Model Validation and Testing in Simulation: a Literature Review

Naoum Tsioptsias[1], Antuela Tako[2], and Stewart Robinson[3]

1 School of Business and Economics Loughborough University, Loughborough, Leicestershire, LE11 3TU, United Kingdom
n.tsioptsias@lboro.ac.uk
2 School of Business and Economics Loughborough University, Loughborough, Leicestershire, LE11 3TU, United Kingdom
a.takou@lboro.ac.uk
3 School of Business and Economics Loughborough University, Loughborough, Leicestershire, LE11 3TU, United Kingdom
s.l.robinson@lboro.ac.uk

Abstract

Model validation is a key activity undertaken during the model development process in simulation. There is a large body of literature on model validation, albeit there exists little convergence in terms of the definitions, types of validity, and tests used. Yet it is not clear what standards should be taken into consideration to avoid developing what could be considered to be invalid or wrong models. In this paper we examine existing literature on model validation with the view to identifying the existing validation approaches and types of tests used to assess model validity. In this review we focus our attention on three domains that usually overlap in methods and techniques: general Operational Research (OR), Modelling & Simulation (M&S) and Computer Science (CS). We analyze each field to identify the aspects of validity considered including the tests used, the validation approach taken, i.e. the suggested level of validity achieved (if this applies) and the reported outcome. The analysis shows that there are common validation practices used in all three fields as well as new ideas that could be adopted in discrete event simulation. Some main points of concurrence include the lack of universal validation, the continuous need for validation, and, the indispensable need for modelers and users to work closely together during the model validation process. This review provides an initial categorization of literature on model validation which can in turn be used as a basis for future work in investigating how and to what extent models are considered sufficiently valid.

1998 ACM Subject Classification I.6.4 Model Validation and Analysis

Keywords and phrases Validation, Simulation, Literature review, Types of validity, Field Comparisons

Digital Object Identifier 10.4230/OASIcs.SCOR.2016.6

1 Introduction

Validation is considered an important activity as part of the model development process [11, 15, 20, 24, 25]. The validation process is undertaken in order to ensure that the model developed is sufficiently accurate for the purpose at hand [23, 24]. The importance of model validity becomes more crucial when undertaking modelling in facilitated modelling workshops with stakeholders in what is called facilitated simulation modelling [25, 31]. If the model is not considered accurate by the client that would mean that it would not be possible to proceed with planned workshop activities. How could the model validation process be carried

5th Student Conference on Operational Research (SCOR'16).
Editors: Bradley Hardy, Abroon Qazi, and Stefan Ravizza; Article No. 6; pp. 6:1–6:11
Open Access Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

out as part of a facilitated modelling process? In order to answer this question this paper explores model validation, definitions, process and types of tests that are involved.

While there is a large body of literature that considers validation, the existing approaches and main concepts used differ significantly [11]. To address this lack of homogeneity, we examine existing literature on the topic of model development and validation focusing on simulation. This paper presents a different categorization of validity yet our analysis corroborates previous studies. More specifically, we expand the differences not only to Operations Research (OR) but also to Modelling & Simulation (M&S) and Computer Science (CS). We identify definitions, approaches and tests utilized in existing literature, as well as the validation levels achieved – if applicable. In view of the theoretical background analyzed, we focus our attention on authors recognized for their fundamental contributions and papers with multiple citations. Additionally, we categorize and report validity types in order to establish a basis for validating models in facilitated modelling.

The aim of this paper is to acquire a better understanding of model validation by exploring three different fields (OR, M&S, and CS) with the view to identifying common practices and areas for improvement of validation in simulation. The work presented here, explores the process and approach taken during validation with the view to understanding how the decision that a model is valid is derived. This in turn can inform how wrong models developed during facilitated workshops could still be used to create learning amongst workshop participants.

This paper is structured as follows: Section 2 presents the existing work on model validation, the criteria employed for selecting papers and the areas of investigation. Section 3 provides the analysis of the literature review in order to define commonalities and differences (definitions, validity types, tests, and other aspects of validation), also demonstrating our categorizations of the concepts involved. Section 4 concludes the study with limitations and future research directions.

2 Existing work and literature selection criteria

There are very few studies where authors have undertaken a literature review analysis on model validation to form a coherent description of the differences and to infer commonalities. Pala et al. [21] compare between Hard OR, Soft OR and System Dynamics by mentioning aspects and functions of validation per case. Another relevant study [11] categorizes tests by creating a hierarchy pattern in order to identify the point at which to cease the validation process using a heuristics approach. Nance and Arthur [18] compare three life-cycle models in simulation and highlight the lack of software requirements engineering employment in M&S studies. A second more relevant study researches social sciences' validation, explaining the many problems that occur but doesn't specify techniques [13].

These studies have approached model validation from different perspectives. We address problems and definitions encountered, along with providing a taxonomy of certain ideas. In some occasions these are not clear and the classifications are made based on the authors' interpretation and terms used, especially for types of validity and analysis.

The papers included in our review are selected based on three different criteria: (i) demonstrate a modelling process/life-cycle including validation, (ii) present tests that can be used in validation and relevant procedures (e.g. [2]), (iii) contain examples or cases where validation was implemented explicitly in the paper (quantitatively or qualitatively). The selection or exclusion of the papers analyzed here was performed based on their content of validation and tests as specified from these criteria. Since we are not approaching a systematic review, but rather focus on the most important authors especially in theoretical explanations, our aim is to provide a clear understanding of the main concepts suggested thus far. To the best of our knowledge, we have sought to include all of the key sources attributed to model validation. Three fields are investigated in this paper: general Operational Research (OR) – also including management and references to other social sciences, Modelling & Simulation (M&S), and Computer Science (CS). The reason we choose these specific domains is due to their overlap with simulation – as it can be indicated by various studies we utilize in the literature (e.g. [34, 27, 12] for CS and [14, 21] for OR). Our selection is justified initially by the apparent use of M&S techniques in OR, and subsequently by the commonalities existing between CS and M&S. More specifically, Balci [2] mentions important differences that "*exist between simulation software engineering and other types of software engineering*" (an opinion shared by M&S [32] and CS [5] papers) but concludes that they can still be applied sideways. Based on this opinion, we could extrapolate the work of Zelkowitz & Wallace [36] where the authors investigate and categorize literature on technology validation, and consider that CS validation outcomes – although different as they may seem – can still be acknowledged as useful and relevant in OR and M&S.

3 A literature review analysis

In this Section we give the most usual definitions and types of validity found in the literature. Also, we comment on the tests used and other aspects relevant to validation as identified in literature.

3.1 Definitions

The evolution of OR – described by Landry et al. [14] – witnessed many additions to the concept of validation. Other than verification, which is mostly attributed to the programming part of model development, we encounter the ideas of independent verification and validation (IV&V) [29], validation, verification, and testing (VV&T) [2], terms like utility [33, 9], certification [33], representativeness [14], assessment [9, 32], accreditation [4], etc. Unlike OR and M&S, the CS papers examined do not try to distinguish that far into the different facets of validation and relevant concepts – even the ones reviewing literature.

Validation itself is viewed as a process and an evidence for "*building the right model*" [2, 23, 20, 13]. One of the earliest examined works – still tracing OR back to its military roots – emphasizes how validation involves the comparison of model outputs with our knowledge of the real world or system [33]. This idea is repeated in other OR [9, 10] and CS [22, 5, 27, 13] adaptations. Sargent states how a model should produce adequate accuracy for its intended use and experimental pre-conditions [28, 29]. This view is also encountered both in the early 1980s [14, 10, 30] and in more recent papers [32, 11] where a necessary "*degree*" of adequacy is considered in order to achieve fidelity. Another suggestion is the usefulness that validation should bring to the model, along with usability, cost consideration and representativeness [9]. Groesser & Schwaninger [11] add that validation is integrated in the model building process. These different views are similar with part of Pala et al.'s analysis [21] – though they see it from a different point of view and purpose of categorization. We should also notice the absence of clear validation analysis in some papers – even those tackling theoretic grounds (e.g. [19, 30]). The same applies for most articles presenting a case study.

Verification is the next most referenced concept found in life-cycle models. It is concerned with "*building the model right*" [2, 23, 20, 13] or achieving a consistent and debugged code [10, 21, 32], to ensure the model runs as intended [33, 9, 20, 29, 13]. It is also accepted that verification regards the correct transition from the designer's conceptual description

[10, 1, 32, 29] or the mathematical/logical model [33, 9, 32] to a numerical one. Gass references both ideas in different papers [9, 10] and Thacker in the same [32]. Even more importantly, an imminent problem occurs when accounting how in social sciences there is a mixed cross-examination of the notion. This overlapping is pointed out by Hahn [13], where the author equates the terms of verification and internal validity. We find more examples in "*verificational validity*" [19, 2], or white-box used as a validity type rather than a technique [23].

Confidence and credibility, used interchangeably, have been connected with the users' idea of whether the model is considered as having credentials for use [14, 9, 10, 21, 20, 29]. Gass [9] and Pace [20] highlight the subjectivity and bias underlying this notion. Other authors [19, 20, 32] view confidence as model accuracy in quantified results. A minimum confidence threshold is suggested [14], with credibility's initial number set to nil [10].

Accreditation is explained analytically in Gass [10], as a broader idea with a comparable content to validation: specific acceptability criteria for an intended use. This is supported by other works [20, 32, 29]. Accreditation is also considered as a similar activity to certification [10, 32] and it is carried out on existing documentation [10]. Certification is considered to be the written assurance of a model's operational applicability [33, 19, 32].

Documentation, accordingly, receives more attention in earlier papers. The necessity and need for using evaluation criteria for all interested parties is pointed out [33], from the beginning of the process. Gass [9] states that it is the "*sine qua non of model validation and assessment*". Other papers supplement these statements [9, 10, 29]. Based on the introductory explanation of how literature handles validation and relevant terminology, we notice that different words are used in the same context and the same word may have different descriptions. The actual problem though isn't just the lexicographic differences that we find, but rather the confusion that arises for modellers and users.

3.2 Types of validity

We now investigate the different aspects of validity encountered in literature in the three fields. We found forty or so different names for validity types, some of which held an elaborate explanation, others were simply referenced, while the rest appeared in contextual description. We implement the same strategy and present the most common ones used, along with our categorization.

Data validity is most often cited in papers. It involves activities leading to: appropriateness, accuracy, completeness, correctness, impartiality, sufficiency, maintainability, reliability, limited cost, and availability of soft and hard, non-biased and biased data [33, 14, 19, 2, 23, 21, 29]. It becomes apparent how potential modellers would get confused over which features should receive the highest priority. Other data aspects involve its transformation [33]. Both Sargent [29] and Gass [10] reference possible constraining issues to consider. Gass [9] distinguishes between raw and structured data. Essentially, data validation should be occurring throughout any and all stages of model development providing some sort of data usage.

Conceptual validity is also considered as an important validation activity. The definitions found so far can be grouped as follows: those concerned with correctness (these include: credibility, relevance, completeness) and the assumptions of the conceptual model, those referring to sufficiency (e.g. level of details/characteristics, fidelity and scope), and the ones explaining relationships (stated and implied variables, system theories and their reasonableness) [33, 14, 9, 2, 23, 21, 20, 29]. Each problem should be validated with a specific purpose in mind and the outcome compared with the real world [9, 23, 29]. Other than these common points, we mostly find dissimilar explanations for the concept. For example, Landry et al. [14] suggest the use of multiple conceptual models to address each problem, Oral & Kettani [19] mention that this validity should be concerned with how mental databases (that is a type of database capturing the knowledge of the participants of a modelling process) are gathered and utilized, the US GAO report [33] refers to internal logic check etc.

Operational validity is another set of definitions found in the literature. Sufficiency is again found as a common concept, this time on the quality and correspondence of the model's outputs as opposed to real world's data [14, 9, 30, 23, 21, 20, 29, 13]. The main motive is to provide decision-makers with adequate information to accept or reject the model [14, 19, 30, 21]. It also takes into consideration model usability, usefulness, timeliness, synergism, justified time, effort, and costs of the model [33, 14, 19, 30].

Logical – also used as mathematical – validity refers to the translation of conceptual model, through numerical interpretation [14, 9, 19, 21]. This type's content could be considered as one of the reasons for the confusion between validation and verification, as Hahn [13] points out. Indeed, Landry et al. [14] and Pala et al. [21] attribute verification as part of validating the logical process, with the second paper mentioning how "*Sargent calls this step 'computerized model verification'*".

The final most commonly found aspect is that of experimental validity. Authors refer to this type as the process considering quality, efficiency, sufficient accuracy and robustness of solutions, mechanisms and techniques used [14, 19, 21]. The experimenting phase of model development bridges any inconsistencies between the formal model acquired from verification and the final model to be presented in the operational part. We should note once more that "*the level of insight gained*" mentioned by Oral & Kettani [19] is another indication for the need to quantify fidelity in validation.

Having explained the most usual types of validity found, we now focus our attention on other types introduced or referenced by authors. Structural validity is identified by various works [9, 4, 21] as the adequate matching in structure between model and system behavior. Not far away from that lies behavioral validity, acknowledged in the testing section [4, 21, 11] for pattern and structural examination. The idea of theoretical or theory validity is given by the USA GAO report [33] and Sargent [29] and is also mentioned in Smith [30] as bridging the real and modeled world through consistency in theories and assumptions. Gass [9] makes references to technical, model, and dynamic aspects, while Smith [30] apposes comfort, pragmatic, convergent, axiomatic, and criterion validity. We find the use of "*formulational*" and the more confusing "*verificational*" types in Oral & Kettani [19] and Balci [2]. Another set of validation types are mentioned in Hahn [13], including some similar adaptations with terminology such as "*translation*", "*content*", "*discriminate*" etc. Experimental design as part of validation seems to be a case relevant to experimental validity [2, 16, 32]. An interesting category is that of "*communicative*' and "*presentation*" VV&T by Balci [2], which we call "*representation validity*". Predictive validity has two different meanings [9, 13] and it can be also considered as a technique [14, 2]. The terms "*white-box*" and "*black-box*" are indicated both as techniques [2] and as validation steps [23]. The same occurs with face validity mentioned as a technique (e.g. [2, 29]) and as a validation type [10, 13].

From the above, it can be concluded that there is a lack of coherence between the different validation activities. We propose the following initial taxonomy for the aspects of validity encountered thus far in literature:

- Conceptual validity (C): Conceptual, Model, Theoretical, Theory, Formulational, and Specification validity
- Logical validity (L): Logical, Mathematical, and Internal validity

- Experimental validity (E): Experimental, Solver, and Experiment design validity
- Operational validity (O): Operational, Implementer, Dynamic, Comfort, Pragmatic, Convergent, Axiomatic, Criterion, Aptness, Results, Implementation, Replication, White & Black box, External, Convergent, Discriminant, Concurrent, and Predictive validity
- Behavioral validity (B): Structural, Behavioral, Legitimational, Verificational, Construct, Translation, Face, and Content validity
- Representation validity (R): Communicative and Presentation validity
- Data validity (D): Data, Technical, and Technological validity

Concepts can and may overlap, yet we choose to distinguish them based on the required clarity. No distinction can be perfect, yet we employ this grouping to divide the dissimilar phases of the development process with the scope of further using it in future works. Indeed, this classification puts together akin ideas, allowing us to proceed with the tests and approaches part.

3.3 Tests for validation and in-paper employment with results

The plethora of existing tests throughout the literature would make it impossible and out of scope to cite and describe each and every one of them. Balci [2] has created an in-depth analysis for a long list of tests presented there. We will instead have a brief introduction to their usage and discuss the results that are attributed to them.

Tests vary from simple guideline methods for conceptual implementation, to techniques on data checking, verification, experimental and other operational implementations for proving a model capable of sufficient prediction. We consider that some of these approaches can be attributed to general methodologies for the construction and validation of models, like linear programming [33], Toulmin Framework [19, 30], System Dynamics [33], IV&V/third party validation [30, 29], Bayesian analysis [32, 5], artificial intelligence, expert systems, genetic algorithms, fuzzy logic, machine learning [20] etc. We find Gass [10] stating certain criteria against which the model should be scored (based on threshold values and intensity level). Also, there are quite a few references to tests for verification (e.g. [32, 29]).

The actual issue occurs when and if papers follow suggestions for which test to use per validity type. Most of the studies that contain life-cycle models and validation types in OR and M&S provide general guidelines rather than factual data to exemplify. Some exceptions with examples exist, but are mostly challenged qualitatively [9, 30]. Barlas [4] and Gass [10] on the other hand provide examples, focusing mainly on accreditation. Case studies are significantly fewer and only employ or mention partially the theoretical dictations of the previous papers. For instance, Athanassopoulos [1] compares empirical results for a fire department, on a qualitative analysis based on Oral & Kettani's tetrahedron [19], while Longaray et al. [17] approach a Soft OR scenario. However, their results are left vague.

In contrast, an analytical explanation is rarely encountered in CS. Numerical examples on frameworks and quantified results are often presented (e.g. [6, 5, 27, 12, 26] from different areas of CS domain), while the level of validation – if any – is just mentioned or simply omitted.

To quote Groesser & Schwaninger [11], "*the existing categorization of the validation tests as well as the validation processes proposed in the literature are often perceived as too abstract and unspecific to be readily applied*". Table 1 presents the main work reviewed in order to acquire a first impression of how studies perceive tests and the level of validation. The columns depict the three fields we have examined. These are compared with regards to: the usual content found (guidelines, examples, and case studies – row "*Contents*"), the frequency that definitions appear and are explained (row "*Definitions*"), the areas of application of the cases examined (i.e. where do the authors apply their theoretic implementations or their examples – row "*Areas of application*"), examples that concern validation or a similar concept (row "*Examples*"), the level of validation achieved (row "*Level of validation*"), and tests often recurring in literature (row "*Common validation tests*"). Note that on "*Level of validation*", we cite the outcome that papers present on validation in view of their approach – in other words how the authors consider their validation level. It refers to the general validity of a model under consideration, rather than a unique type of validity. We term as "*vague*" guidelines or case studies where validation is not subjected to an actual outcome.

Based on Table 1 we can draw some conclusions on the common features of each area. We find that OR and M&S provide general guidelines rather than using data to infer validity, contrary to CS papers. This is also depicted on many "*vague*" outcomes existing in the first two fields. On the other hand, CS case studies do not expand on definitions and types of validation. We also notice how OR and M&S areas of applications have many commonalities.

The analysis indicates common validation practices used in the three fields. This could also lead to new ideas adopted in simulation, and more specifically discrete event simulation (DES). For example, we could employ the list of available techniques mentioned here to validate each step during the process and the discrete variables of simulation. Additionally, based on the lack of concrete definitions for validation levels and in combination with our proposed validation types, we could create quantifiable levels for the different validity aspects of such a discrete system in order to describe a model as "*sufficiently*" valid or "*wrong*". Finally, the variety of areas implemented especially in CS provides the opportunity for DES to be tested as well in various fields not yet explored.

The most important conclusion is that our hypothesis about the confusion occurring in test selection is found in all of the fields. This is due to the lack of practical cases in order to quantify sufficiently the appropriateness of each approach [34].

3.4 Other aspects identified in the literature

We now refer to some common ideas found in the literature.

A number of papers refer to the need for identifying acceptable ranges of degrees of validation [9, 10, 19, 30, 21, 11, 29] or adequate trust on the model's behavior and usability (i.e. not entirely correct but at least functional models) [33, 30, 23, 32, 11]. Also papers point out how study objectives should drive the validation requirements [14, 2, 34, 5, 26]. Another concept regards the clients' involvement (as for example teams of users, decision or policy makers, managers, stakeholders, etc.) where close interaction and coordination between modeller and users is recommended (e.g. [33, 14, 9, 10, 19, 30, 21, 16, 32, 29, 35]) with Balci [2] proposing the utilization of Decision Support Systems for a non-technical explanation of the outcomes. Also, Groesser & Schwaninger [11] mention that modellers learn from participating in the process. We could comment that although studies spend a lot of effort highlighting the importance of user involvement, they rarely explain or discuss the exact role of model creators or builders. For example, evaluators are mentioned as independent investigators and multidisciplinary teams for model evaluation [33]. Landry et al. [14] state how model assessors are "*a group independent of both model builders and model users*". Oral & Kettani [19] suggest different types of OR analysts. Lastly, we encounter three very important concepts. First, the lack of perfect or unique validation in models [33, 14, 9, 19, 2, 22, 23, 21, 16, 32, 5, 11, 29] leads to a trade-off in validity [1, 11]. This occurs due to the fact that high validation level of one type doesn't necessarily imply high validation level to another [14, 19, 3, 29]. Second, the need for keeping validation iterative

Table 1 Comparison between fields.

	OR	**M&S**	**CS**
Content	Mostly guidelines and/or examples	Mostly guidelines and/or examples	Mostly empirical results from case studies
Definitions	Many definitions usually divergent	Many definitions usually divergent	Lack of definitions
Areas of application	Policy analysis and decision making (e.g. public, government, military) [33, 9] Hard OR [21] Soft OR [17, 21] Systems Dynamics [21]	Simulation models (e.g. government, industry, military) [2, 6, 23, 20, 32, 34, 27, 29] Hard OR [21] Soft OR [21] Systems Dynamics [11, 21]	Simulation models [6, 34, 27] Time-domain models [22] Caching computer-memory strategies [6] Knowledge-based systems [5] Software process model [12] Social networks [26]
Examples	Expected Utility model [30] Fire department (public services) [1]	Epidemics [4] Market growth [4]	Reference generator programme [6] Blast furnace control [16] Mechanical network [34] Thermal system [34] Spot weld [5] Hypersonic wind tunnel [27]
Level of validation	Vague [33, 14, 9, 19, 21, 13, 17] Scores for validation (Poor to Superior) [10] Qualitative evaluation of validity [30]	Vague [2, 23, 21, 20, 32, 11, 29] Validation assumed as "*already passed*" by author [4]	Vague [5, 13] Mathematical conditions for validation [22] Analytical model is robust [6] Model completed and validated [16] Various outcomes – good and bad [34] Metric comparison [12, 26]
Common validation tests	Comparison with (existing) data [33, 14, 9, 21, 13] Statistical tests [33, 14, 9, 21] Face validity [33, 14, 9, 10, 21] Turing test [14] Graphics/Animation [21] Qualitative analysis [30]	Comparison with (existing) data [14, 23, 21, 20, 34, 29] Statistical tests [14, 4, 23, 21, 20, 27, 29] Face validity [14, 21, 20, 34, 29] Turing test [14, 23, 11, 29] Graphics/Animation [4, 21, 29] Qualitative analysis [4, 20]	Comparison with (existing) data [6, 34, 13] Statistical tests [34, 27] Face validity [34]

We consider papers [14] and [21] having references on both OR and M&S.
We consider papers [6], [34] and [27] having references on both M&S and CS.
We consider paper [13] having references on both OR and CS.
Balci [2] is not referenced in tests.
Simulation and general methodologies cited earlier are excluded from tests.

throughout all of the modeling process is often mentioned [33, 14, 19, 2, 3, 8, 23, 32, 5, 11, 29]. And finally, validation is considered to be inseparable from the modelling process [14, 2, 11].

4 Conclusion

This paper explored the existing literature on model validation. Papers from three scientific domains – OR, M&S, and CS – that have commonalities with simulation have been included in the analysis.

The main aspects considered were: definition of validation, the types of validation, the tests used, along with their perspective towards outcomes and suggested level of validity. The most imminent concerns that arose included the use of insufficient terminology for definitions, different types of validity, disparate use of types, confusion of validity type names with techniques, questions on which validity tests should be employed per case, the failure to establish a proper linkage between theory and practice, and the lack of empirical studies with results being vague and the level of validity undetermined.

Needless to say, this initial work is still subjected to limitations. The existing literature especially in the field of CS can be further explored. Also, more tests can be identified to create a more extensive list of techniques and matching aspects. Additional case studies are also required in OR and M&S to verify our hypothesis. The analysis indicates the common validation practices used in the three fields as well as new ideas that could be adopted in simulation, and more specifically DES. This study presents the preliminary analysis of our results. Brooks & Tobias [8] suggest using a set of criteria for selecting a "*correct*" model, contrary to Box & Draper who start from the hypothesis that "*all models are wrong*" and question "*how wrong do they have to be to not be useful*" [7]. Also, Groesser & Schwaninger [11] ask "*how can we transfer to the client the knowledge about the relationship between model structure and the behavior it produces*". Based on this, we will next look facilitated simulation modelling and the definition of models' level of validity in order to define (i) model fidelity and (ii) model sufficiency in providing useful information to clients.

References

1 A. D. Athanassopoulos. Decision Support for Target-Based Resource Allocation of Public Services in Multiunit and Multilevel Systems. *Management Science*, 44(2):173–187, 1998. doi:10.1287/mnsc.44.2.173.

2 O. Balci. Validation, verification, and testing techniques throughout the life cycle of a simulation study. *Annals of Operations Research*, 53:121–173, 1994. doi:10.1109/WSC.1994.717129.

3 O. Balci. Principles and techniques of simulation validation, verification, and testing. In *Winter Simulation Conference Proceedings, 1995.*, pages 147–154, 1995. doi:10.1109/WSC.1995.478717.

4 Y. Barlas. Multiple tests for validation of system dynamics type of simulation models. *European Journal of Operational Research*, 42(1):59–87, 1989. doi:10.1016/0377-2217(89)90059-3.

5 M. J. Bayarri, J. O. Berger, R. Paulo, J. Sacks, J. A. Cafeo, J. Cavendish, C.-H. Lin, and J. Tu. A Framework for Validation of Computer Models. *Technometrics*, 49(2):138–154, 2007. doi:10.1198/004017007000000092.

6 A. J. Bennett, T. Field, and P. Harrison. Modelling and validation of shared memory coherency protocols. *Performance evaluation*, 27&28:541–563, 1996.

7 G. E. P. Box and N. R. Draper. *Empirical model-building and response surfaces*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1987.

8 R. J. Brooks and A. M. Tobias. Choosing the best model: Level of detail, complexity, and model performance. *Mathematical and Computer Modelling*, 24(4):1–14, 1996. doi:10.1016/0895-7177(96)00103-3.

9 S. I. Gass. Decision-Aiding Models: Validation, Assessment, and Related Issues for Policy Analysis. *Operations Research*, 31(4):603–631, 1983. doi:10.1287/opre.31.4.603.

10 S. I. Gass. Model accreditation: A rationale and process for determining a numerical rating. *European Journal of Operational Research*, 66(2):250–258, 1993. doi:10.1016/0377-2217(93)90316-F.

11 S. Groesser and M. Schwaninger. Contributions to model validation: hierarchy, process, and cessation. *System Dynamics Review*, 28(2):157–181, 2012. doi:10.1002/sdr.

12 H. Gull, S. Alrashed, and S. Z. Iqbal. Validation of Usability Driven Web based Software Process Model using Simulation. *Procedia Computer Science*, 62:487–494, 2015. doi:10.1016/j.procs.2015.08.520.

13 H. A. Hahn. The conundrum of verification and validation of social science-based models. *Procedia Computer Science*, 16:878–887, 2013. doi:10.1016/j.procs.2013.01.092.

14 M. Landry, J.-L. Malouin, and M. Oral. Model validation in operations research. *European Journal of Operational Research*, 14(3):207–220, 1983. doi:10.1016/0377-2217(83)90257-6.

15 A. M. Law and D. M. Kelton. *Simulation Modeling and Analysis*. McGraw-Hill Higher Education, 3rd edition, 1999.

16 M. Le Goc, C. Frydman, and L. Torres. Verification and validation of the SACHEM conceptual model. *International Journal of Human-Computer Studies*, 56(2):199–223, 2002. doi:10.1006/ijhc.2001.0521.

17 A. A. Longaray, L. Ensslin, S. R. Ensslin, and I. O. da Rosa. Assessment of a Brazilian public hospital's performance for management purposes: A soft operations research case in action. *Operations Research for Health Care*, 5:28–48, 2015. doi:10.1016/j.orhc.2015.05.001.

18 R. E. Nance and J. D. Arthur. Software Requirements Engineering: Exploring the Role in Simulation Model Development. In *Proceedings of the Third Operational Research Society Simulation Workshop (SW06)*, pages 117–127, 2006.

19 M. Oral and O. Kettani. The facets of the modeling and validation process in operations research. *European Journal of Operational Research*, 66(2):216–234, 1993. doi:10.1016/0377-2217(93)90314-D.

20 D. K. Pace. Modeling and simulation verification and validation challenges. *Johns Hopkins APL Technical Digest (Applied Physics Laboratory)*, 25(2):163–172, 2004.

21 Ö. Pala, J. A. M. Vennix, and J. P. C. Kleijnen. Validation in Soft OR, Hard OR and System Dynamics: A Critical Comparison and Contribution to the Debate. In *The 17th International Conference of The System Dynamics Society and the 5th Australian & New Zealand Systems Conference*, 1999. URL: http://www.systemdynamics.org/conferences/1999/PAPERS/PARA199.PDF.

22 K. Poolla, P. Khargonekar, A. Tikku, J. Krause, and K. Nagpal. A time-domain approach to model validation. *Transactions on Automatic Control*, 39(5):951–959, 1994. doi:10.1109/9.284871.

23 S. Robinson. Simulation model verification and validation: Increasing the users' confidence. In *Proceedings of the 1997 Winter Simulation Conference*, pages 53–59, 1997. doi:10.1145/268437.268460.

24 S. Robinson. *Simulation: The Practice of Model Development and Use*. Palgrave Macmillan, 2nd edition, 2014.

25 S. Robinson, C. Worthington, N. Burgess, and Z. J. Radnor. Facilitated modelling with discrete-event simulation: Reality or myth? *European Journal of Operational Research*, 234(1):231–240, 2014. `doi:10.1016/j.ejor.2012.12.024`.

26 N. Ronald, T. Arentze, and H. Timmermans. Towards process validation for complex transport models: A sensitivity analysis of a social network-enhanced activity-travel model. *Computers, Environment and Urban Systems*, 55:24–32, 2016. `doi:10.1016/j.compenvurbsys.2015.09.005`.

27 C. J. Roy and W. L. Oberkampf. A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. *Computer Methods in Applied Mechanics and Engineering*, 200(25-28):2131–2144, 2011. `doi:10.1016/j.cma.2011.03.016`.

28 R. G. Sargent. Some approaches and paradigms for verifying and validating simulation models. In *Proceedings of the 2001 Winter Simulation Conference*, pages 106–114, 2001. `doi:10.1109/WSC.2001.977367`.

29 R. G Sargent. Verification and validation of simulation models. *Journal of Simulation*, 7(1):12–24, 2012. `arXiv:10`, `doi:10.1057/jos.2012.20`.

30 J. H. Smith. Modeling muddles: Validation beyond the numbers. *European Journal of Operational Research*, 66(2):235–249, 1993. `doi:10.1016/0377-2217(93)90315-E`.

31 A. A. Tako and K. Kotiadis. PartiSim: A multi-methodology framework to support facilitated simulation modelling in healthcare. *European Journal of Operational Research*, 244(2):555–564, 2015. `doi:10.1016/j.ejor.2015.01.046`.

32 B. H. Thacker, S. W. Doebling, F. M. Hemez, M. C. Anderson, J. E. Pepin, and E. A. Rodriguez. Concepts of Model Verification and Validation. Technical report, Los Alamos National Laboratory, University of California, 2004. `doi:10.2172/835920`.

33 U.S. G.A.O. Guidelines for Model Evaluation. Technical Report January, National Criminal Justice Reference Service (NCJRS), 1979. URL: `https://www.ncjrs.gov/App/abstractdb/AbstractDBDetails.aspx?id=84469`.

34 A. S. White and R. Sinclair. Quantitative validation techniques a database. (I). Simple examples. *Simulation Modelling Practice and Theory*, 12(6):451–473, 2004. `doi:10.1016/j.simpat.2004.06.001`.

35 T. R. Willemain. Insights on Modeling from a Dozen Experts. *Operations Research*, 42(2):213–222, 1994. `doi:10.1287/opre.42.2.213`.

36 M. V. Zelkowitz and D. R. Wallace. Experimental model for validating technology. *IEEE Computer*, 31(5):23–31, 1998.