Simulation Based Studies in Software Engineering: A Matter of Validity

Breno Bernard Nicolau de França, Guilherme Horta Travassos

ESE Group / COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil
{bfranca,ght}@cos.ufrj.br

Abstract. CONTEXT: Despite the possible lack of validity when compared with other science areas, Simulation-Based Studies (SBS) performed in the context of Software Engineering (SE) have supported the achievement of some results in the field. However, as it happens with any other sort of experimental study, to increase their validity and strength confidence in the results the threats to validity must be identified and dealt. OBJECTIVE: To identify potential threats to SBS validity in SE and suggest ways to mitigate them. METHOD: To apply a secondary qualitative analysis in a dataset resulted from the aggregation of data from a *quasi*-systematic literature review combined with information surveyed *ad-hoc* regarding other science areas. RESULTS: 28 different threats to validity and concerned with SBS in SE were identified and classified according Cook and Campbell's categories. Besides, 12 verification and validation procedures applicable to SBS were analyzed to understand their ability to detect these 28 threats to valitidy. After this, the observed behaviors were used to improve guidelines regarding the planning and reporting of SBS in SE. CONCLUSIONS: Simulation based studies have different threats to validity when compared with traditional studies. They are not well known. Therefore, it is not easy to realize all of them without explicit guidance yet. So, recommendations and guidelines are presented in this paper to support their identification and mitigation.

Keywords: Simulation-based studies, simulation models, threats to validity.

1 Introduction

Simulation-Based Studies consist of a series of activities aiming at observing a phenomenon instrumented by a simulation model. Thomke [2] reported the adoption of SBS as an alternative strategy to support experimentation in different areas, such as automotive industry and drugs development. Criminology is another field where researches have taken place with the support of SBS [3].

In the direction of these potential benefits, Software Engineering (SE) community has presented some initiatives. Indeed, apart from some interesting results, the SBS presented in the context of SE [4] allowed us to observe its initial maturity stage when compared with SBS concerned with the aforementioned areas. Lack of research protocols, *ad-hoc* experimental designs and output analysis, missing relevant information on reports are some examples of issues that can be observed into this context.

Based on the findings of our review [4] and also on existing Empirical Software Engineering (ESE) and simulation guidelines from other research areas, we proposed a preliminary set of guidelines aiming at providing guidance to researchers when reporting SBS into the SE context [5]. These guidelines have evolved to comprehend planning issues such as the problem, goal, context and scope definitions; model description and validation; experimental design and output analysis issues; the supporting environment and tools; and reporting issues such as background knowledge and related works, applicability of results, conclusions and future works.

Another expected contribution of the guidelines' application is the reduction (or identification) of potential threats to validity that may bias the study. However, in order to perceive such reduction we believe it is worth to know some common threats already reported by the SE community when performing SBS. Therefore, we have conducted a secondary analysis of the potential threats to validity found in those SBS, which we are now presented in this paper. As far as we are aware, there is no other work like this into the context of experimental software engineering involving SBS. Such threats to validity compose the body of knowledge, organized as the new version of the proposed guidelines. Additionally, we have related these threats to Verification and Validation (V&V) procedures for simulation models previously identified in the technical literature in order to illustrate how to deal with such threats in SBS. Finally, we deliver some recommendations for using such knowledge when planning and reporting SBS, which also compose a bigger set of guidelines (in progress).

The remaining sections of this methodological paper is organized as follows: section 2 presents the adopted research methodology; section 3 presents the threats to validity identified through a qualitative analysis performed on a set of SBS, both in the SE technical literature and papers from other areas discussing this subject; section 4 presents a list of technical V&V procedures applicable for simulation models; section 5 presents an analysis on how the threats and the V&V procedure relate in order to provide more reliable SBS and deliver some recommendations in this sense. Finally, section 6 presents the final remarks and the way ahead.

2 Research Methodology

Recently, we undertook a *quasi*-Systematic Literature Review aiming at characterizing SBS into the context of SE [4]. It identified 108 relevant research papers, from which 57 consist in SBS. The remaining papers rely on simulation model proposals.

After analyzing the content of these 57 studies, we proposed a preliminary set of reporting guidelines for SBS in Software Engineering, with the purpose of orientating researchers in such simulation studies activities [5]. Also, these guidelines should help researchers to identify (*a priori)* potential threats to the study validity. For that, in the current paper we performed a secondary analysis over the 57 studies, making use of some procedures of a qualitative approach, namely the Constant Comparison Method [6] to identify common threats across the studies. Additionally, we search for research papers outside SE discussing threats to validity in SBS and included two more threats in this analysis. It was necessary to define an additional information extraction form, containing the study environment, whether *in virtuo* or *in silico* [1], and the potential threats description (identified in the research papers as limitations, assumptions or threats to validity). The environment is important since *in virtuo* contexts are supposed to be risky, mainly by the involvement of human subjects.

This way, first we extracted the threats to validity descriptions and grouped them by paper. After that, we performed an initial (open) coding, assigning concepts to chunks of the whole extracted text. So, for each new code, it was compared to the other ones to understand whether it was or not about the same concept. Furthermore, we reviewed the codes and then started to establish relationships among codes (axial coding) to generate reasoning about the threat description. This way, each reasoning is written as a threat to validity, which the code represents the name of the threat in the next section. Finally, we grouped these open codes into categories (with an initial set based on the classification proposed by [7], but that could be extended in case we understand that it was needed). Note that no selective coding was performed, since the main goal was to identify and categorize the threats to validity.

The main result of this secondary analysis is a list containing the potential threats to SBS validity, labeled using the grounded codes and organized according the classification proposed by Cook and Campbell, as presented in [7].

Additionally, we performed an analysis by matching threats to validity and V&V procedures for simulation models. The goal of such analysis is to identify whether the procedures can fully prevent from threats occurrences. And, finally, deliver some recommendations on how to avoid them, all grounded on the findings of the systematic review and additional information collected from the literature on Simulation.

3 Threats to Simulation Studies Validity

The identified threats to validity are organized according to the classification presented in [7], in the following subsections. The title (in bold) for each threat to validity reflects the generated codes (categories) in the qualitative analysis. We do not present the analysis in detail, but concentrate on their results. It is important to notice that we did not analyze threats of validity for each study, but only collected what have been reported. Indeed, it is possible to observe many other potential threats to validity in each study, but we decided not to judge them based on the research paper only. For sake of avoiding repeating threats already discussed in others Experimental Software Engineering forums, we will concentrate on threats more related to *in virtuo* and *in silico* studies and not discussed on SE papers yet.

3.1 Conclusion Validity

This validity refers 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. 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 [3]**: different from the threats previously mentioned, this one was observed 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. An example of this threat can be observed in [3], where the authors 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 being examined. 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 [3]**: also observed into the *in silico* context, this threat involves the use of inappropriate procedures for output analysis. It should be avoided comparing single values from simulated to empirical outcomes. It is recommended to use proper statistical tests or measures to compare distributions with a certain level of confidence.

Other threats to conclusion validity are both observed at in virtuo environments: a small population sample hampering the application of statistical tests, which is similar to the one mentioned by Wohlin et al [7] as "Low statistical power"; and the uneven outcome distribution (high variance) due to purely random subjects assignment, which is mentioned in [7] as "Random heterogeneity of subjects".

3.2 Internal Validity

This validity refers to the assurance that the treatment causes the outcome, rather than an uncontrolled external factor, i.e., avoid the indication of a false relationship 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 supporting data, human subjects manipulating the model when performing in virtuo experiments, or bias introduction by the simulation model itself. Thus, the main internal validity identified threats in SBS are:

- **Inappropriate experimental design (missing factors) [15-18]**: 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. So, dropping out factors to simplify the experimental design, as in fractional factorial designs, should be done with caution.
- **Simulation model simplifications (assumptions) forcing the desired outcomes [3,10,12,20,23,24]**: this is the most recurrent threat reported in the analyzed papers. Often occurring into the *in silico* context, it is concerned with the simulation model itself. In this threat, the simulation model contains assumptions implemented in a

way that they directly produce the desired effect, establishing the intended behavior or hypothesis as truth directly from the input to output variables. For instance, in one of the six times we observed this threat (reported as an assumption) the authors 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 production decreases, the quality increases, and the maintenance time decreases are directly introduced in the model as assumptions. It goes in the wrong direction of SBS, where there is a theory with a defined mechanism that explains a phenomenon, i.e., how these interactions between variables occur. In such case, there is no room for simulation, since the outcomes are predictable without run the simulations. Such a *black box* (without mechanisms) approach is the typical situation where *in vitro* experiments are more applicable.

- **Distinct datasets (context) for model calibration and experimentation [21]**: it is difficult to realize how a controlled computer environment (*in silico*) may be influenced by external or disturbing factors. Nevertheless, the supporting dataset, often required by the simulation models, may disturb the results whether data form different contexts are been compared. This is the case when a specific dataset, reflecting the context of a particular project, product, or organization is used to calibrate a model that will be used in an experiment for another distinct context. For example, 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, similar to the ones already mentioned in [7]: lack of SE knowledge hiding possible implications due to unknown disturbing factors; insufficient time to subjects familiarize with the simulation tool; premature stage of the simulation tool (instrumentation effect); non-random subjects' dropout after the treatment application (mortality); different number of simulation scenarios (instruments) for each treatment and available time to their performing (instrumentation effect); maturation effect by the application of same test both before and after treatments; different level of expertise required by the instruments for both control and treatments groups (instrumentation effect).

3.3 Construct Validity

This validity refers to the assurance that 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 in the study. Davis et al [8] claim that the nature of simulation models tends to improve construct validity, since it requires formally defined constructs (and their measurement) and algorithmic representation logic 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 [15-18]**: this happens when simulation models are compared to any other kind of model not only in terms of their output variables, but also in nature, like analytic models.
- **Inappropriate application of simulation [15-18]**: in the *in virtuo* context, it is possible to identify situations where the model building could be more effective than the model usage, considering that SBS involve both parts.
- **Inappropriate cause-effect relationships definition [10]**: this threat is associated to the proper implementation of the causal relationships between simulation model constructs explaining the mechanism under study.
- **Inappropriate real-world representation by model parameters [10]**: the choice of input parameters should reflect real-world situations, assuming suitable values that can be observed in practice and ware worthy for the analysis.
- **Inappropriate model calibration data and procedure [10]**: 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 up or down the effects.
- **Hidden underlying model assumptions [10]**: if assumptions are not explicit in 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 [19]**: this threat is related to 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 must be checked later, it is not an adequate “device” by which one can reduce model complexity or scope.
- **The simulation model does not capture the corresponding real world building blocks and elements [10]**: this threat concerns with 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 [9]**: This threat is related to simulation studies where a simulation model is chosen without proper information about its validity. So, no conclusion can be draw about the phenomenon, but only about the model itself. So, the simulation model plays the role of an object of study, rather than an instrument. As an example, the authors in [9] 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].”

The inappropriate measurement for observed constructs is also observed in SBS. Wohlin [7] 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.

3.4 External Validity

This validity involves the possibility of generalization of results outside the experimental settings scope. In simulation studies, it is particularly interesting to know if results can be also observed in different simulation studies (called simulated external validity [3]) or it can predict real-world results (called empirical external validity [3]). Thus, the five identified (all into the *in silico* context) threats to external validity are:

- **Simulation results are context-dependent, since there is a need for calibration [10]**: simulation modeling involves the definition of both conceptual and executable models. So, to run simulations, the model needs to be calibrated with data representing the context where the conclusions will be drawn. Results are as general as the supporting data. In other words, if simulations are performed with specific organization, project, or product data, results will only be applicable to that scope.
- **Simulation may not be generalizable to other same phenomena simulations [3]**: 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 can be shown in such 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 [3,10]**: when simulation outcomes sufficiently differ from empirical outcomes, we may say that simulated results have no external validity. One example of such threat was said in [10]: "*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 [8,22]**: 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 [8].

4 Verification and Validation of Simulation Models

Among the possible approaches to avoid the occurrence of the threats to validity mentioned in the previous section we have: the procedures adopted to verify and validate the simulation model and the experimental design. It is reflection of the nature of computer-based controlled environment, where the phenomenon under investigation is observed essentially through the simulation model execution. This way, the only possible changes are on the input data or the simulation model. Consequently, the validity aspects concentrate on both the simulation model validity and data validity. For the scope of this paper, we are considering only issues regarding the model validity impacting on the study validity. Also, it is relevant to mention that we made no analysis regarding the possible interaction among these threats to validity, in the sense that mitigating one threat may impose on the occurrence of others. However, we believe that threats related to model validity and those which can be mitigated by the application of V&V procedures do not present this sort of property, since these procedures when performed together increase the level of validity, having no impact on the results of applying any of them. Maybe other kind of threats, like the one caused by issues on the experimental design or supporting data may present side effects.

Since the SBS validity is highly affected by the validity of the simulation model, using a model that cannot be considered valid will bring invalid results, regardless the mitigation actions applied to deal with other possible validity threats. In other words, the simulation model itself represents the main threat to study validity.

In [4], we identified nine verification and validation (V&V) procedures applied to simulation models in the context of SE. Besides, we merged these procedures with the ones in [11], which are twelve V&V procedures often performed for discrete-event simulation models in several domains. In fact, Sargent [11] presents fifteen procedures for V&V. However, we understand that three of them are useful instruments to perform verification and validation activities, rather than procedures or techniques. These three procedures are: use of animations to graphically display the model behavior, operational graphics to present values for the model variables and outputs, and traces of the simulation runs to describe the whole variables change in every cycle. This way, Table 1 presents the merge from the remaining thirteen with the procedures identified in the systematic literature review, and the merge process was based on the reasoning about procedures' descriptions, where some of them were grouped together.

The procedure "Comparison to Other Models" was found both in the review and the list presented by [11]. Besides, we merged the software testing related procedures together in the procedure "Testing structure and model behavior", where we grouped "Degenerate Tests" and "Extreme Condition Tests", from [11].

Face validity is an expert-based evaluation approach. However, it does not have a systematic script or a set of steps. A review, an interview or even a survey may work in the same way, asking the expert about how reasonable that model and outputs are.

Most of comparisons among simulated and actual data rely on historical or predictive validation. Sargent also mentions a group called "Historical Methods", which is composed by three V&V approaches for simulation models: Rationalism; Empiricism, that "requires every assumption and outcome to be empirically validated"; and Positive Economics, that "requires that the model be able to predict the future, rather than concerned with model's assumptions or causal relationships (mechanism)".

We agree that Rationalism may contribute in V&V of simulation models. However, for the empiricism, it has a general description and seems to be just a characteristic or a type of verification, since such it can be reworded as the Historical Validation or Predictive Validation procedures, for instance. It is also true for the Positive Economics, being a matter of perspective or abstraction. Finally, Sargent [11] also presents the "Multistage Validation" procedure that consists in performing the "Historical Methods", namely, Rationalism, Empiricism and Positive Economics sequentially.

As an example of application of such V&V procedures, Abdel-Hamid [12] submitted his model to several of them. The basis for developing his Software Project Integrated Model, using the System Dynamics (SD) approach, was field interviews with software project managers in five organizations, supplemented by an extensive data

Table 1. 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 exist, 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's 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 structure and model 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.

base of empirical findings from the technical literature. Additionally, tests were performed to verify the fit between the rate/level/feedback structure of the model and the essential characteristics of the real software projects dynamics. The project managers involved in the study confirmed this fit. However, the procedures for tests and reviews performed were not described in the paper. Besides, the results were not reported either. So, one may ask among other questions, "*What kinds of test were performed? How many discrepancies were identified by the project managers?*"

Another procedure performed was the comparison against reference behaviors. In this case, the behavior was textually and graphically described and the model representation was presented in System Dynamics diagrams. The reference behavior in this case is the 90% syndrome, where developers use to miscalculate the required effort for a task, and always underestimate it.

Also, the simulation results in [12] were plotted in sequence run charts to compare against the expected behavior. Thus, the results seem to indicate the fit between the reference behavior and simulation results. Reference behaviors reproduced by the model included a diverse set of behavior patterns observed both in the organizations studied as well as reported in the literature.

The author also reports extreme condition simulations, i.e., to "test whether the model behaves reasonably under extreme conditions or extreme policies" [12].

Additionally, the author conducted a case study at NASA. According to him, the DE-A project case study, which was conducted after the model was completely developed, forms an important element in validating model behavior as NASA was not part of the five organizations studied during model development.

It is important to note, as also pointed out by the author, that one of these procedures alone may not provide enough validity for this simulation model. However, taking them together can represent a solid group of positive results [12].

5 Recommendations for the improvement of Simulation Studies

Considering the V&V procedures mentioned in the previous section, now we relate them to the threats to validity identified in the context of SE simulation studies (section 3). The goal of such matching is to provide explanations about how different bias imposed by the threats can be avoided by performing specific V&V procedures and also to highlight that the use of such procedures cannot avoid all the threats to simulation studies validity. From these explanations, we make some recommendations to guide researchers for SBS planning.

The V&V procedures presented in the previous section can be divided into two perspectives: black and white box. The Face Validity procedure is the only one from Table 1 with a white box perspective. Such procedure enables the investigation of internal properties and behaviors of a simulation model, rather than dealing with it as a black box, in which just the combinations of input and output are evaluated. Thus, threats to construct validity, involving the mechanisms that explain the phenomenon captured by the simulation model, may be identified in advance by domain experts.

Examples of such threats are the inappropriate definition of cause-effect relationships and failure on capturing the corresponding real world building blocks and elements.

Recommendation 1*. Make use of Face Validity procedures, involving domain experts, to assess the plausibility of both conceptual, executable models and simulation outcomes, using proper diagrams and statistical charts as instruments respectively.*

The first threat can also be better avoided when the model propositions or causal relationships are based on empirical evidence, what sounds good to have at least one empirical evidence report regarding the embedded cause-effect relationships.

Araújo *et al* [13] performed a set of systematic reviews aiming at reinforcing the validity of their SD model for observation of software evolution. In that opportunity, the reviews supported the identification of sixty reports of evidence for different relationships among the characteristics (e.g., eight reports of evidence for the relationship between characteristics Complexity and Maintainability) defined in their model.

Recommendation 2. *Try to support model (causal) relationships, as much as possible, with empirical evidence to reinforce their validity and draw more reliable conclusions.*

Using Face Validity in combination with Sensitivity Analysis can assist the proper selection of model's input parameters. Parameters that are sensitive should be made accurate prior to using the simulation model.

Recommendation 3. *Use results from Sensitivity Analysis to select valid parameters' settings when running simulation experiments, rather than model "fishing".*

In the same sense, it can be used along with the Rationalism to assess model's assumptions regarding the underlying concepts. However, when the model assumptions are hidden or not clearly stated, no Face Validity can be applied. For these cases, procedures like Comparison to Reference Behaviors and Testing Structure and Model Behavior are more suitable. The baseline or expected behaviors can give insights about the hidden model assumptions are affecting its results.

Recommendation 4. *Always verify model assumptions, so the results of simulated experiments can get more reliable.*

From the black box perspective, Comparison to Reference Behaviors can also help to identify situations where simulation results differ from the outcomes of empirical observations. However, for those cases in which there is a mismatch between the simulated and empirical outcomes, procedures like Historical Data Validation and Predictive Validation are more suitable, as long as enough data is available.

Recommendation 5. *When comparing actual and simulated results, be aware about data validity and that the data been compared came from the same or similar measurement contexts.*

Black box approaches may also assist on data validity issues. Often, simulation models have a calibration procedure, and using it inappropriately may cause strange behaviors or invalid results. Turing tests may help with these situations, once these results should resemble the actual ones. If a phenomenon expert cannot identify such a difference, the results have a minimum degree of confidence.

For stochastic simulations, these models have their particularities and the main difference that has to be validated is the amount of internal variation on the outcomes. The threat of considering only one observation when dealing with stochastic simulation, rather than central tendency and dispersion measures can bias or blind the user or experimenter on the interpretation of results. The V&V procedure "Internal Validity" (the term adopted by Sargent is overloaded with the Cook and Campbell classification of threats to validity, but they have complete different meanings) helps on the understanding and measuring the amount of internal variation of stochastic models by running the model with the same input configuration and calculating both central and dispersion statistics. The results should be compared to real phenomenon observations to understand whether the both amounts of variation are proportional.

Performing one procedure or another can bring some validity to the study. The simulation models should be considered valid, based on evidence regarding its validity. It is important for not reducing the findings only to the simulations themselves.

Recommendation 6. *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, augmenting the precision of results.*

Understood that V&V procedures may help to perform more confident simulation studies, it should also be pointed out that it is not a silver bullet. We still can mention a series of threats that do not directly relate to such procedures, but to the adopted experimental design for the study. For instance, not using proper statistics when comparing simulated to empirical distributions; missing factors; distinct datasets (context) for model calibration and experimentation; naturally different treatments (unfair) comparison; inappropriate use of simulation; simulation results are context-dependent, since there is a need for calibration; and simulation may not be generalizable to other simulations of the same phenomena.

Recommendation 7. *When designing the simulation experiment, consider as factors (and levels) not only the simulation model's input parameters, but also internal parameters, different sample datasets and versions of the simulation model, implementing alternative strategies to be evaluated.*

At last, there is a recurrent threat to construct validity that is hard to identify: the simulation model simplifications (assumptions) forcing desired outcomes. From the viewpoint of simulation outputs compared to empirical observations, this one does not represent any threat. When both empirical and simulated and values are statistically similar, everything seems to be perfect. The problem lies on such limited black box view. The reason for reaching the desired output cannot be explained by a reasonable causal model or mechanism, but an explicitly generation from the input parameters to the output variables. So, one is not capable of explain how to get such outcomes in real life, since there is no mechanism for a theoretical explanation. In summary, there is no way of make interventions to reproduce such behavior in real world, because the reasoning is missing and the result has probably occurred by chance. Comparison-based procedures cannot capture this type of threat. Just white box procedures like Face Validity involving simulation experts may help to identify such threat.

6 Final Remarks

Taking simulation as a complementary research strategy for the evolution of Software Engineering knowledge, mainly in contexts where *in vivo* or *in vitro* experiments are unfeasible or risky, researchers should be aware about possible threats involved in this sort of study. The results reported on this paper advance the current state in ESE, by exposing such threats to SBS validity and matching them to V&V procedures for simulation models. Besides, seven recommendations, all of them grounded in technical literature acquired data, emerged for planning the tasks intending to reduce the possibility of occurrence of threats to validity.

We believe that the identification and compilation of such threats complemented by their discussion and analysis offers an evolved perspective that can contribute for the maturity of SBS, where most of time the main tasks have been performed *ad-hoc* due the lack of orientation, especially regarding model experimentation. Additionally, the possibility of detecting some of these threats by using V&V procedures, the understanding of how to avoid them and presenting a set of recommendations configure an interesting contribution. As far as we are aware, there is no other work offering this sort of discussion in the experimental software engineering technical literature.

Our efforts have been directed by the organization of knowledge available in the technical literature regarding simulation studies in SE through secondary studies. This organization involves synthesis and knowledge representation as guidelines for the planning and reporting of SBS, which is not a simple task.

As future directions we are investigating how the Design of Experiments can contribute to improve the quality and confidence of simulation based studies in SE. Not only in the perspective presented by [9] and [14], but also as an enabler to explore more ambitious results than just anticipating *in vitro* and *in vivo* experiments.

Acknowledgment. Authors thank CNPq for supporting this research with research grants.

References

1. Travassos, G. H.; Barros, M. O.: Contributions of *In Virtuo* and *In Silico* Experiments for the Future of Empirical Studies in Software Engineering. In: WSESE03, Fraunhofer IRB Verlag, Rome (2003)
2. Thomke, S.: Experimentation Matters: Unlocking the Potential of New Technologies for Innovation. Harvard Business School Press, Boston (2003)
3. Eck, J. E.; Liu, L.: Contrasting simulated and empirical experiments in crime prevention. J Exp Criminol, 4, 195-213 (2008)
4. de França, B. B. N.; Travassos, G. H.: Are We Prepared for Simulation Based Studies in Software Engineering Yet? CLEI eletronic journal, vol. 16, n. 1, paper 8 (2013)
5. de França, B. B. N.; Travassos, G. H.: Reporting guidelines for simulation-based studies in software engineering. In: 16th International Conference on Evaluation & Assessment in Software Engineering, pp. 156 – 160, Ciudad Real (2012)

6. Corbin J.; Strauss, A.: Basics of qualitative research: Techniques and procedures for developing grounded theory. Sage Publications (2007)
7. Wohlin, C.; Runeson, P.; Host, M.; Ohlsson, M.; Regnell, B.; Wesslen, A.: Experimentation in Software Engineering: An Introduction. Kluwer Academic Publishers (2000)
8. Davis, J. P.; Eisenhardt, K. M.; Bingham, C. B.: Developing Theory Through Simulation Methods. Academy of Management Review, 32, 2, 480-499 (2007)
9. Houston, D. X.; Ferreira, S.; Collofello, J. S.; Montgomery, D. C.; Mackulak, G. T.; Shunk, D. L.: Behavioral characterization: Finding and using the influential factors in software process simulation models. Journal of Systems and Software, 59, 259-270 (2001)
10. Garousi, V.; Khosrovian, K.; Pfahl, D.: A customizable pattern-based software process simulation model: Design, calibration and application. Software Process Improvement and Practice, 14, 3, 165 – 180 (2009)
11. Sargent, R. G.: Verification and Validation of Simulation Models. In: Proc. of the Winter Simulation Conference, pp. 166 – 183, Baltimore (2010)
12. Abdel-Hamid, T.: Understanding the "90% syndrome" in software project management: A simulation-based case study. Journal of Systems and Software, 8, 319-33 (1988)
13. Araújo, M. A. P.; Monteiro, V. F.; Travassos, G. H.: Towards a model to support in silico studies of software evolution. In: Proc. of the ACM-IEEE international symposium on Empirical software engineering and measurement, pp. 281-290, New York (2012)
14. Wakeland, W. W.; Martin, R. H.; 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, 107–119 (2004)
15. 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, 3, 283 – 298 (2001)
16. 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, 4, 367 – 395, (2003)
17. Rodríguez, D.; Sicilia, M. Á.; 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, 4, 451–463 (2006)
18. 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. Infor. and Software Technology, 46, 2, 127–147, (2004)
19. Stopford, B.; Counsell, S.: A Framework for the Simulation of Structural Software Evolution. ACM Transactions on Modeling and Computer Simulation, 18 (2008)
20. Thelin, T.; Petersson, H.; Runeson, P.; Wohlin, C.: Applying sampling to improve software inspections. Journal of Systems and Software, 73, 2, 257 – 269 (2004)
21. Alvarez, F.; Cristian, G. A.: Applying simulation to the design and performance evaluation of fault-tolerant systems. In: Proc. of the IEEE Symposium on Reliable Distributed Systems, pp. 35–42, Durham (1997).
22. Rahmandad, H.; Weiss, D. M.: Dynamics of concurrent software development. System Dynamics Review, 25, 3, 224–249 (2009).
23. 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, 345 – 360, (2006)
24. 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, 11, 610 – 618 (2006)