Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

BETTER SAFE THAN SORRY – AN EVALUATION FRAMEWORK FOR SIMULATION-BASED THEORY CONSTRUCTION

Marvin Auf der Landwehr
Maik Trott
Maylin Wartenberg
Christoph von Viebahn

Hochschule Hannover
Faculty IV – Business and Informatics
Ricklinger Stadtweg 120
Hannover, 30459, GERMANY

ABSTRACT

Simulation constitutes a fundamental methodology that is commonly used to analyze the dynamic and emergent behaviors of complex systems. To increase faith in simulation-based theory-building, research needs to be able to demonstrate and evaluate its validity by means of true findings and correct recommendations. However, simulation credibility is pluralist, which means there are different forms of validity depending on context and domain. Hence, based on a systematic literature analysis, this study seeks to dissect simulation evaluation routines and rationales in scientific research. Having critically synthesized 1,609 articles on simulation-based theory-building, we describe the methods used to evaluate simulations. Finally, based on the literature insights, we compile two evaluation frameworks that enable researchers to plan multifarious evaluation episodes and advance a pluralist approach to simulation evaluation.

1 INTRODUCTION

Due to the fact that social and technical aspects are inextricably intertwined and thus, do not constitute clear boundaries between each other (Tolk et al. 2021), they need to be conceptualized as complex socio-technical interactions (Beese et al. 2019). Information processing depends on the mutual interplay of humans and IT artefacts, requiring dynamic, adaptive and distinctive methodological approaches to capture, study and predict their behavioral interdependencies (Tolk et al. 2021). Accordingly, simulation-based research is gaining increasing relevance in various research disciplines such as the natural sciences and information systems research (e.g,, Küppers and Lenhard 2005; Beese et al. 2019). As "method for using computer software to model the operations of real-world processes, systems, or events" (Davis et al. 2007, p. 481), computer simulation allows the isolation and variation of influencing factors, time mechanisms and boundaries of a given system within a controlled environment, while at the same time producing massive amounts of data, which aid in the description, analysis and evaluation of non-linear system relations (Law and Kelton 2000). In contrast to deductive theoretical analysis paradigms or inductive approaches, simulation modelling can be regarded as a differentiated, alternative scientific methodology, which captures analytical reflections by means of mathematical models and provides data itself (Harrison et al. 2007). Consequently, scholars have exploited the potential of the simulation methodology to advance theory construction in manifold contexts and research disciplines (e.g., Butler et al., 2014; Guo et al. 2016).

The computer simulation methodology features both, deductive (e.g., using existing theory to build a conceptual model – Winsberg 2003) as well as inductive properties (e.g., validating simulation results with

empirical data – Sargent 2013), which entails careful reflection of real-world problems as well as their theoretical implications for building a simulation model and an accurate validation of experimental results for analyzing simulation studies (Beese et al., 2019). Understanding modelling and simulation as potent problem-solving technique, the methodology's rigor and validity needs to be assessed on a structural, contextual as well as technical basis to ensure a high degree of simulation quality and provide useful results (Davis et al. 2007; Sargent, 2020). While verification and validation (V&V) processes, which have been extensively reviewed by simulation scholars (e.g., Balci, 2004; Sargent, 2013), address some of these quality-related concerns, namely the technical efficiency and correctness (verification) as well as the conceptual fidelity and credibility of a simulation artefact (validation), there are few examples of holistic regimes that support the development of evaluation strategies by synthesizing evaluation mechanisms and structuring the heterogeneous range of evaluation applications. To address the need for structural guidance and assist researchers in the identification of comprehensive evaluation episodes that suit the particular context of a simulation study, we adopt the context, content, process (CCP) view proposed by Stockdale and Standing (2006) to develop a comprehensive evaluation framework that depicts *the when, who, what and how of evaluation approaches for simulation artefacts with the purpose of theory-buildin*g.

Simulation evaluation can be classified as ex-ante, which is a formative approach concerning evaluation processes before and during simulation development, and ex-post, referring to a summative procedure for evaluation upon completion of model building and experimentation. Indubitably, formative and summative V&V can improve a simulation's overall utility, fidelity and credibility. However, the heterogeneity of available methods makes it unnecessarily difficult for simulation researchers to get an overview across the wide range of methods and identify those that suit a simulation study's particular context. "The real task of validation is finding an appropriate set of actions (Van Horn 1971, p. 257). Correspondingly, this study proposes a framework that embeds common V&V methods in a CCP perspective and supports scholars in planning evaluation episodes before, during and after simulation development. Our objective is to assist researchers by providing an extensive overview of the dimensions, elements and methods that are relevant to evaluate a simulation project. Since the absence of structural guidance on the development of evaluation strategies hampers the adoption of integrated evaluation episodes, many scholars follow the pragmatic paradigm "performance beats theoretical accuracy" (Küppers and Lenhard 2005, p. 6), which results in a lack of prediction precision, as the performance of the model on the computer is valued over its accuracy of calculation. To answer our research question and support the evaluation of simulation artefacts for theory-building, we conduct a systematic literature review (SLR) and synthesize the methods employed by researchers to verify, validate, and evaluate simulations within a recapitulatory framework. By this means, our contribution provides structural comprehension on the wide variety of V&V methods and helps to establish holistic evaluation episodes. The framework considers the CCP of evaluation and includes a set of factors that address *what* simulation elements are being evaluated, *who* is involved in the evaluation, *when* the evaluation is conducted and *how* the evaluation is to be carried out in terms of V&V.

2 RELATED WORK

To ensure "that the computer program of the computerized model and its implementation are correct" (Sargent 2013, p. 166) researchers employ a variety of V&V techniques. Verification describes the internal (i.e. technical) correctness of the computer program and its implementation, while validation is concerned with the external (i.e. conceptual) accuracy of the simulation model and its results based on the exemplary real-world system (Sargent 2020). In terms of verification, related research has proposed several techniques and frameworks to address technical issues related to the black-box nature of simulations (e.g., Tolk et al. 2021). However, existing research is scarce in respect of validation frameworks or guidelines. Literature on simulation validation generally suggests the use of multiple validation methods (Harrison et al. 2007; Sargent 2013). Testing the predictive reliability of numerical simulation models for complex systems is practically not possible, as these always embed several uncertainties, which in turn undermine the predictive reliability of a model and its results (Burton and Obel, 2011). These model uncertainties can be categorized as theoretical (difficult to understand), empirical (difficult to measure), parametrical (difficult to reproduce)

and temporal (difficult to replicate time behavior) (Oreskes 1998). Thereby, validity is not absolute but relative to the purpose of the simulation artefacts: When considering simulation as an evaluation method, it is typically not important whether the underlying model is an adequate depiction of reality; however, to test different assembly strategies in a complex production plant, a high degree of model fidelity is required to derive reliable recommendations. Hence, to develop useful evaluation guidelines, we exclusively focus on simulation artefacts that address theory-building and experimentation. Concerning the cornerstones of our research, it needs to be noted that the importance of simulation evaluation (the '*why'*) is contingent on the source of theory that forms the basis of the simulation. If the theory is mainly based on empirical evidence featuring a high degree of external validity, evaluation is less important than in cases where theory is based on non-empirical arguments or evidence from distant scientific disciplines (Davis et al. 2007).

Common V&V techniques for simulations include comparison to other models, internal validity, historical data validation, sensitivity analyses, operational graphics and predictive validation (Sargent 2013; Beese et al. 2019). Depending on the individual simulation technique, simulation use, the characteristics of the empirical target as well as associated epistemic challenges, the efficiency and usefulness of V&V methods heavily varies, requiring scholars to carefully define and select appropriate procedures. Moreover, to address simulation quality holistically, internal verification and empirical validation need to be combined in a circular process, testing both, model, experiments and results (Burton and Obel 2011, Sargent 2013). With the growing importance of simulations in scientific research, there is an increasing need for structured and substantiated evaluation processes that go beyond the scope of traditional V&V. While V&V help to assess simulation methods and experiments "in terms of their capability to deliver the intended outcomes, partially taking into account various influencing factors such as data quality or model fidelity" (Auf der Landwehr et al. 2020), individually, they are not able to assess simulation artefacts within the integrated scope of CCP. V&V techniques relate to program efficiency, model fidelity and result credibility and, therefore, lack the capability to guide simulation evaluation across multiple dimensions such as timing, subject, object and method(s). Moreover, the focus of V&V is to validate a set of new insights rather than the mechanism that generated these insights (Nan and Johnston 2009), which may negatively impact a simulations maintainability and re-usability, ultimately impairing future theory construction based on the same simulation object. Research concerning evaluation of simulation methods, models and experiments in their entirety is limited and mainly comes down to abstract guidelines (e.g., Davis et al. 2007) or V&V-specific recommendations (e.g., Sargent 2020). To the best of our knowledge, there is hardly any guidance on the structural process of evaluation for the CCP of simulation development, which would be required to operationalize evaluation strategies and thus, improve the rigor and validity of simulation-based research.

3 METHODOLOGY

To address our research question, we followed the rigorous guidelines of Webster and Watson (2002) to conduct a SLR with multiple stages. The research process opted to provide a representative set of papers from the large body of related publications and comprised a total of four SLR cycles: (1) A keyword search in top-rated journals (*A according to VHB-Rating*) from the operations research, production, logistics and information systems domains as well as (2) the proceedings of selected conferences (*WSC, ICIS, ECIS, CIRP*) and a (3) forward (i.e., identifying articles that quote relevant publications discerned during the keyword search) and backward (i.e., assessing citations from the literature results) search.

The keyword search was conducted between July 2021 and February 2022 for the fields abstract, title, and keywords across all selected journals and conference proceedings using Scopus. After excluding a total of 282 duplicates and 12 non-English or non-peer reviewed publications, we obtained a set of 1,609 papers. Subsequently, the first and second author of this paper independently read the abstract and introduction of each article and determined its relevance (i.e., coded) by assessing whether a publication is concerned with simulation evaluation in general or describes V&V methods or evaluation strategies to assess a simulation method, model or experiment in a particular application context. In this step, we excluded papers that reference simulations without linkage to V&V or evaluation mechanism or use simulation in contexts that do not involve computation. Ultimately, we identified 212 publications applying and 46 articles

conceptually assessing simulation evaluation. The coding was validated by calculating the interrater reliability, yielding an agreement percentage of 97,7% for the coded attribute 'simulation reference only' (i.e., scholars referencing simulation without engaging in evaluation), 100% for 'simulation evaluation overview' (i.e., scholars reviewing simulation V&V or evaluation irrespective of an application context) and 97.7% for 'simulation evaluation application' (i.e., scholars applying a simulation and evaluating it). Furthermore, the interrater agreement between the authors was high, featuring agreement levels calculated by Cohen's Kappa and Krippendorf's Alpha between 0.9245 and 0.9246 (Lombard et al. 2002).

Proceeding iteratively, we first assessed publications that elaborate on V&V or evaluation conceptually (simulation evaluation overview). Thereafter, we performed multiple analysis cycles to assess the individual evaluation patterns of scholars who adapted an evaluation approach to their methodologies (simulation evaluation application). Finally, we inductively drafted a categorization scheme based on the insights of 'simulation evaluation overview' articles (i.e., time, subject, object, and method of simulation), before we compiled the proposed V&V methods of all 'simulation evaluation application' publications in a concept matrix that combined the collated insights. A synopsis of our search design is given in Table 1.

Table 1: Systematic Literature Review.

Procedure	**Search framework**		**Identification of relevant articles**				
Phase	Search strings	Search fields	Total hits	Exclusion due to … non-english	no peer review	duplicates	Final sample
Keyword search journals (1)	"Simulation" AND ("Validation" OR "Verification" OR "Evaluation")	Title, abstract, keywords	378	378	378	322	75
Keyword search conferences (2)			1,438	1,438	1,438	1,213	173
Forward/backward search (3)			51/36	46/31	45/30	44/30	6/4
Total number of search results			*1,893*	*1,883*	*1,880*	*1,609*	*258*

4 EVALUATION FRAMEWORK

4.1 Identification of simulation evaluation methods

Out of the 1,609 analyzed publications, 46 articles elaborate on the conceptual side, while 212 papers actually apply V&V techniques. 339 publications fall out of the scope of this study as they report V&V in research settings that relate to the evaluation of other designs and methods. The remaining 1.012 studies that reference simulation as a major research component do not report any V&V or evaluation technique. Frequently used approaches are comparing the predictive accuracy of simulation results with real-world observations (60) or extant research (35), conducting sensitivity analyses (52), examining a model's internal validity (42), calibrating the model with empirical data (13) or comparing it to other (validated) models (25), conducting extreme tests (22) and checking operational behaviors via animations (22).

Table 2 provides an exemplary overview of publications applying simulation evaluation in terms of V&V as well as the individual method(s) used within these paradigms. For overview purposes, only selected references from the collated database are shown. In total, 95 studies employ a single evaluation technique, while 67 publications report on two different methods. In contrast, few articles reference the application of three (25), four (10) or more than five (15) methods. Regarding specific simulation techniques, sensitivity analysis (38%) and calibration (33%) are the preferred options for agent-based modelling (ABM), while predictive validity (PV) (real-world) and PV (research) are favored for analytical simulation (AS) (40% and 25%) as well as Monte Carlo simulation (MCS) (33% each). In terms of discrete-event simulation (DES), scholars commonly employ face validity (FV) (expert interview) (15%), graphical animation (15%), model comparison (15%), internal validity (18%), PV (real-world) (18%) and sensitivity analysis (18%). System dynamics (SD) studies generally build on extreme tests (40%) and sensitivity analysis (40%) and for petri-net simulation (PNS), activities are distributed equally, including different methods for event validity (EV). Hereinafter, we briefly outline the applied methods as to their specific application context in the case of simulation evaluation.

Table 2: Exemplary overview of selected publications evaluating simulations (■ addressed).

Reference	Verification						Validation																	
							EV			FV											PV			
	Calibration	Degeneracy test	Extreme test	Model comp.	Sensitivity anal.	Str. walkthrough	Observation	Real-life log	Synthetic log	Application	Exp. interview	Focus group	Narr. interview	Involvement	Graphical animat.	Hist. data validity	Internal validity	Logical reasoning	Operational graph.	Real-world	Quasi-real	Research	Traces
Aguirre-Urreta & Rönkko (2018)	■			■																			
Akkermans et al. (2021)			■		■						■		■							■			
Altazin et al. (2020)																				■			
Bapna et al. (2003)																				■			
Bitomsky et al. (2019)					■																		
Bosse et al. (2014)			■		■															■			
Brünnet et al. (2014)					■															■			
Butler et al. (2014)	■			■																■			
Canessa & Riolo (2006)					■																		
Chang et al. (2010)	■				■												■						
Choi et al. (2010)		■			■	■																	
De Vreede (1997)													■	■	■								
Domschke et al. (2022)					■																		
Dutta (2001)							■													■			
Dutta et al. (2007)					■							■											
Fuentes et al. (2021)		■			■																		
Gay et al. (2005)																	■						
Gerrits et al. (2017)						■						■			■			■					■
Glatt et al. (2019)				■																	■		
Guo et al. (2016)																	■	■					
Hauser et al. (2017)					■																		
Havakhor et al. (2018)		■		■	■																	■	
Janssen & Verbraeck (2005)						■					■				■	■				■			■
Jiao et al. (2016)											■											■	
Johnson et al. (2014)	■																■						
Koch et al. (2018)							■													■			
Konana et al. (2000)		■																					
Kontoyiannakis et al. (2009)	■	■													■	■			■				
Kropp et al. (2019)					■					■													
Kumar et al. (2008)				■																■			
Lindberg et al. (2020)	■				■										■								
Mes & Koot (2019)						■		■								■		■		■			
Nan & Johnston (2009)	■																	■				■	
Oh et al. (2016)				■												■							
Pfeiffer et al. (2016)					■			■															
Pierce et al. (2018)					■																		
Port & Bui (2009)			■		■			■							■					■			
Ren & Kraut (2011)	■				■															■			
Ridler et al. (2022)				■															■				
Ross et al. (2019)			■		■										■		■						■
Schroer et al. (2022)	■																			■			
Semelhago et al. (2021)		■			■																		
Sen et al. (2009)	■																■						
Shah et al. (2020)									■														
Smits et al. (2011)																				■		■	
Torres-Jiménez et al. (2015)																					■		
Troy et al. (2017)			■				■				■				■			■	■				
Utomo et al. (2020)					■															■			
Wawrzyniak et al. (2020)	■															■							
Wöste et al. (2021)					■																		

4.1.1 Verification Dimension

Verification is a crucial meta-component of simulation evaluation, as it addresses the internal structure and correctness of a simulation and can be applied in six forms: *Calibration* describes the integrative use of historical data or validated models for the construction of a simulation model (e.g., Butler et al. 2014), supporting a structural development routine and avoiding false dependencies or subliminal manipulations through the modeler, which is typically inherent to simulation modelling due to its positivistic nature (Mingers and Standing 2020). In contrast to *historical data validity* or *model comparisons*, which are employed summatively after model building (ex-post), *calibrations* are applied formatively during (ex-ante) the simulation development process (Sargent 2013). *Degeneracy tests* measure the degree to which individual components of the simulation model have a negative effect on the holistic model performance or behavior (Beese et al. 2019). The process entails the appropriate selection of input parameter values and is the technical (internal) counterpart to *tracing*. Our sample included 12 publications that reported on the use of *degeneracy tests* for simulation verification (e.g., Kontoyiannakis et al. 2009). *Extreme tests* assess the internal correctness of simulation models in the event of extreme or unlikely model parameters or factor combinations (Sargent 2013). If a simulation is able to behave reasonably well and produce realistic outputs even under extreme conditions, its internal structure is likely to be correct (e.g., Bosse et al. 2014; Port and Bui 2009). Similarly, *sensitivity analyses* examine a simulation model's capability to reproduce equal relationships for different value combinations of input parameters. The relationships or interdependencies can be measured on a model-related, behavioral (e.g., Lindberg et al. 2020) as well as an output-specific, downstream (e.g., Fuentes et al. 2021) level. A *structured walkthrough* includes formal and structural testing of individual model components by the model developer, either individually or by involving relevant peers/stakeholders, to assure that the model structure is consistent with the descriptive knowledge about the phenomenon that is being modelled (Choi et al. 2010). Our sample includes 13 publications where a simulation model or its conceptual representation are verified by means of a *structured walkthrough* (e.g., Janssen and Verbraeck 2005). Ultimately, simulation verification can be performed through *model comparison*. In this context, the simulation model is compared to other models in order to assess its internal correctness (Mingers and Standing 2020). To assure a high degree of robustness for this verification technique, the reference models should exclusively be objects that have been thoroughly validated themselves beforehand. Hence, we do not consider instances that lack empirical grounding or structured V&V as appropriate means of *model comparison* in our analysis, as false model constructs may potentially match false simulation constructs, leading to undesired outcomes (Havakhor et al. 2018). In our study, 42 articles employ validated models or their results to verify the internal correctness (e.g., Ridler et al. 2022).

4.1.2 Validation Dimension

In contrast to the process of verification, validation addresses the external correctness of a simulation model, its results or its propositions, i.e., whether it is a credible representation of the real system. In this study, we observed 17 methods to judge simulation validity: *EV* refers to the comparison of simulated events with events from the real world. The higher the degree of similarity between model and real system, the likelier is the simulation model to be externally correct (Sargent 2013). During our SLR, we found three distinctive methods to measure *EV*. First, simulation modelers can use or conduct *observations* to collect reference values on conceptually important events (e.g., Troy et al. 2017; Koch et al. 2018). Second, they can refer to *real-life logs*, which are recordings of event characteristics such as business process execution logs (e.g., Pfeiffer et al. 2016). Finally, also *synthetic* logs that have been produced from other models can be employed to measure the *EV* (e.g., Shah et al. 2018). In contrast to model comparison as verification procedure, *EV with synthetic logs* does not refer to technical (i.e., code) or structural (i.e., networks) model components, but the conceptual composition (i.e., event setup). Correspondingly, *FV* involves stakeholder feedback to assess whether the model and its behavior are a reasonable representation of the given system. Concerning our sample, we found 12 methods that are used to involve *FV* in the validation process of

simulations. *Expert interviews* enable scholars to collect data from expert stakeholders (e.g., Troy et al. 2017), while *narrative interviews* include a broader base of stakeholders (i.e., project managers, users) responding to the simulation model's correctness and credibility (e.g., Akkermans et al. 2021). *Involvement* of stakeholders ex-ante model building is another possibility to increase simulation validity (e.g., De Vreede 1997). Unlike *calibration* as verification method, *involvement* in the course of *EV* neglects model structure (as stakeholders are typically not familiar with formal simulation model building) and rather examines the simulation's conceptual setup (e.g., De Vreede 1997). Equally, model *application* (e.g., gamification – Kropp et al. 2019) through stakeholders can aid in ensuring that the simulation is a valid representation of the real-world system. Ultimately, *focus groups*, in which a group of stakeholders (usually six to ten) jointly elaborates the simulation validation, are another constructive measure to ensure external correctness (e.g., Dutta et al. 2007). A distinctive advantage of simulation modelling is its inherent visualization potential (Beese et al. 2019). Hence, the model's time-advancing behavior can be displayed graphically (*graphical animation*) to check the system behavior and ensure that relevant sequences or processes are represented as intended. Our sample includes 21 articles that use *graphical animation* to validate a simulation model (e.g., Gerrits et al. 2017; Troy et al. 2017). *Historical data validity* relates to the use of historical system data to build and test a model. For this purpose, given data sets are split and employed separately for model development and validation (Sargent 2013). Unlike *EV*, *historical data validity* always refers to real-world data and does not relate to behavioral elements (i.e., events) but output parameters of a simulation. Our sample contains 20 publications reporting on *historical data validity* as validation technique (e.g., Yi et al. 2019). *Internal validity* involves several replications of stochastic models to determine the statistical variability. A high degree of consistency across multiple simulation runs generally indicates a good fit between model and system, while a large amount of variability indicates less external correctness (Sargent 2013). In our sample, *internal validity* is assessed as validation measure by 43 publications (e.g., Gay et al. 2005). *Logical reasoning* deals with the logical deduction of a simulation model (e.g., based on empirical evidence or prior research), including clearly rational and reasonable modelling assumptions (e.g., Koch et al. 2018), while *operational graphics* are concerned with displaying the values of performance measures for the system over time (e.g., Troy et al. 2017). Contrary to *graphical animation*, *operational graphics* do not relate to visual behavior tracking, but to observing simulation results or key performance indicators (Sargent 2013). The most commonly used validation method in our review is the one of *PV*. It describes a simulation model's capabilities to mimic the behavior of the real-world system on a macroscopic level. Unlike *EV*, which assesses individual simulation events, *PV* examines the holistic model behavior and its aptitude to predict/replicate results obtained from system data. This system data can be *real-world* information from existing systems (e.g., Utomo et al. 2020), *quasi-real-world* specifications (i.e., assumptions from domain experts), which are artefacts from real-life objects that do not directly relate to the system to be simulated (e.g., Ridler et al. 2022), and insights from (theoretical) *research* contributions (e.g., Havakhor et al. 2018). Finally, *traces* represent the activity of tracking the behavior of specific entities in a simulation to descry the degree to which the behavior of the entity is logical and reasonable (e.g., Gerrits et al. 2017). Unlike *EV* or *graphical animation*, *tracing* focuses on individual model units rather than compiled (i.e., events) or inter-linked (i.e., processes) system states. Regarding our sample, ten articles utilized *traces* to validate the external simulation correctness (e.g., Ross et al. 2019).

4.2 Formulation of a simulation evaluation framework

Our literature analysis reveals the methodological aspects of evaluation in terms of V&V. However, reliable and comprehensive evaluation of simulation artefacts requires an overarching approach featuring interaction and linkage between CCP that allows to explore the complicated procedure of evaluation in multiple ways. Hence, to provide structural guidance and extent the scope of V&V, we propose an evaluation framework that combines the diversity of V&V methods currently employed by researchers to evaluate simulation artefacts and extent these methods by CCP-specific dimensions (see Figure 1). Simulation evaluation revolves around four items: The *when* (i.e., *time*), *who* (i.e., *subject*), *what* (i.e., *object*) and *how* (i.e., *method*) of evaluation. These four dimensions were inductively deduced from the

articles in our SLR. Accordingly, we analyzed the descriptions of the evaluation attempts by means of rationalism to develop a unified understanding of simulation evaluation across the entire sample. During the analysis, two researchers independently coded the articles from the sample based on their individual evaluation characteristics. Subsequently, the results were consolidated in four iterations. Finally, we clustered the outcomes of these iterations to reveal recapitulatory evaluation dimensions and characteristics.

Environment and setting	**Method:** ABM, AS, DES, MCS, SD, PNS ⟷	**Purpose:** Explain, discover, predict, prescribe, guide, criticize, prove

Time	**Stance**	Ex-ante																							
	Timing	Continuous																							
Subject	**Subject**	Stakeholder																							
	Involvement	Subject involved in simulation development																							
	Experience	Practitioners																							
Object	**Type**	Theory			Conceptual Model				Sim. inputs					Sim. model					Sim. outputs						
	Coverage	Exhaustive						Selective							Representative										
Method	**Criteria**	Correspondence							Coherence					Pragmatism						Consensus					
	Reference	Internal correctness						External correctness																	
	Method							EV			FV										PV				
		Calibration	Degeneracy test	Extreme test	Model comparison	Sensitivity anal.	Str. walkthrough	Observation	Real-life log	Synthetic log	Application	Exp. interview	Focus group	Narr. interview	Involvement	Graphical animat.	Hist. data validity	Internal validity	Logical reasoning	Operational graph.	Real-world	Quasi-real	Research	Traces	

Figure 1: Simulation Evaluation Framework.

Evaluation timing (i.e. when?): Simulation artefacts can be evaluated ex-ante (i.e., theory) or ex-post instantiation of the simulation model or experiments. The evaluation can be carried out continuously for the entire lifetime of the simulation project or at discrete points of time (i.e., upon completion of relevant milestones). Regardless of the timing, scholars should integrate multiple evaluation episodes throughout a single iteration of a development process to achieve a high level of assessment credibility.

Evaluation subject (i.e. who?): Evaluation can be conducted by scholars or practitioners that are involved in simulation development (i.e., modeler) or by stakeholders who are introduced to the simulation object for the sole purpose of evaluation. Involving a mix of stakeholders and modelers is particularly useful as it allows to unveil inconsistencies, flaws and errors in the use of a method.

Evaluation object (i.e. what?): Evaluation can be aimed at a variety of objects, including the underlying theory of the system (i.e., system design), the conceptual model, input data, the computerized simulation model, simulation experiments and the quantitative simulation results. Contingent on the context, an object can be evaluated on an exhaustive (i.e., entire object), selective (i.e., selected components) or representative (i.e., components relating to real-world) basis, both individually as well as collectively.

		Stakeholder	**Modeler**
Veri-fication	**Ex-ante**	Model comparison (C)	Calibration (C,I,S,T), Model comparison (C), Structured walkthrough (C)
	Ex-post	Model comparison (S)	Degeneracy test (S), Extreme test (O,S), Model comparison (S), Sensitivity analysis (O)
Vali-dation	**Ex-ante**	Face validity (C, I, T)	Logical reasoning (C), Traces (C)
	Ex-post	Face validity (S,O), Traces (S), Graphical animation. (S), Operational graphic (S), Predictive validity (O),	Event validity (S), Graphical animation (S), Historical data/ internal validity (O, S), Logical reasoning (O,S), Operational graphic (S), Predictive validity (O), Traces (S)

Legend: *C – Conceptual model; I – Inputs; O – Outputs; S – Simulation model; T – Theory*

Figure 2: V&V framework for exemplary items of time (left), subject (top) and object (parenthesis).

Evaluation method (i.e. how?): Simulation artefacts can be evaluated based on their correspondence (i.e., results) or coherence (i.e., model) to the real-world system or according to pragmatism (i.e., compliance with purpose) or consensus about the operational credibility. The evaluation process can target the structural, internal (i.e., verification) and behavioral, external (i.e., validation) correctness of the simulation object by employing various methods that we have been outlined previously.

Finally, the combinations of different timings, subjects, objects, and methods result in diverse evaluation strategies. Depending on timing, subject, object and methodic reference (verification/validation), Figure 2 suggests different evaluation methods that fit the context of the respective evaluation episode.

5 DISCUSSION AND CONCLUSION

Simulation-based research is increasingly gaining relevance in various scientific disciplines such as information systems research (which is also indicated by the large body of information systems literature in our sample) and features valuable properties for knowledge creation (e.g., creation or evaluation of design-artifacts). Evaluation is an essential activity for conducting rigorous research and applies to simulation artefacts in the same way as to other artefacts. Prior work provides an equally comprehensive and valuable body of knowledge on evaluation strategies, while leaving a gap between abstract evaluation strategies and concrete operationalization and implementation of these strategies (e.g., Kleijnen 1995, Sargent 2020). To make a genuine contribution, research focusing on simulation-based theory-building needs to take every effort to demonstrate that its representations and results are valid, both, internally as well as externally. By comprehensively reviewing current evaluation patterns for simulation artefacts, we created a framework that helps scholars to interlink CCP-related evaluation properties and extend the boundaries of (inherent) standard routines (e.g., code reviews) and detached V&V. The latter rather focus on the proposition of individual methods for assessing internal and external correctness rather than on guiding evaluation practices that address the interplay between evaluation time, subject, object and methods. While we agree with the general notion of simulation researchers that there is a broad range of V&V techniques that can be applied in different circumstances (e.g., Balci 2004; Sargent 2013), we also consider the selection of an appropriate V&V "basket" and application scope to be essential to build sufficient confidence in any given use of simulation. By bundling the multiplicity of V&V techniques and structuring them in a framework, we tie in with Mingers and Standing (2021) and support a pluralist approach to simulation evaluation that is based on coherence and rigor underlying the multifarious V&V approaches. In doing so, our study seeks to make the knowledge base on evaluation accessible to scholars that pursue to employ simulation artefacts in order to improve the quality and credibility of simulation projects. The main contribution of our study to simulation research are the proposed evaluation frameworks, which can guide scholars in applying different evaluation episodes with holistic focus on the setting, bridging the gap between standalone V&V and an integrated contextual view, namely the when, who, what and how of evaluation. Correspondingly, our study advances prior works on theory-based simulation evaluation by Davis et al. (2007), who proposed a set of evaluation guidelines in terms of 'theoretical contribution' and 'strength of method' and establishes a structural and informative basis that helps researchers to consider all stages for developing substantial simulation evaluation strategies. Anymore, our findings on the use and reporting of V&V show that many techniques are often disregarded. Thus, we want to reemphasize the need for thorough and well-documented simulation evaluation, as also stressed by other scholars (e.g., Tolk et al. 2021). From a practical viewpoint, the proposed frameworks can be employed to identify V&V methods that fit the particular needs of a given project in a structured manner, while at the same enabling practitioners to adopt a greater variety of evaluation approaches (that they may not have been aware of or used to before). In addition, the frameworks can be used to guide evaluation for simulation-components within the context of digital twin development and thus increase confidence in these systems.

Like all research, our study also holds some limitations. First, it needs to be emphasized that our SLR does not guarantee a complete overview of evaluation practices addressed by scholars, as the documentation of V&V processes is neither always applicable nor useful. Simulation-based research entails a constant trade-off between application of appropriate evaluation techniques and depth of reporting practices. While

some techniques may be regarded as standard workflow practice, others are not documented due to space restrictions or context. For instance, the study of Guo et al. (2016) has a strong grounding in empirical data. Hence, the authors have rightly chosen to leverage this grounding to build confidence in their simulation results (falling under the pragmatism criterion in our framework) rather than adapting and reporting numerous V&V practices. When using our framework, scholars need to be aware of the importance to balance evaluation choices and reporting practices based on the given context. Even though the scope of our keyword search was limited to the selected journals and conferences, we believe that our focus on high-quality publications ensures that our sample provides a representative and valuable synthesis of relevant insights. Finally, our frameworks require copious application to gain more confidence in their credibility. We reckon that this will not just aid in advancing our frameworks, but also in developing prescriptive knowledge that explains which evaluation sequences and methods are best suited in different situations.

REFERENCES

Aguirre-Urreta, M. I., and M. Rönkkö. 2018. "Statistical inference with PLSc using bootstrap confidence intervals". *MIS quarterly* 42(3):1001-1020.

Akkermans, H., W. van Oppen, B. Vos, and C. XJ Ou. 2021. "Reversing a relationship spiral: From vicious to virtuous cycles in IT outsourcing". *Information Systems Journal* 31(2): 231-267.

Altazin, E., S. Dauzère-Pérès, F. Ramond, and S. Tréfond. 2020. "A multi-objective optimization-simulation approach for real time rescheduling in dense railway systems". *European Journal of Operational Research* 286(2):662-672.

Auf der Landwehr, M., M. Trott, and C. von Viebahn. 2020. "Computer Simulation as Evaluation Tool of Information Systems: Identifying Quality Factors of Simulation Modeling". In *Proceedings of the 22nd Conference on Business Informatics*, June 22nd-24th, Antwerp, Belgium, 211-220.

Balci, O. 2004. "Quality assessment, verification, and validation of modeling and simulation applications". In *Proceedings of the 2004 Winter Simulation Conference*, edited by R .G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, 1-8. Piscataway, New Jersey: IEEE.

Bapna, R., P. Goes, and A. Gupta. 2003. "Replicating online Yankee auctions to analyze auctioneers' and bidders' strategies". *Information Systems Research* 14(3): 244-268.

Beese, J., M. K. Haki, S. Aier, and R. Winter. 2019. "Simulation-based research in information systems". *Business & Information Systems Engineering* 61(4):503-521.

Bitomsky, L., J. Huhn, W. Kratsch, and M. Roeglinger. 2019. "Process Meets Project Prioritization - A Decision Model For Developing Process Improvement Roadmaps". In *European Conference on Information Systems*, June 8th-14th, Stockholm, Sweden.

Bosse, S., C. Schulz, and K. Turowski. 2014. "Predicting availability and response times of IT services". In *Proceedings of the European Conference on Information Systems*, June 9th-11th, Tel Aviv, Israel, 1-14.

Brünnet, H., N. Lyubenova, M. Müller, J. E. Hoffmann, and D. Bähre. 2014. "Verification and application of a new 3D finite element approach to model the residual stress depth profile after autofrettage and consecutive reaming". *Procedia CIRP* 13:72-77.

Burton R. M., and B. Obel. 2011. "Computational modeling for what-is, what-might-be, and what-should-be studies—and triangulation". *Organization Science* 22(5):1195-1202.

Butler, B. S., P. J. Bateman, P. H. Gray, and E. I. Diamant. 2014. "An attraction–selection–attrition theory of online community size and resilience". *MIS Quarterly* 38(3):699-729.

Canessa, E., and R. L. Riolo. 2006. "An Agent–Based Model of the Impact of Computer–Mediated Communication on Organizational Culture and Performance: An example of the Application of Complex Systems Analysis Tools to the Study of CIS". *Journal of Information Technology* 21(4):272-283.

Chang, R. M., W. Oh, A. Pinsonneault, and D. Kwon. 2010. "A network perspective of digital competition in online advertising industries: A simulation-based approach". *Information Systems Research* 21(3):571-593.

Choi, J., D. L. Nazareth, and H. K. Jain. 2010. "Implementing service-oriented architecture in organizations". *Journal of Management Information Systems* 26(4):253-286.

Davis, J. P., K. M. Eisenhardt, and C. B. Bingham. 2007. "Developing theory through simulation methods". *Academy of Management Review* 32(2):480-499.

De Vreede, G. J. 1997. "Collaborative business engineering with animated electronic meetings". *Journal of Management Information Systems* 14(3):141-164.

Domschke, P., O. Kolb, and J. Lang. 2022. "Fast and reliable transient simulation and continuous optimization of large-scale gas networks". *Mathematical Methods of Operations Research*. Advance online publication.

Dutta, A. 2001."Business planning for network services: A systems thinking approach". *Information Systems Research* 12(3):260-283.

Dutta, A., A. Kankanhalli, and R. Roy. 2007. "The dynamics of sustainability of electronic knowledge repositories". In *International Conference on Information Systems*, December 9th-12th, Montreal, Canada.

Fuentes, A. T., M. Kipfmüller, C. Burghart, M. A. J. Prieto, T. Bertram, M. Bryg, and T. Bergmann. 2021. "Stable operation of arm type robots on mobile platforms". *Procedia CIRP* 99:104-109.

Gay, R. H., R. A. Davis, D. T. Phillips, and D. Z. Sui. 2005. "Modeling paradigm for the environmental impacts of the digital economy". *Journal of Organizational Computing and Electronic Commerce* 15(1):61-82.

Gerrits, B., M. Mes, and P. Schuur. 2017. "An agent-based simulation model for autonomous trailer docking". In *Proceedings of the 2017 Winter Simulation* Conference, edited by W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer and E. Page, 1324-1335. Piscataway, New Jersey: IEEE.

Glatt, M., D. Kull, B. Ravani, and J. C. Aurich. 2019. "Validation of a physics engine for the simulation of material flows in cyber-physical production systems". *Procedia CIRP* 81:494-499.

Guo, H., H. K. Cheng, and K. Kelley. 2016. "Impact of network structure on malware propagation: A growth curve perspective". *Journal of Management Information Systems* 33(1):296-325.

Harrison, J. R., Z. Lin, G. R. Carroll, and K. M. Carley. 2007. "Simulation Modeling in Organizational and Management Research". *Academy of Management Review* 32(4):1229-1245.

Hauser, F., J. Hautz, K. Hutter, and J. Füller. 2017. "Firestorms: Modeling conflict diffusion and management strategies in online communities". *Journal of Strategic Information Systems* 26(4):285-321.

Havakhor, T., A. A. Soror, and R. Sabherwal. 2018. "Diffusion of knowledge in social media networks: effects of reputation mechanisms and distribution of knowledge roles". *Information Systems Journal* 28(1):104-141.

Janssen, M., and A. Verbraeck. 2005. "Evaluating the information architecture of an electronic intermediary." *Journal of Organizational Computing and Electronic Commerce* 15(1):35-60.

Jiao, A., K. Egorova, J. Hahn, and G. Lee. 2016. "The effects of Spatial and Temporal dispersion on Virtual Teams' Performance". In *European Conference on Information Systems*, June 12th-15th, Istanbul, Turkey, 1-15.

Johnson, S. L., S. Faraj, and S. Kudaravalli. 2014. "Emergence of power laws in online communities." *MIS Quarterly* 38(3):795-808.

Kleijnen, J. P. 1995. "Verification and validation of simulation models". *European Journal of Operational Research* 82(1):145-162.

Koch, J.-A., J. Lausen, and M. Kohlhase. 2018. "Towards Internalizing the Externalities of Overfunding – Introducing a 'Tax' on Crowdfunding Platforms". In *European Conference on Information Systems*, June 23rd-28th, Portsmouth, UK, 1-15.

Konana, P., A. Gupta, and A. B. Whinston. 2000. "Integrating user preferences and real-time workload in information services". *Information Systems Research* 11(2):177-196.

Kontoyiannakis, K., E. Serrano, K. Tse, M. Lapp and A. Cohn. 2009. "A simulation framework to evaluate airport gate allocation policies under extreme delay conditions". In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R.G. Ingallis, 2332-2342. Piscataway, New Jersey: IEEE.

Kropp, L. A., J. J. Korbel, M. M. Theilig, and R. Zarnekow. 2019. "Dynamic Pricing of Product Clusters: A Multi-Agent Reinforcement Learning Approach". In *European Conference on Information Systems*, June 8th-14th, Stockholm, Sweden.

Kumar, R. L., S. Park, and C. Subramaniam, 2008. "Understanding the value of countermeasure portfolios in information systems security". *Journal of Management Information Systems* 25(2):241-280.

Küppers, G., and J. Lenhard. 2005. "Validation of simulation: Patterns in the social and natural sciences". *Journal of Artificial Societies and Social Simulation* 8(4):1-13.

Lindberg, T., F. Johansson, A. Peterson, and A. Tapani. 2020. "Microsimulation of bus terminals: a case study from Stockholm". In *Proceedings of the 2020 Winter Simulation Conference*, edited by K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, 1206-1217. Piscataway, New Jersey: IEEE.

Lombard, M., J. Snyder-Duch, and C. C. Bracken. 2002. "Content analysis in mass communication. Assesment and reporting of intercoder reliability". *Human Communication Research* 28(4):587-604.

Mes, M. R., and M. Koot. 2019. "Simulation solution validation for an integrated emergency post". In *Proceedings of the 2019 Winter Simulation Conference*, edited by N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, 1160-1171. Piscataway, New Jersey: IEEE.

Mingers, J., and C. Standing. 2020. "A framework for validating information systems research based on a pluralist account of truth and correctness". *Journal of the Association for Information Systems* 21(1):117-151.

Nan, N., and E. W. Johnston. 2009. "Using multi-agent simulation to explore the contribution of facilitation to GSS transition". *Journal of the Association for Information Systems* 10(3):252-277.

Oh, W., J. Y. Moon, J. Hahn, and T. Kim. 2016. "Research note—Leader influence on sustained participation in online collaborative work communities: A simulation-based approach". *Information Systems Research* 27(2):383-402.

Oreskes, N. 1998. "Evaluation (not validation) of quantitative models". *Environmental Health Perspectives* 106(6):1453-1460.

Pfeiffer, A., D. Gyulai, B. Kádár, and L. Monostori. 2016. "Manufacturing lead time estimation with the combination of simulation and statistical learning methods". *Procedia CIRP* 41:75-80.

Pierce, M. E., U. Krumme, and A. M. Uhrmacher. 2018. "Building simulation models of complex ecological systems by successive composition and reusing simulation experiments". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 2363-2374. Piscataway, New Jersey: IEEE.

Port, D., and T. Bui. 2009. "Simulating mixed agile and plan-based requirements prioritization strategies: proof-of-concept and practical implications". *European Journal of Information Systems* 18(4):317-331.

Ren, Y., and R. E. Kraut. 2011. "A simulation for designing online community: Member motivation, contribution, and discussion moderation". *Information Systems Research*.

Ridler, S., A. J. Mason, and A. Raith. 2022. "A simulation and optimisation package for emergency medical services". *European Journal of Operational Research* 298(3):1101-1113.

Ross, B., L. Pilz, B. Cabrera, F. Brachten, G. Neubaum, and S. Stieglitz. 2019. "Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks". *European Journal of Information Systems* 28(4):394-412.

Sargent, R. G. 2013. "Verification & validation of simulation models". *Journal of Simulation* 7(1):12-24.

Sargent, R. G. 2020. "Verification and validation of simulation models: An advanced tutorial." In *Proceedings of the 2020 Winter Simulation Conference*, edited by K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, 16-29. Piscataway, New Jersey: IEEE.

Schroer, K., W. Ketter, T. Y. Lee, A. Gupta, and M. Kahlen. 2022. "Data-Driven Competitor-Aware Positioning in On-Demand Vehicle Rental Networks." *Transportation Science* 56(1):182-200.

Semelhago, M., B. L. Nelson, E. Song, and A. Wächter. 2021. "Rapid discrete optimization via simulation with Gaussian Markov random fields". *INFORMS Journal on Computing* 33(3): 915-930.

Sen, S., T. S. Raghu, and A. Vinze. 2009. "Demand heterogeneity in IT infrastructure services: Modeling and evaluation of a dynamic approach to defining service levels". *Information Systems Research* 20(2):258-276.

Shah, V., L. Gulikers, L. Massoulié, and M. Vojnović. 2020. "Adaptive matching for expert systems with uncertain task types". *Operations Research* 68(5):1403-1424.

Smits, M., W. J. van den Heuvel, and C. Nikolaou. 2011. "Redesign and performance of service networks: A systems dynamics approach". In *European Conference on Information Systems*, June 9th-11th, Helsinki, Finland, 1-12.

Stockdale, R., and C. Standing. 2006. "An interpretive approach to evaluating information systems: A content, context, process framework". *European Journal of Operational Research* 173(3):1090-1102.

Tolk, A., J. E. Lane, F. L. Shults, and W. J. Wildman. 2021. "Panel on ethical constraints on validation, verification, and application of simulation". In *Proceedings of the 2021 Winter Simulation Conference*, edited by Panel on ethical constraints on validation, verification, and application of simulation, 1-15. Piscataway, New Jersey: IEEE.

Torres-Jiménez, M., C. R. García-Alonso, L. Salvador-Carulla, and V. Fernández-Rodríguez. 2015. "Evaluation of system efficiency using the Monte Carlo DEA: The case of small health areas". *European Journal of Operational Research* 242(2):525-535.

Troy, P., L. Westaway, A. Grondin, and T. Rezanowicz. 2017. "Rationalizing healthcare budgeting when providing services with mandated maximum delays: A simulation modeling approach". In *Proceedings of 2017 Winter Simulation Conference*, edited by W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer and E. Page, 2740-251. Piscataway, New Jersey: IEEE.

Utomo, D. S., A. Gripton, and P. Greening. 2020. "Long haul logistics using electric trailers by incorporating an energy consumption meta-model into agent-based model". In *Proceedings of the 2020 Winter Simulation Conference*, edited by K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, 147-158. Piscataway, New Jersey: IEEE.

Van Horn, R. L. 1971. "Validation of simulation results". *Management Science* 17(5):247-258.

Wawrzyniak, J., M. Drozdowski, and É. Sanlaville. 2020. "Selecting algorithms for large berth allocation problems". *European Journal of Operational Research* 283(3):844-862.

Webster, J., and R. T. Watson. 2002. "Analyzing the Past to Prepare for the Future: Writing a Literature Review". *MIS Quarterly* 26(2):xiii-xxiii.

Winsberg, E. 2003. "Simulated experiments: Methodology for a virtual world". *Philosophy of Science* 70(1):105-125.

Wöste, F., T. Siebrecht, and P. Wiederkehr. 2021. "Evaluation of a novel approach for considering damping effects in a process force model of a geometric physically-based milling simulation". *Procedia CIRP* 103:188-193.

AUTHOR BIOGRAPHIES

MARVIN AUF DER LANDWEHR is full-time Ph.D. student and researcher at the University of Applied Sciences and Arts of Hannover. He holds an honored M.Sc. in strategic corporate development. His research interests focus on Logistics and Supply Chain Simulation, Modelling Paradigms and Digital Nudging. His e-mail address is marvin.auf-der-landwehr@hs-hannover.de.

MAIK TROTT is part-time Ph.D. student at the University of Applied Sciences and Arts of Hannover and Innovation Manager at DB Schenker Deutschland AG. He holds a B.Sc. in Business Informatics and an honored M.Sc. in strategic corporate development. His interests focus on Simulation, Logistics and Enterprise Innovations. His email address is maik.trott@hs-hannover.de.

MAYLIN WARTENBERG is professor for business informatics at the University of Applied Sciences Hannover. She has a mathematical background and profound experience in the financial and automotive industry. Her research areas include Data Science, Business Intelligence and Artificial Intelligence. She is also active in Science Communication in the field of AI. Her email address is maylin.wartenberg@hs-hannover.de. Her website is http://www.das-hub.de.

CHRISTOPH VON VIEBAHN is professor for business informatics at the University of Applied Sciences Hannover. He is specialized in supply chain management, business administration and dynamic computer simulation. Since 2016, he is elected chairperson for the chapter Lower Saxony by the Bundesvereinigung Logistik (BVL) and is member of the urban logistics working group since 2017. His email address is christoph-von.viebahn@hs-hannover.de. His website is http://www.das-hub.de.