies. Empirical Software Engineering, n. 13, 2008. 97–121.

KITCHENHAM, B.; PFLEEGER, S.; PICKARD, L.; JONES, P.; HOAGLIN, D.; EL EMAM, K.; ROSENBERG, J. Preliminary guidelines for empirical research in software engineering. IEEE Transactions on Software Engineering, 28, n. 8, 2002. 721-734.

KITCHENHAM, B.; TRAVASSOS, G.; MAYRHAUSER, A.; NIESSINK, F.; SCHNEIDEWIND, N.; SINGER, J.; TAKADA, S.; VEHVILAINEN, R.; YANG, H. Towards an ontology of software maintenance. Journal of Software Maintenance, 11, n. 6, 1999. 365-389.

KLEIJNEN, J. An overview of the design and analysis of simulation experiments for sensitivity analysis. European Journal of Operational Research, 164, n. 2, 2005. 287-300.

KLEIJNEN, J. P. C. Statistical design and analysis of simulation experiments. Informatie, 17, n. 10, October 1975. 531-535.

KLEIJNEN, J. P.; SANCHEZ, S. M.; LUCAS, T. W.; CIOPPA, T. M. State-of-the-Art Review: A User’s Guide to the Brave New World of Designing Simulation Experiments. INFORMS Journal on Computing, 17, n. 3, 2005. 263-289.

KONTIO, J.; BRAGGE, J.; LEHTOLA, L. The Focus Group Method as an Empirical Tool in Software Engineering. In: SHULL, F.; SINGER, J.; SJøBERG, D. I. K. Guide to Advanced Empirical Software Engineering. London: Springer-Verlag, 2008. Cap. 4, p. 93-116.

LAUER, C.; GERMAN, R.; POLLMER, J. Discrete event simulation and analysis of timing problems in automotive embedded systems. IEEE International Systems Conference Proceedings, SysCon 2010. San Diego, CA, USA: IEEE. 2010. p. 18 – 22.

LAW, A.; KELTON, W. Simulation modeling and analysis. 3. ed. New York: McGraw-Hill, 2000.

LUCKHAM, D.; KENNEY, J.; AUGUSTIN, L.; VERA,.; BRYAN, D.; MANN, W. Specification and analysis of system architecture using rapide. IEEE Transactions on Software Engineering, 21, n. 4, April 1995. 336-355.

MADACHY, R. J. System dynamics modeling of an inspection-based process. Proceedings of the 18th International Conference on Software Engineering. Berlin, Germany: IEEE. 1996.

MADACHY, R. J. Software process dynamics. New Jersey: John Wiley & Sons, 2007.

MARIA, A. Introduction to Modeling and Simulation. Proceedings of the 1997 Winter Simulation Conference. Atlanta, Georgia, USA: IEEE. 1997.

MELIS, M.; TURNU, I.; CAU, A.; CONCAS, G. Evaluating the impact of test-first programming and pair programming through software process simulation. Software Process Improvement and Practice, 11, 2006. 345–360.

MIAN, P. G.; TRAVASSOS, G. H.; ROCHA, A. R. C. D.; NATALI, A. C. C. Towards a Computerized Infrastructure for Managing Experimental. Actas de las Jornadas Iberoamericanas de Ingeniería del Software y del Conocimiento. Madrid: Polytechnic University of Madrid. 2004. p. 475-487.

MONTGOMERY, D. C. Design and analysis of experiments. New York: John Wiley & Sons, 2008.

MÜLLER, M.; PFAHL, D. Simulation methods. In: SHULL, F.; SINGER, J.; SJøBERG, D. I. Guide to Advanced Empirical Software Engineering. London: Springer London, 2008. p. 117-152.

ÖREN, T. I. Concepts and criteria to assess acceptability of simulation studies: a frame of reference. Communications of the ACM, 24, n. 4, 1981. 180-189.

PAI, M.; MCCULLOCH, M.; GORMAN, J. D.; PAI, N.; ENANORIA, W.; KENNEDY, G.; THARYAN, P.; COLFORD, J. M. J. Systematic Reviews and meta-analyses: An illustrated, step-by-step guide. The National Medical Journal of India, 17, n. 2, March-April 2004. 86-95.

PAIKARI, E.; RUHE, G.; SOUTHEKEL, P. Simulation-Based Decision Support for Bringing a Project Back on Track: The Caseof RUP-Based Software Construction. Proceedings of International Conference on Software and System Process. Zurich, Switzerland: Springer. 2012.

PERRY, D.; SIM, S.; EASTERBROOK, S. Case Studies for Software Engineering. Proceedings of the 29th Annual IEEE/NASA Software Engineering Workshop - Tutorial Notes. Washington, DC, USA: IEEE. 2005. p. 96-159.

PETERSEN, K. Measuring and predicting software productivity: A systematic map and review. Information and Software Technology, 53, n. 4, 2011. 317-343.

PETERSEN, K.; WÖHLIN, C. Context in Industrial Software Engineering Research. Proceedings 3rd International Symposium on Empirical Software Engineering and Measurement. Orlando, USA: IEEE. 2009. p. 401-404.

PFAHL, D.; KLEMM, M.; RUHE, G. A CBT module with integrated simulation component for software project management education and training. Journal of Systems and Software, 59, n. 3, 2001. 283–298.

PFAHL, D.; LAITENBERGER, O.; DORSCH, J.; RUHE, G. An Externally Replicated Experiment for Evaluating the Learning Effectiveness of Using Simulations in Software Project Management Education. Empirical Software Engineering, 8, n. 4, 2003. 367–395.

PFAHL, D.; LAITENBERGER, O.; RUHE, G.; DORSCH, J.; KRIVOBOKOVA, T. Evaluating the learning effectiveness of using simulations in software project management education: Results from a twice replicated experiment. Information and Software Technology, 46, n. 2, 2004. 127–147.

PFAHL, D.; RUHE, G. IMMoS: a methodology for integrated measurement, modelling and simulation. Software Process: Improvement and Practice, 7, n. 3-4, 2002. 189-210.

PSAROUDAKIS, J. E.; EBERHARDT, A. A discrete event simulation model to evaluate changes to a software project delivery process. IEEE 13th Conference on Commerce and Enterprise Computing (CEC). Luxembourg: IEEE. 2011.

RAFFO, D. Software project management using PROMPT: A hybrid metrics, modeling and utility framework. Information and Software Technology, 47, 2005. 1009-1017.

RAHMANDAD, H.; STERMAN, J. D. Reporting guidelines for simulation-based research in social sciences. System Dynamics Review, 28, n. 4, 2012. 396-411.

RAHMANDAD, H.; WEISS, D. Dynamics of concurrent software development. System Dynamics Review, 25, n. 3, 2009. 224–249.

RODRÍGUEZ, D.; SICILIA, M.; CUADRADO-GALLEGO, J.; PFAHL, D. e-learning in project management using simulation models: A case study based on the replication of an experiment. IEEE Transactions on Education, 49, n. 4, 2006. 451–463.

RUNESON, P.; HÖST, M. Guidelines for conducting and reporting case study research in software engineering. Empirical Software Engineering, 14, 2009. 131–164.

SARGENT, R. G. Validation and Verification of Simulation Models. Proc. of the 1999 Winter Simulation Conference. Phoenix, AZ, USA: IEEE. 1999.

SEVERANCE, F. System Modeling and Simulation: An Introduction. Chichester, England: John Wiley & Sons Ltd, 2001.

SOKOLOWSKI, J.; BANKS, C. Principles of Modeling and Simulation: A Multidisciplinary Approach. Hoboken, New Jersey: John Wiley & Sons, Inc, 2009.

STERMAN, J. Business dynamics: systems thinking and modeling for a complex world. Boston: Irwin/McGraw-Hill, 2000.

STOPFORD, B.; COUNSELL, S. A Framework for the Simulation of Structural Software Evolution. ACM Transactions on Modeling and Computer Simulation, 18, 2008.

THELIN, T.; PETERSSON, H.; RUNESON, P.; WÖHLIN, C. Applying sampling to improve software inspections. Journal of Systems and Software, 73, n. 2, 2004. 257–269.

THOMKE, S. H. Experimentation matters: unlocking the potential of new technologies for innovation. Boston: Harvard Business Press, 2003.

TRAVASSOS, G.; BARROS, M. Contributions of in virtuo and in silico experiments for the future of empirical studies in software engineering. 2nd Workshop on Empirical Software Engineering. Stuttgart, Germany: Fraunhofer IRB Verlag. 2003. p. 117-130.

TRAVASSOS, G.; DOS SANTOS, P.; NETO, P.; BIOLCHINI, J. An environment to support large scale experimentation in software engineering. 13th IEEE International Conference on Engineering of Complex Computer Systems. Belfast, Ireland: IEEE. 2008.

TURNU, I.; MELIS, M.; CAU, A.; SETZU, A.; CONCAS, G.; MANNARO, K. Modeling and simulation of open source development using an agile practice. Journal of Systems Architecture, 52, n. 11, 2006. 610–618.

UZZAFER, M. A simulation model for strategic management process of software projects. Journal of Systems and Software, 86, n. 1, 2013. 21-37.

WAKELAND, W.; MARTIN, R.; RAFFO, D. Using design of experiments, sensitivity analysis, and hybrid simulation to evaluate changes to a software development process: a case study. Software Process: Improvement and Practice, 9, n. 2, 2004. 107-119.

WÖHLIN, C.; RUNESON, P.; HÖST, M.; OHLSSON, M.; REGNELL, B.; WESSLÉN, A. Experimentation in software engineering. London: Springer Science & Business Media, 2012.

ZELKOWITZ, M. V. Techniques for Empirical validation. In: BASILI, V., et al. Empirical Software Engineering Issues. Critical Assessment and Future Directions. Berlin: Springer Berlin Heidelberg, 2007. p. 4-9.

ZHANG, H.; KITCHENHAM, B.; PFAHL, D. Reflections on 10 years of software process simulation modeling: a systematic review. In: WANG, Q.; PFAHL, D.; RAFFO, D. Making globally distributed software development a success story. Berlin: Springer Berlin Heidelberg, 2008. p. 345-356.

ZHANG, H.; KLEIN, G.; STAPLES, M.; ANDRONICK, J.; ZHU, L.; KOLANSKI, R. Simulation modeling of a large-scale formal verification process. Proceedings of the International Conference on Software and System Process. Zurich, Switzerland: IEEE Press. 2012. p. 3-12.

APPENDIX A

Consent Form for the Simulation Guidelines Evaluation

Informed Consent Form for Participation in Research

Study regarding Guidelines for Conducting Simulation-Based Experiments in Software Engineering

I declare to be over 18 years old and agree to participate in the studies conducted by the researchers Breno Bernard Nicolau de França and Prof. Guilherme Horta Travassos, as part of the activities of the Experimental Software Engineering course at COPPE/UFRJ.

This study aims at improving the understanding of which aspects are relevant for planning simulation-based experiments, in the context of a Software Engineering research. For that, we intend to observe the use of the proposed guidelines for this sort of study.

PROCEDURE

The proposed activities for this study should be performed individually and voluntarily. Each subject will receive the proposed scenario for the simulation study, the simulation model specification proposed by Chernoguz, as well as the required training for the Vensim tool and the executable simulation model. Furthermore, each subject will receive a set of planning guidelines for simulation-based experiments.

Each subject will need a computer with the Vensim tool installed and the simulation model file. Having the mentioned instruments in hands, the subject should:

1. Read the problem description and the general goal for the simulation-based experiment to be planned based on the proposed scenario and, then, start to elaborate the study plan (according to the Study Plan Template presented in class), using the simulation model as support and, in considering it relevant, the planning guidelines for simulation-based experiments. In the case the subject decides for using the guidelines, s/he will inform which guidelines were used to support the Study Plan elaboration in one section or another. 2. After all subjects deliver the plans, each subject should review the study plan from another subject supported by the planning guidelines. 3. After the review, the subject will join a group dynamics, in which subjective issues regarding their experience will be discussed.

It is extremely important that subjects do not discuss their tasks among themselves.

CONFIDENTIALITY

All information collected in this study is classified, and my name will not be identified at any moment. Similarly, I commit not to communicate my results before the study ends, as well as to keep confidentiality regarding the presented techniques and documents under the scope of this study.

BENEFITS AND FREEDOM TO WITHDRAW

I understand the benefits I will receive in this study are limited to the learning regarding the available material, regardless my participation on this study, but the researchers expect to comprehend better topics about simulation-based studies, and the benefits from this study for the context of Software Engineering.

I understand that I am free to ask questions at any moment and to request my information not to be included in the study. Still, I understand that I am participating in this study by freewill, only intending to contribute for the progress and development of Software Engineering research.

I understand that I am not obligated to contribute with information regarding my performance on these tasks, and that I may withdraw my results at any moment with no penalty or losses for me. I understand there will be no extra advantage or benefit in case of join the study.

RESPOSIBLE PROFESSOR

Prof. Guilherme Horta Travassos
Systems Engineering and Computer Science - COPPE/UFRJ

____________________________ ____________________________

Subject Name Subject Signature

APPENDIX B

Subject Characterization Form for the Simulation Guidelines Evaluation

Subject name:

E-mail:

Instruction level: ( ) Master Student ( ) Doctorate Student

1. How many software development projects do you participate in?

( ) None
( ) Until 1 project
( ) From 1 to 3 projects
( ) From 3 to 5 projects
( ) From 5 to 7 projects
( ) More than 7 projects

2. How much academic experience time do you have in the Software Engineering research area?

( ) None
( ) Until 1 year
( ) From 1 to 3 years
( ) From 3 to 5 years
( ) From 5 to 7 years
( ) More than 7 years

3. How much industry experience time do you have in the Software Engineering area?

( ) None
( ) Until 1 year
( ) From 1 to 3 years
( ) From 3 to 5 years
( ) From 5 to 7 years
( ) More than 7 years

4. How much academic experience time do you have in the Software Development Processes research area?

( ) None
( ) Until 1 year
( ) From 1 to 3 years
( ) From 3 to 5 years
( ) From 5 to 7 years
( ) More than 7 years

5. How much industry experience time do you have in the Software Development Processes area?

( ) None
( ) Until 1 year
( ) From 1 to 3 years
( ) From 3 to 5 years
( ) From 5 to 7 years
( ) More than 7 years

6. Have you worked (developed or used) with any simulation model? How many?

( ) None
( ) 1 simulation model
( ) 2 simulation models
( ) 3 simulation models
( ) 4 or more simulation models

7. Choose on the alternatives below the simulation approaches that you have some experience with:

( ) System Dynamics
( ) Discrete-Event Simulation
( ) Agent-Based Simulation
( ) State-Based Simulation
( ) Hybrid Simulation (Combinations)
( ) Other: ___________________________

APPENDIX C

Template for Study Plans

1. IDENTIFICATION
Title, Topic, Technical area, Author, Affiliation, Local, Date.
Used guidelines or complementary material in this section:___________________

2. CHARACTERIZATION
Type of study, Domain, Language, Partners (Institution, Address, Phone Number, E-mail and Internet Address), Links, Estimated Number of Trials, Glossary of Terms.
Used guidelines or complementary material in this section:___________________

3. INTRODUCTION
Background, Problem Definition, Document Organization.
Used guidelines or complementary material in this section:___________________

4. STUDY DEFINITION

- **Object of Study**
  Used guidelines or complementary material in this section:___________________
- **Main Goal**
  Used guidelines or complementary material in this section:___________________
- **Specific Goals**
  - Analyze
  - For the purpose of
  - With respect to
  - From the point of view
  - In the following context

  Used guidelines or complementary material in this section:___________________
- **Quality Focus**
  Used guidelines or complementary material in this section:___________________
- **Context**
  Used guidelines or complementary material in this section:___________________
- **Questions and Metrics**
  Used guidelines or complementary material in this section:___________________
- **Open questions**
  Used guidelines or complementary material in this section:___________________

5. PLANNING

- **Hypotheses Formulation**
  Used guidelines or complementary material in this section:___________________
- **Variables Selection**
  - Dependents
  - Independents

  Used guidelines or complementary material in this section:___________________
- **Subjects' Selection**
  - Selection Criteria
  - Required Experience
  - Criteria for Groups' Selection
  - Probabilistic Sampling Techniques
  - Non-Probabilistic Sampling Techniques

  Used guidelines or complementary material in this section:___________________
- **Resources**
  - Software
  - Hardware
  - Questionnaires

  Used guidelines or complementary material in this section:___________________
- **Experimental Design**

    - Objetcs
    - Measurements
    - Guidelines
    - Techniques
    - Factors
    - Treatments

  Used guidelines or complementary material in this section:____________________

- **Instruments**
    - Description
    - Justification
    - Advantages and Disadvantages
    - Limitations
    - Support to Quantitative Analysis
    - Support to Qualitative Analysis
    - Observation Criteria
    - Artifacts (Questionnaires, Procedures, etc)

  Used guidelines or complementary material in this section:____________________

- **Analysis Mechanisms**
    - Statistical Tests
    - Outliers' Removal Criteria

  Used guidelines or complementary material in this section:____________________

- **Results Validity**
    - Internal Validity
    - External Validity
    - Conclusion Validity
    - Construct Validity

  Used guidelines or complementary material in this section:____________________

6. TRAINING

- Training Definition and Procedures
    - Mentors
    - Participants
- Artifacts

  Used guidelines or complementary material in this section:____________________

7. EXECUTION PROCEDURE

- Execution Procedure for the Experimental Study
- Artifacts (Instructions, Documents, etc)

  Used guidelines or complementary material in this section:____________________

8. PLAN'S ASSESSMENT

- Goals
- Participants
- Execution Procedure
- Input Artifacts
- Output Artifacts (Lessons Learned, Change Suggestions for the Plan)

  Used guidelines or complementary material in this section:____________________

9. COSTS PLANNING

- Experimental Study Costs
    - Planning Costs
        - The plan
        - Instruments
        - Training material
        - Plan Assessment
    - Execution Costs
        - Deslocamentos
        - Training
        - Human Resources

        - Material Resources
    - Analysis Costs
    - Packaging Costs

Used guidelines or complementary material in this section:___________________

10. BIBLIOGRAPHY

Used guidelines or complementary material in this section:___________________

11. ANNEXES

Used guidelines or complementary material in this section:___________________

APPENDIX D

Organizational Scenario for the Evaluation of Simulation Guidelines

Organizational Scenario

An embedded software development organization, committed to their products and processes improvement, continually invests in practices capable of returning positive results regarding both quality and financial perspectives. However, the organization identified recent projects having schedule overruns on product deliveries. Therefore, its relation with clients and reputation have been negatively affected, resulting also on financial losses.

The projects identified characterized as significantly behind the schedule, usually, starts with a small number of veterans (more or less four senior professionals) when compared to the number of rookies (about 30 novice professionals). Moreover, it is possible to add manpower, when requirements and early architecture design are almost complete. In general, rookies allocated during the project take about 30 days to become as productive as veterans.

In general, the QA manager perceived that when the project progress reach about 30% of the estimated schedule and the project is already late, the project managers often react by adding more developers. However, he also notice that such approach has not solved the problem yet. Besides, it increases the project budget by allocating additional manpower. Thus, the organization hired you, as a software engineering consultant, and requested you to give a short-time diagnostic for the current situation, explaining the reasons for the raised issues and proposing a feasible solution for the problem, i.e., which strategy can be adopted in future similar projects to reduce the losses.

The observed effect by adding more developers reminds the behavior described by Frederick Brooks, in 1975: “Adding manpower to a late software project makes it later”. This behavior is the so-called Brooks’ Law (see next section).

Understanding the problem should be investigation in a short period and there is projects’ historical data available, a feasible alternative is to conduct (plan, execute, and analyze) simulation-based experiments in order to understand the problem affecting the deliveries and testing possible alternative solutions.

Regarding the historical dataset from the organization, it is possible to identify information for the following metrics, by project:

- Project Schedule (chronogram, in days);
- Functional size (in function points);
- Tem productivity, total and stratified in rookies and veterans (in function points / person-month);
- Team size, total and stratified in rookies and veterans;
- Training effort (in person-day);
- Mean time for individual learning (in days);
- Communication overhead, percent of communication needs and tasks coordination effort (in person-day);

Brooks' Law

In the project beginning, project managers estimate the required effort, negotiate schedule and budget and, in the case things deviate from the planned, plan corrective action to put the project back on track. In these actions, project managers should be aware of what to change. Some actions may have negative impact, as when professionals become overloaded or more manpower is added to a late project. When working overloaded, it triggers a vicious cycle, in which professionals become exhausted and consequently, it increases the number of defects, rework and productivity decay. For the other case, the paradox effect of adding more manpower, or the Brooks' Law, refers to the behavior observed by Frederick Brooks and described in his book titled, *The Mythical Man-Month*, where the author alerts project managers regarding attempts of bringing the project back on track,

> "When schedule slippage is recognized, the natural (and traditional) response is to add manpower . . .Like dousing a fire with gasoline, this makes matters worse, much worse. Oversimplifying outrageously, we state Brooks' Law: Adding manpower to a late software project makes it later".

Brooks explains his law by the very nature of programming work, which is "*more like having a baby than picking cotton"*:

> "Ten people can pick cotton ten times as fast as one person because the work is almost perfectly partitionable, requiring little communication or coordination. But nine women can't have a baby any faster than one woman can because the work is not partitionable". Unlike manufacturing, software construction is an inherently systemic effort: it cannot be easily partitioned into isolated, independent tasks. The complexity of software development "... creates the tremendous learning and understanding burden that makes personnel turnover a disaster".

When new people is added to a software development projects, it is required domain knowledge and project architecture information, as well as organizational policies and procedures, team responsibilities, and other relevant information. Usually, a veteran helps a rookie to become part of the team. Mentoring activities involves the veteran to deviate from his current tasks and increase the general communication overhead. The increase on training and in the communication and coordination overhead are contributing factors for the Brooks' Law, due to the extra workload. Training affects productive work reduction for the veterans. In general, the Brooks' Law states that while the project communication complexity and effort increase by the square of team size, the workload increases linearly. This way, the effort spent in training, coordination, and communication is greater than the return of waiting until the rookies become effectively productive.

Simulation Model for the Brooks' Law

Chernoguz (2011) proposed a model for the observation of the Brooks' Law phenomenon, including several improvements w.r.t. models previously published in the Software Engineering technical literature. Such model can be used as basis for planning simulation-based studies. Figure 1 presents a causal model for the Brooks' Law. Figure 2 presents the SD's stock and flows diagram.

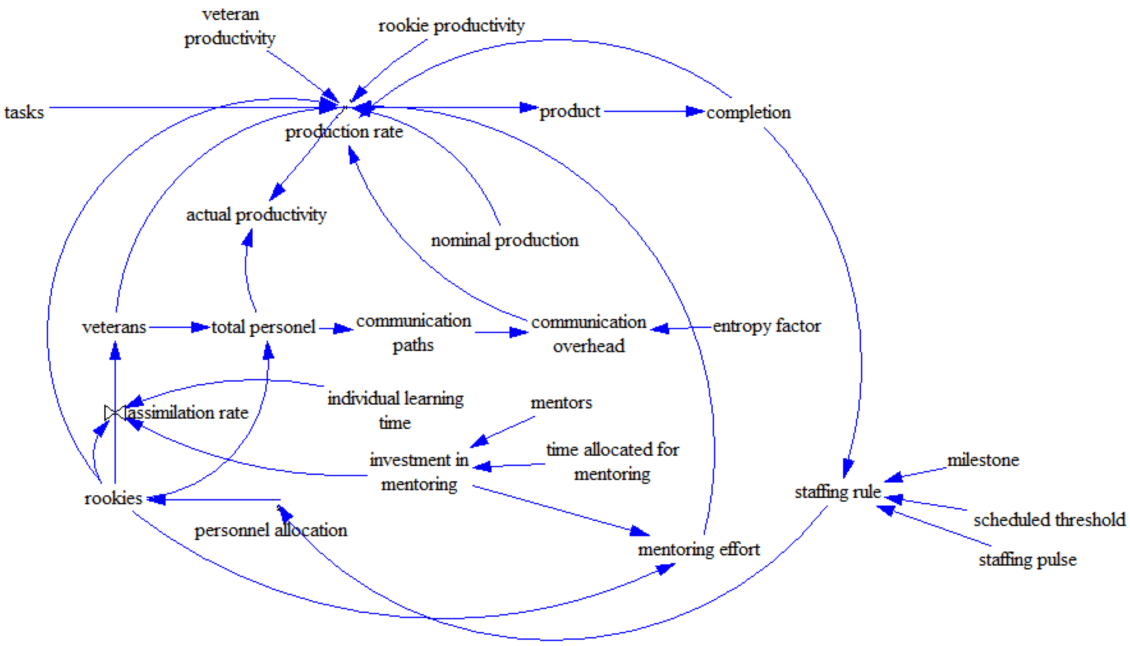


Figure 1. Cause-effect diagram for the Brooks' Law.

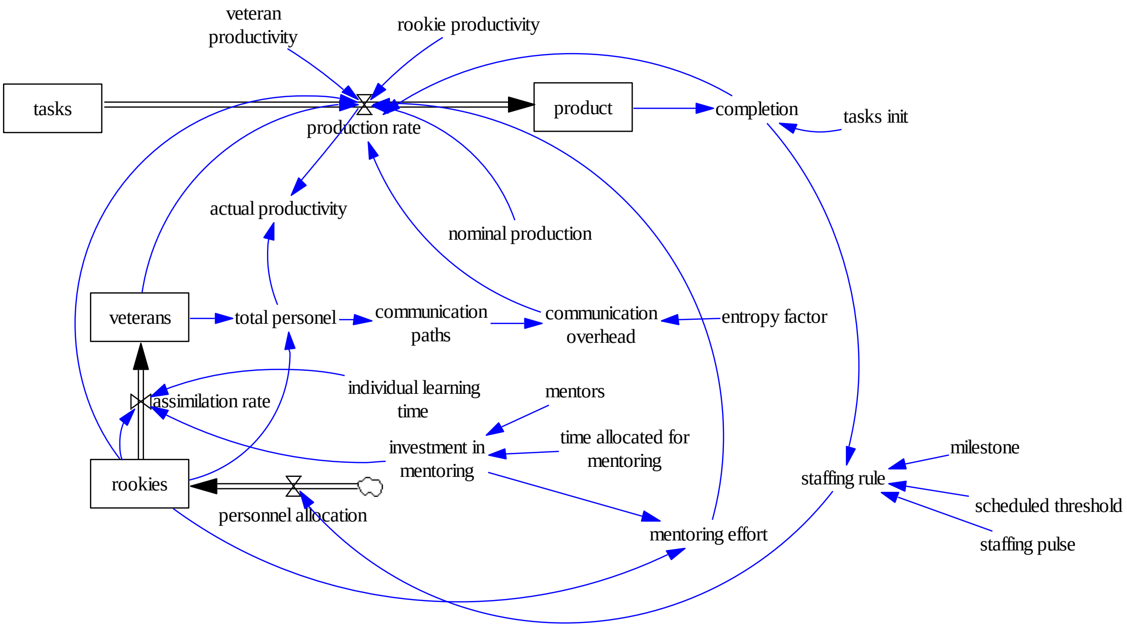


Figure 2. Stocks and Flows diagram for the Brooks' Law.

Model Inputs (Constants)

Parameters	Description	Intervals
Veteran productivity	Mean productivity for veterans	[1.0 – 5.0]
Rookie productivity	Mean productivity for novices	[0.2 – 1.0]
Nominal production	Daily effort (person-day) of a developer	[0.05 – 0.2]
Entropy factor	Entropy factor[10] for project communication	[0.03 – 0.06]
Individual learning time	Number of workdays required for rookies' assimilation into the project without team investment in training	[5 - 60]
Mentors	Number of experienced members allocated for training	[1 – all veterans]
Time allocated for mentoring	A fraction of veterans' time reallocated away from production to mentor rookies	[0.125 – 1.0]
Staffing pulse	Amount of manpower to be added during the project.	[4 – 20]
Schedule threshold	Project progress percent, from which the project is considered to be late (*time step*, simulation step).	[22 – 45]
Veterans (Initial value)	Amount of experienced professionals in the project beginning.	[4-10]
Rookies (Initial value)	Amount of novice professionals in the project beginning.	[28-32]

Intermediate Variables

Variable	Description
Tasks	Tasks to be performed in the software development.
Product	Amount of performed tasks, i.e., amount of work done.
Task init	Initial amount of tasks (*Tasks*).
Veterans	Amount of experienced professionals during the projects.
Rookies	Amount of unexperienced professionals during the projects.
Assimilation rate	Mean time for rookies become as productive as veterans.
Personnel Allocation	Amount unexperienced professionals allocated per day.
Communication overhead	Effort spent on communication between members.
Communication paths	Amount of communication channels between each pair of members.
Total personnel	Total team size.
Investment in mentoring	Percentage of daily effort spent by mentors.
Mentoring effort	Daily effort dedicated for mentoring including all rookies.
Staffing rule	Rule determining if new members should be allocated to the project.
Milestone	Simulation step in which new members are allocated to the project.

[10] É uma medida do grau de desorganização que pode levar a falência de um sistema.

Model Outputs

Variable	Description
Production rate	Daily project team effort measured in number of tasks (i.e., function points) by team and by day.
Actual productivity	Actual productivity observed.
Schedule	Execution time for the project, in days. Observed by the simulation cycles (*dt*), with each cycle being ¼ of a day.

References

Brooks, Frederick P. The mythical man-month. Vol. 1995. Reading, MA: Addison-Wesley, 1975.

David G Chernoguz. 2011. The system dynamics of Brooks' Law in team production. *Simulation*87, 11 (November 2011), 947-975. DOI=10.1177/0037549710382423.

APPENDIX E

Proof of concept for the Planning Guidelines

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 life cycle 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 of costs and possibly on decay of product quality.

Goal and Research Questions

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

To analyze *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 the product quality?

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).

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 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 1.

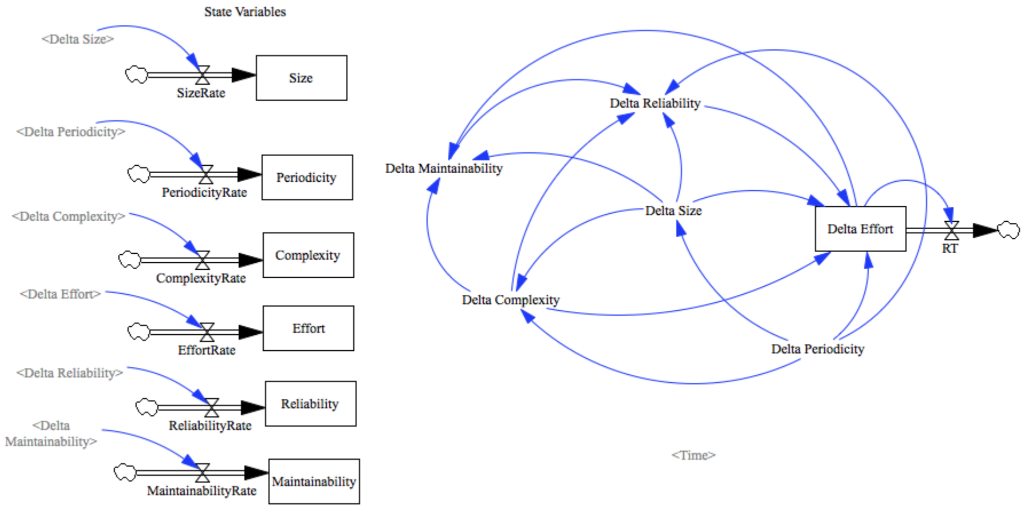


Fig 1. Software Evolution Model adapted from (Araújo et al, 2012).

The model was developed over six state variables, which 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 life cycle 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 the model validity, the authors collected evidence for each relationship amongst model variables from the technical literature (SG24). For the complete set of evidence, see (Araújo *et al*, 2012). Besides, the model was successfully assessed 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 ran 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.

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.

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 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 1, it is possible to see that there are first-order and other higher-order possible effects of Periodicity on both Reliability and Maintainability. Therefore, in this experiment we will explore full factorial designs for questions Q1 and Q2 as shown in the design matrix (Table 1).

Table 1. 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)

For the scenarios (SG13) concerning with 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, being a total of 300 runs for the variable-duration scenarios.

In order to determine the number of required trials, we adopted the method by Law and Kelton (2008). Basically, it consists in choosing an initial number of runs (sample size), from which a confidence interval will be estimated. Such number of run should be increased until the confidence interval present an estimated error less or equal than the allowable percentage error between the simulated and real means (0.01 in our experiment). This procedure should be performed for each output variables. In our case, Maintainability and Reliability. The estimated mean and confidence intervals for each variable are presented in Fig 2 and Fig 3.

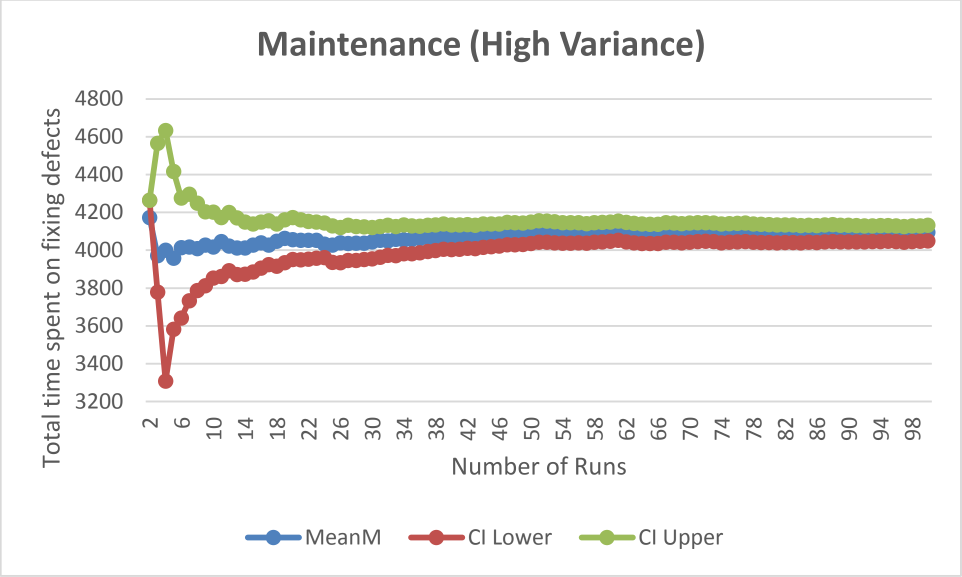


Fig 2. Determination of number of runs for Maintenance.

Actually, the confidence interval for Maintenance converges with the mean in 99 runs, while Reliability converges in 100 runs. Both numbers were considered for the worst case (high variance). However, the number of runs should be the same for each scenario. This way, we selected 100 runs for each stochastic scenario in the experimental design.

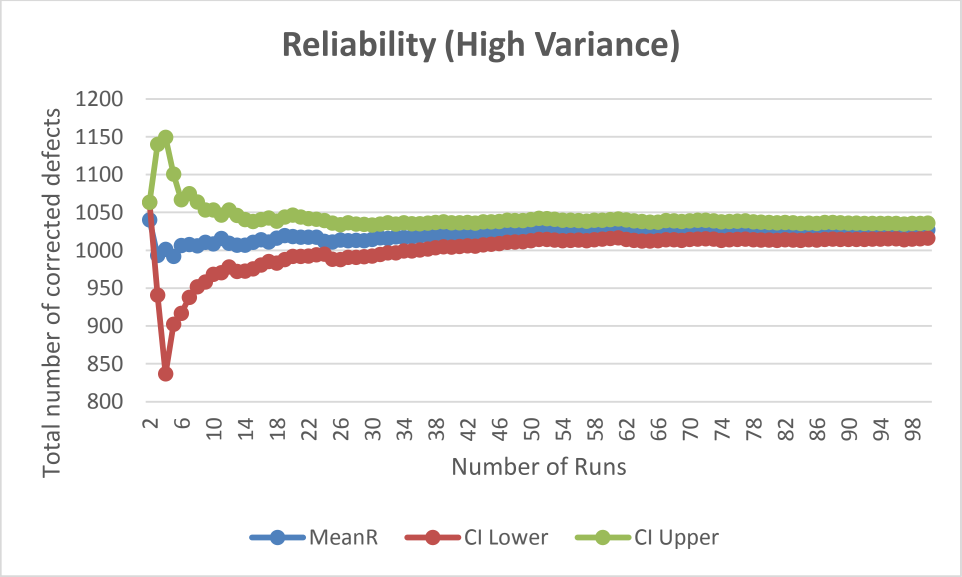


Fig 3. Determination of number of runs for Reliability.

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.

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. Besides, 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 1.3.

The system releases resulted 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.

Output Analysis

For output analysis (SG31), statistical charts are used, namely histograms and sequence run charts, to characterize response variable behavior. 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 analyze 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 done with the other factors or interactions concerning 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.

Simulation Results

As mentioned in Section 1.5, a total of 303 simulation runs are 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 4 shows, the shorter periodicity scenario (1) has a higher number of corrected defects over six months. This result is explained by two main reasons: shorter cycles are mostly related to corrective and adaptive 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.

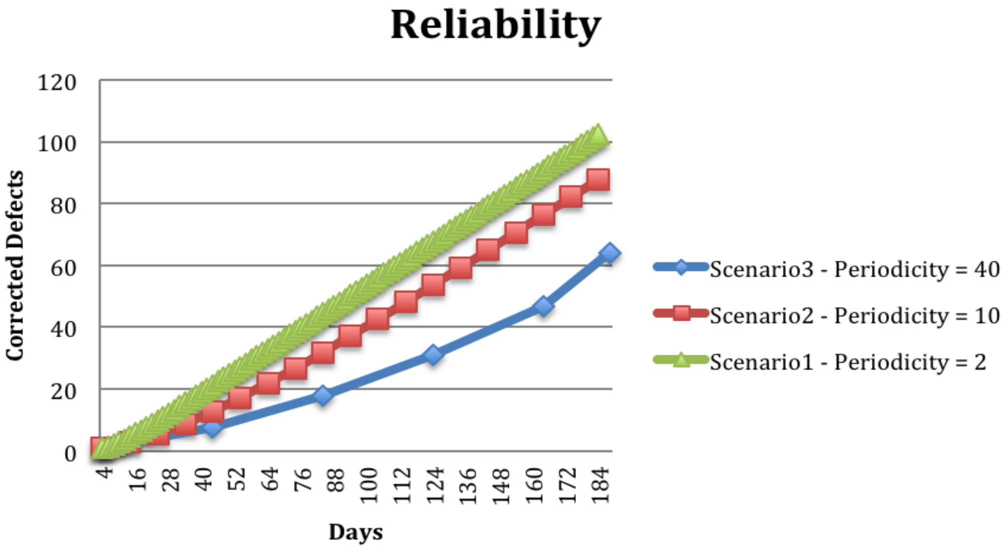


Fig 4. Reliability Output for Fixed-Durations.

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. Therefore, 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, it should be clear 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 behavior for the variable-duration strategy. This behavior 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.

Threats to Validity (SG19 and SG32)

As a general limitation, the model adopted has a perspective abstracting the process-level details and presents only the behavior of continuous variables involved in its causal model. Therefore, 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 behavior. However, we can also see both short enhancement cycles and long correction cycles in the initial dataset. This kind of behavior 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.

References

Araújo MA, Monteiro V, Travassos GH (2012) Towards a Model to Support *in silico* Studies regarding Software Evolution. In: ESEM 2012.

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.

Law, A. M., Kelton, W. D. (2000). *Simulation modeling and analysis* (Vol. 3). New York: McGraw-Hill.

APPENDIX F

List of papers captured in the *quasi*-Systematic Review

[1] R. Ahmed, T. Hall, P. Wernick, S. Robinson, M. Shah, “Software process simulation modelling: A survey of practice,” *Journal of Simulation*, vol. 2, pp. 91 – 102, 2008.

[2] H. Zhang, B. Kitchenham, D. Pfahl, “Reflections on 10 years of software process simulation modeling: A systematic review,” *LNCS*, vol. 5007, pp. 345-356, 2008.

[3] T. Abdel-Hamid, “Understanding the "90% syndrome" in software project management: A simulation-based case study,” *Journal of Systems and Software*, vol. 8, pp. 319-33, 1988.

[4] R. Martin, D. Raffo, “Application of a hybrid process simulation model to a software development project,” *Journal of Systems and Software*, vol. 59, pp. 237– 46, 2001.

[5] B. Lee, J. Miller, “Multi-project management in Software Engineering using simulation modeling,” *Software Quality Journal*, vol. 12, pp. 59-82, Kluwer Academic Publ, 2004.

[6] S. W. Ormon, C. R. Cassady, A. G. Greenwood, “A simulation-based reliability prediction model for conceptual design,” in *Proceedings of the Annual Reliability and Maintainability Symposium*, Philadelphia, PA, United states, 2001, pp. 433 – 436.

[7] E. O. Navarro, A. V. D. Hoek, “Design and evaluation of an educational software process simulation environment and associated model,” in *Proc. of 18th Conference on Software Engineering Education and Training*, CSEE and T 2005, Ottawa, ON, Canada, 2005, pp. 25 – 34.

[8] S. Grillinger, P., Brada, P., Racek, “Simulation approach to embedded system programming and testing,” in *Proc. of 11th IEEE International Conference and Workshop on the Engineering of Computer-Based Systems*, ECBS 2004, Brno, 2004, pp. 248–254.

[9] I. Turnu, M. Melis, A. Cau, A. Setzu, G. Concas, K. Mannaro, “Modeling and simulation of open source development using an agile practice,” *Journal of Systems Architecture*, vol. 52, no. 11, pp. 610 – 618, 2006.

[10] D. Pfahl, K. Lebsanft, “Using simulation to analyze the impact of software requirement volatility on project performance,” *Information and Software Technology*, vol. 42, pp. 1001–1008, 2000.

[11] F. Padberg, “A software process scheduling simulator,” in *Proc. ICSE*, pp. 816-817, 2003.

[12] D. M. Raffo, “Software project management using PROMPT: A hybrid metrics, modeling and utility framework,” *Information and Software Technology*, vol.47, pp.1009–1017, 2005.

[13] B. Stopford, S. Counsell, “A Framework for the Simulation of Structural Software Evolution,” *ACM Transactions on Modeling and Computer Simulation*, vol. 18, 2008.

[14] Y. X. Chen, Q. Liu, “Hierarchy-based team software process simulation model,” *Wuhan University Journal of Natural Sciences*, vol. 11, pp. 273– 77, 2006.

[15] T. Abdel-Hamid, “The economics of software quality assurance: A simulation-based case study,” *MIS Quarterly: Management Information Systems*, vol. 12, pp. 395-410, 1988.

[16] T. Abdel-Hamid, “Dynamics of software project staffing: A system dynamics based simulation approach,” *IEEE Trans. on Software Engineering*, vol. 15, pp. 109–119, 1989.

[17] T. Abdel-Hamid, “Investigating the cost/schedule trade-off in software development,” *IEEE Software*, vol. 7, pp. 97-105, 1990.

[18] T. Abdel-Hamid, “A multiproject perspective of single-project dynamics,” *Journal of Systems and Software*, vol. 22, pp. 151-165, 1993.

[19] T. Abdel-Hamid, S. Madnick, “Impact of Schedule Estimation on Software Project Behavior,” *IEEE Software*, vol. 3, pp. 70–75, 1986.

[20] A. Drappa, J. Ludewig, “Quantitative modeling for the interactive simulation of software projects,” *Journal of Systems and Software*, vol. 46, pp. 113 – 122, 1999.

[21] S. Ferreira, J. Collofello, D. Shunk, G. Mackulak, "Understanding the effects of requirements volatility in software engineering by using analytical modeling and software process simulation," *Journal of Systems and Software*, vol. 82, pp. 1568–1577. 2009.

[22] B. G. Ambrosio, J. L. Braga, and M. A. Resende-Filho, "Modeling and scenario simulation for decision support in management of requirements activities in software projects," Journal of Software Maintenance and Evolution, vol. 23, no. 1, pp. 35 – 50, 2011.

[23] S. Setamanit, W. Wakeland, D. Raffo, "Using simulation to evaluate global software development task allocation strategies," *Software Process Improvement and Practice*, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom, vol. 12, pp. 491 – 503, 2007.

[24] K. Choi, D. Bae, T. Kim, "An approach to a hybrid software process simulation using the DEVS formalism," *Software Process Improvement and Practice*, vol. 11, pp. 373 – 383, 2006.

[25] S. Setamanit, D. Raffo, "Identifying key success factors for globally distributed software development project using simulation: A case study," *Lecture Notes in Computer Science*, Leipzig, Germany, vol. 5007 LNCS, pp. 320 – 332, 2008.

[26] M. Melis, I. Turnu, A. Cau, G. Concas, "Evaluating the impact of test-first programming and pair programming through software process simulation," *Software Process Improvement and Practice*, vol. 11, pp. 345 – 360, 2006.

[27] D. X. Houston, S. Ferreira, J. S. Collofello, D. C. Montgomery, G. T. Mackulak, D. L. Shunk, "Behavioral characterization: Finding and using the influential factors in software process simulation models," *Journal of Systems and Software*, vol. 59, pp. 259-270, 2001.

[28] W. W. Wakeland, R. H. Martin, D. Raffo, "Using Design of Experiments, sensitivity analysis, and hybrid simulation to evaluate changes to a software development process: A case study," *Software Process Improvement and Practice*, vol. 9, pp. 107–119, 2004.

[29] D. Pfahl, M. Klemm, G. Ruhe, "A CBT module with integrated simulation component for software project management education and training," *Journal of Systems and Software*, vol. 59, no. 3, pp. 283 – 298, 2001.

[30] V. Garousi, K. Khosrovian, D. Pfahl, "A customizable pattern-based software process simulation model: Design, calibration and application," *Software Process Improvement and Practice*, vol. 14, no. 3, pp. 165 – 180, 2009.

[31] K. G. Kouskouras, A. C. Georgiou, "A discrete event simulation model in the case of managing a software project," *European Journal of Operational Research*, vol. 181, no. 1, pp. 374 – 389, 2007.

[32] M. Ruiz, M. Ramos, I. Toro, "A Dynamic Integrated Framework for Software Process Improvement," *Software Quality Journal*, vol. 10, no. 2, pp. 181–194, 2002.

[33] D. Williams, T. Hall, M. Kennedy, "A framework for improving the requirements engineering process management," *Software Quality Journal*, vol. 8, no. 2, pp. 133–147, 1999.

[34] M. Höst, B. Regnell, C. Tingström, "A framework for simulation of requirements engineering processes," in *EUROMICRO 2008 – Proc. of the 34th EUROMICRO Conference on Software Engineering and Advanced Applications*, SEAA 2008, Parma, 2008, pp. 183–190.

[35] N. Hanakawa, K. Matsumoto, K. Torii, "A knowledge-based software process simulation model," *Annals of Software Engineering*, vol. 14, no. 1–4, pp. 383–406, Dec. 2002.

[36] C. Chen, C. Wang, Y. Lee, Y. Lee, "A process-oriented system dynamics model for software development project prediction," in *Proc. of 4th International Conference on Networked Computing and Advanced Information Management*, NCM 2008, Gyeongju, Korea, Republic of, 2008, vol. 2, pp. 126 – 131.

[37] T. L. Landers, H. A. Taha, C. L. King, "A reliability simulation approach for use in the design process," *IEEE Transactions on Reliability*, vol. 40, no. 2, pp. 177 – 181, 1991.

[38] Z. Wang, W. Haberl, A. Herkersdorf, M. Wechs, "A simulation approach for performance validation during embedded systems design," *Communications in Computer and Information Science*, Porto Sani, Greece, vol. 17 CCIS, pp. 385 – 399, 2008.

[39] S. Cook, R. Harrison, P. Wernick, “A simulation model of self-organising evolvability in software systems,” in *Proc. of the 2005 IEEE International Workshop on Software Evolvability*, Budapest, Hungary, 2005, pp. 17 – 22.

[40] S. Balsamo, M. Marzolla, “A Simulation-Based Approach to Software Performance Modeling,” in *Proc. of the Joint European Software Engineering Conference (ESEC) and SIGSOFT Symposium on the Foundations of Software Engineering* (FSE-11), Helsinki, Iceland, 2003, pp. 363 – 366.

[41] N. Smith, A. Capiluppi, J. F. Ramil, “A study of open source software evolution data using qualitative simulation,” *Software Process Improvement and Practice*, vol. 10, no. 3, pp. 287–300, 2005.

[42] D. Pfahl, A. Al-Emran, G. Ruhe, “A system dynamics simulation model for analyzing the stability of software release plans,” *Software Process Improvement and Practice*, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom, vol. 12, pp. 475 – 490, 2007.

[43] S. Kuppuswami, K. Vivekanandan, P. Rodrigues, “A system dynamics simulation model to find the effects of XP on cost of change curve,” *Lecture Notes in Computer Science*, vol. 2675, pp. 54–62, 2003.

[44] D. Pfahl, N. Koval, G. Ruhe, “An experiment for evaluating the effectiveness of using a system dynamics simulation model in software project management education,” in *International Software Metrics Symposium*, Proceedings, London, United kingdom, 2001, pp. 97 – 109.

[45] T. Abdel-Hamid, F. Leidy, “An Expert Simulator for Allocating the Quality Assurance Effort in Software Development,” *Simulation*, vol. 56, no. 4, pp. 233–240, Apr. 1991.

[46] D. Pfahl, O. Laitenberger, J. Dorsch, G. Ruhe, “An Externally Replicated Experiment for Evaluating the Learning Effectiveness of Using Simulations in Software Project Management Education,” Empirical Software Engineering, vol. 8, no. 4, pp. 367 – 395, 2003.

[47] L. Williams, A. Shukla, A. I. Anton, “An initial exploration of the relationship between pair programming and Brooks’ law,” in *Proc. of the Agile Development Conference*, ADC 2004, Salt Lake City, UT, United states, 2004, pp. 11 – 20.

[48] S. Ur, E. Yom-Tov, P. Wernick, “An open source simulation model of software development and testing,” in *Lecture Notes in Computer Science*, Haifa, Israel, 2007, vol. 4383 LNCS, pp. 124 – 137.

[49] T. Thelin, H. Petersson, P. Runeson, C. Wohlin, “Applying sampling to improve software inspections,” Journal of Systems and Software, vol. 73, no. 2, pp. 257 – 269, 2004.

[50] F. Alvarez, Guillermo A., Cristian, “Applying simulation to the design and performance evaluation of fault-tolerant systems,” in *Proc. of the IEEE Symposium on Reliable Distributed Systems*, Durham, NC, USA, 1997, pp. 35–42.

[51] K. Kang, K. Lee, J. Lee, G. Kim, “ASADAL/SIM: An incremental multi-level simulation and analysis tool for real-time software specifications,” *Software Practice & Experience*, vol. 28, no. 4, pp. 445–462, Apr. 1998.

[52] R. Madachy, B. Boehm, J. A. Lane, “Assessing hybrid incremental processes for SISOS development,” *Software Process Improvement and Practice*, Southern Gate, Chichester, West Sussex, United Kingdom, 2007, vol. 12, pp. 461 – 473.

[53] T. Häberlein, “Common structures in system dynamics models of software acquisition projects,” *Software Process Improvement and Practice*, vol. 9, no. 2, pp. 67–80, 2004.

[54] S. Park, K. Choi, K. Yoon, D. Bae, “Deriving software process simulation model from spem-based software process model,” in *Proc. of Asia-Pacific Software Engineering Conference*, APSEC, Nagoya, Japan, 2007, pp. 382 – 389.

[55] P. Sooraj, P. K. J. Mohapatra, “Developing an inter-site coordination index for global software development,” in *Proc. of 3rd IEEE International Conference Global Software Engineering*, ICGSE 2008, Bangalore, India, 2008, pp. 119 – 128.

[56] C. Lauer, R. German, J. Pollmer, “Discrete event simulation and analysis of timing problems in automotive embedded systems,” in *IEEE International Systems Conference Proceedings*, SysCon 2010, San Diego, CA, United states, 2010, pp. 18 – 22.

[57] R. Madachy, B. Khoshnevis, “Dynamic simulation modeling of an inspection-based software lifecycle processes,” *Simulation*, vol. 69, no. 1, pp. 35 – 47, 1997.

[58] H. Rahmandad, D. M. Weiss, “Dynamics of concurrent software development,” *System Dynamics Review*, vol. 25, no. 3, pp. 224–249, Sep. 2009.

[59] A. Al-Emran, D. Pfahl, and G. Ruhe, “DynaReP: A discrete event simulation model for re-planning of software releases,” *Lecture Notes in Computer Science*, Minneapolis, MN, United states, 2007, vol. 4470, pp. 246 – 258.

[60] D. Rodríguez, M. Á. Sicilia, J. J. Cuadrado-Gallego, D. Pfahl, “e-learning in project management using simulation models: A case study based on the replication of an experiment,” *IEEE Transactions on Education*, vol. 49, no. 4, pp. 451–463, 2006.

[61] D. Pfahl, O. Laitenberger, G. Ruhe, J. Dorsch, T. Krivobokova, “Evaluating the learning effectiveness of using simulations in software project management education: Results from a twice replicated experiment,” *Information and Software Technology*, vol. 46, no. 2, pp. 127 – 147, 2004.

[62] S. Setamanit, W. Wakeland, D. Raffo, “Exploring the impact of task allocation strategies for global software development using simulation,” *Lecture Notes in Computer Science*, Shanghai, China, 2006, vol. 3966 LNCS, pp. 274 – 285.

[63] R. Lazarova-Molnar, S., Mizouni, “Floating task: Introducing and simulating a higher degree of uncertainty in project schedules,” in *Proc. of the Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises*, WET ICE, Larissa, 2010, pp. 136–141.

[64] K. Khosrovian, D. Pfahl, V. Garousi, “GENSIM 2.0: A customizable process simulation model for software process evaluation,” in *Lecture Notes in Computer*, Leipzig, Germany, vol. 5007, pp. 294 – 306, 2008.

[65] H. Zhang, R. Jeffery, L. Zhu, “Hybrid modeling of test-and-fix processes in incremental development,” *Lecture Notes in Computer Science*, Leipzig, Germany, vol 5007, pp. 333 – 344, 2008.

[66] D. Merrill, J. S. Collofello, “Improving software project management skills using a software project simulator,” in *Proc. of Frontiers in Education Conference*, Pittsburgh, PA, USA, 1997, vol. 3, pp. 1361 – 1366.

[67] R. Madachy, “Integrated modeling of business value and software processes,” *Lecture Notes in Computer Science*, Beijing, China, 2006, vol. 3840 LNCS, pp. 389 – 402.

[68] S. Sakthivel, R. Agarwal, “Knowledge-Based Model Construction for Simulating Information Systems,” *Simulation*, vol. 59, no. 4, pp. 223–236, Oct. 1992.

[69] N. Hanakawa, S. Morisaki, K. Matsumoto, “Learning curve based simulation model for software development,” in *Proc. International Conference on Software Engineering*, Kyoto, Jpn, 1998, pp. 350 – 359.

[70] N. Powell, A., Murdoch, J., Tudor, “Modeling risk-benefit assumptions in technology substitution,” *Lecture Notes in Computer Science*, vol. 4470, pp. 295–306, 2007.

[71] I. Collofello, James S., Zhen, Yang, Tvedt, John D., Merrill, Derek, Rus, “Modeling software testing processes,” in *Proc. of International Phoenix Conference on Computers and Communications*, Scottsdale, AZ, USA, 1996, pp. 289–293.

[72] R. R. Levary, C. Y. Lin, “Modelling the software development process using an expert simulation system having fuzzy logic,” *Software - Practice and Experience*, vol. 21, no. 2, pp. 133 – 148, 1991.

[73] J. Lee, B., Miller, “Multi-project software engineering analysis using systems thinking,” *Software Process Improvement and Practice*, vol. 9, no. 3, pp. 173–214, 2004.

[74] R. Koci, V. Janouek, “OOPN and DEVS formalisms for system specification and analysis,” in *Proc. of 5th International Conference on Software Engineering Advances*, ICSEA 2010, Nice, France, 2010, pp. 305 – 310.

[75] A. Al-Emran. D. Pfahl, “Operational planning, re-planning and risk analysis for software releases,” *Lecture Notes in Computer Science*, Riga, Latvia, 2007, vol. 4589, pp. 315 – 329.

[76] G. Waters, G., Linington, P., Akehurst, D., Utton, P., Martin, “Permabase: Predicting the performance of distributed systems at the design stage,” *IEEE Software*, vol. 148, no. 4, pp. 113–121, 2001.

[77] M. Nonaka, L. Zhu, M. A. Babar, M. Staples, “Project cost overrun simulation in software product line development,” *Lecture Notes in Computer Science*, Riga, Latvia, 2007, vol. 4589, pp. 330 – 344.

[78] M. Nonaka, Z. Liming, M. A. Babar, M. Staples, “Project delay variability simulation in software product line development,” *Lecture Notes in Computer Science*, Minneapolis, MN, United states, 2007, vol. 4470, pp. 283 – 294.

[79] H. Zhang, M. Huo, B. Kitchenham, R. Jeffery, “Qualitative simulation model for software engineering process,” in *Proc. of the Australian Software Engineering Conference*, ASWEC, Sydney, Australia, 2006, vol. 2006, pp. 391 – 400.

[80] H. Zhang, B. Kitchenham, “Semi-quantitative simulation modeling of software engineering process,” *Lecture Notes in Computer Science*, vol. 3966, pp. 242–253, 2006.

[81] A. Drappa, J. Ludewig, “Simulation in software engineering training,” in *Proc. of International Conference on Software Engineering*, Limerick, Ireland, 2000, pp. 199 – 208.

[82] M. Wu, H. Yan, “Simulation in software engineering with system dynamics: A case study,” *Journal of Software*, vol. 4, no. 10, pp. 1127 – 1135, 2009.

[83] B. Zhang, R. Zhang, “Simulation model of team software process using temporal parallel automata,” in *Proc. 4th International Conference on Internet Computing for Science and Engineering*, ICICSE 2009, Harbin, China, 2010, pp. 283 – 285.

[84] D. Pfahl, A. Al-Emran, G. Ruhe, “Simulation-based stability analysis for software release plans,” *Lecture Notes in Computer Science*, Shanghai, China, 2006, vol. 3966, pp. 262 – 273.

[85] N. Celik, H. Xi, D. Xu, Y. Son, “Simulation-based workforce assignment considering position in a social network,” in *Proc. of Winter Simulation Conference*, Baltimore, MD, United states, 2010, pp. 3228 – 3240.

[86] C. Caulfield, G. Kohli, S. P. Maj, “Sociology in software engineering,” in *ASEE Annual Conference Proc.*, Salt Lake City, UT, United states, 2004, pp. 12685 – 12697.

[87] C. L. . Wernick, P., Hall, T., Nehaniv, “Software evolutionary dynamics modelled as the activity of an actor-network,” *IET Software*, vol. 2, no. 4, pp. 321–336, 2008.

[88] D. M. Raffo, W. Harrison, J. Vandeville, “Software process decision support: Making process tradeoffs using a hybrid metrics, modeling and utility framework,” *ACM International Conference Proceeding Series*, Ischia, Italy, 2002, vol. 27, pp. 803 – 809.

[89] I. Rus, J. Collofello, P. Lakey, “Software process simulation for reliability management,” *Journal of Systems and Software*, vol. 46, no. 2, pp. 173 – 182, 1999.

[90] D.M. Raffo, J.V. Vandeville, R. H. Martin, “Software process simulation to achieve higher CMM levels,” *Journal of Systems and Software*, vol. 46, no. 2, pp. 163–172, 1999.

[91] P. Wernick, M. M. Lehman, “Software process white box modelling for FEAST/1,” *Journal of Systems and Software*, vol. 46, no. 2, pp. 193 – 201, 1999.

[92] C. Y. Lin, T. Abdel-Hamid, J. S. Sherif, “Software-Engineering Process Simulation model (SEPS),” *Journal of Systems and Software*, vol. 38, no. 3, pp. 263 – 277, 1997.

[93] R. Madachy, B. Boehm, J. A. Lane, “Spiral lifecycle increment modeling for new hybrid processes,” *Lecture Notes in Computer Science*, Shanghai, China, 2006, vol. 3966 LNCS, pp. 167 – 177.

[94] J. Zhai, Q. Yang, F. Su, J. Xiao, Q. Wang, M. Li, “Stochastic process algebra based software process simulation modeling,” *Lecture Notes in Computer Science*, Vancouver, Canada, 2009, vol. 5543, pp. 136 – 147.

[95] D.M. Raffo, S. Setamanit, “Supporting software process decisions using bi-directional simulation,” *International Journal of Software Engineering and Knowledge Engineering*, vol. 13, no. 5, pp. 513–530, 2003.

[96] R. J. Madachy, “System dynamics modeling of an inspection-based process,” in *Proc. of International Conference on Software Engineering*, Berlin, Germany, 1995, pp. 376–386.

[97] F. Stallinger, P. Grunbacher, “System dynamics modelling and simulation of collaborative requirements engineering,” *Journal of Systems and Software*, vol. 59, no. 3, pp. 311 – 321, 2001.

[98] G. Kahen, M. M. Lehman, J. F. Ramil, P. Wernick, “System dynamics modelling of software evolution processes for policy investigation: Approach and example,” *Journal of Systems and Software*, vol. 59, no. 3, pp. 271–281, 2001.

[99] F. Huq, “Testing in the software development life-cycle: Now or later,” *International Journal of Project Management*, vol. 18, no. 4, pp. 243–250, 2000.

[100] T. Wernick, P., Hall, “The impact of using pair programming on system evolution: A simulation-based study,” in *IEEE International Conference on Software Maintenance*, ICSM, Chicago, IL, 2004, pp. 422–426.

[101] A. Avritzer, E. J. Weyuker, “The role of modeling in the performance testing of e-commerce applications,” *IEEE Transactions on Software Engineering*, vol. 30, no. 12, pp. 1072 – 1083, 2004.

[102] G. Ruhe, A. Eberlein, D. Pfahl, “Trade-off analysis for requirements selection,” *International Journal of Software Engineering and Knowledge Engineering*, vol. 13, pp. 345 – 366, 2003.

[103] K. Choi, S. Jung, H. Kim, D. Bae, D. Lee, “UML-based modeling and simulation method for mission-critical real-time embedded system development,” in *Proc. of the IASTED International Conference on Software Engineering*, Innsbruck, Austria, 2006, vol. 2006, pp. 160 – 165.

[104] N. Smith, A. Capiluppi, J. Fernandez-Ramil, “Users and developers: An agent-based simulation of open source software evolution,” *Lecture Notes in Computer Science*, Shanghai, China, 2006, vol. 3966, pp. 286 – 293.

[105] M. Ruiz, I. Ramos, M. Toro, “Using dynamic modeling and simulation to improve the COTS software process,” *Product Focused Software Process Improvement*, vol. 3009, F. Bomarius and H. Iida, Eds. Springer-Verlag Berlin, 2004, pp. 568–581.

[106] D. Pfahl, A. Birk, “Using simulation to visualise and analyse product-process dependencies in software development projects,” *Product Focused Software Process Improvement*, vol. 1840, F. Bomarius and M. Oivo, Eds. Springer-Verlag Berlin, 2000, pp. 88–102.

[107] D. Sycamore, J. S. Collofello, “Using system dynamics modeling to manage projects,” in *Proc. IEEE Computer Society’s International Computer Software and Applications Conference*, Phoenix, AZ, USA, 1999, pp. 213 – 217.

[108] E. Katsamakas, N. Georgantzas, “Why most open source development projects do not succeed?” in *First International Workshop on Emerging Trends in FLOSS Research and Development*, FLOSS’07, Minneapolis, MN, United states, 2007.

Year	References	No. of papers
1986	[19]	1
1988	[3][15]	2
1989	[16]	1
1990	[17]	1
1991	[18][37][45][72]	4
1992	[68]	1
1995	[96]	1
1996	[71]	1
1997	[50][57][66][92]	4
1998	[51][69]	2
1999	[20][33][89][90][91][107]	6
2000	[10][81][99][106]	4
2001	[4][6][27][29][44][76][97][98]	8
2002	[32][35][88]	3
2003	[11][40][43][46][95][102]	6
2004	[5][8][28][47][49][53][61][73][86][100][101][105]	12
2005	[7][12][39][41]	4
2006	[9][14][24][26][60][62][67][79][80][84][93][103][104]	13
2007	[23][31][42][48][52][54][59][70][75][77][78][108]	12
2008	[1][2][13][25][34][36][38][55][64][65][87]	11
2009	[21][30][58][82][93]	5
2010	[56][63][74][83][85]	5
2011	[22]	1