An Approach for Face Validity Assessment of Agent-Based Simulation Models Through Outlier Detection with Process Mining

Rob Bemthuis[1,2(✉)] and Sanja Lazarova-Molnar[2,3]

[1] University of Twente, Enschede, The Netherlands
r.h.bemthuis@utwente.nl
[2] Karlsruhe Institute of Technology, Karlsruhe, Germany
rob.bemthuis@partner.kit.edu, sanja.lazarova-molnar@kit.edu
[3] University of Southern Denmark, Odense, Denmark
slmo@mmmi.sdu.dk

Abstract. When designing simulations, the objective is to create a representation of a real-world system or process to understand, analyze, predict, or improve its behavior. Typically, the first step in assessing the credibility of a simulation model for its intended purpose involves conducting a face validity check. This entails a subjective assessment by individuals knowledgeable about the system to determine if the model appears plausible. The emerging field of process mining can aid in the face validity assessment process by extracting process models and insights from event logs generated by the system being simulated. Process mining techniques, combined with the visual representation of discovered process models, offer a novel approach for experts to evaluate the validity and behavior of simulation models. In this context, outliers can play a key role in evaluating the face validity of simulation models by drawing attention to unusual behaviors that can either raise doubts about or reinforce the model's credibility in capturing the full range of behaviors present in the real world. Outliers can provide valuable information that can help identify concerns, prompt improvements, and ultimately enhance the validity of the simulation model. In this paper, we propose an approach that uses process mining techniques to detect outlier behaviors in agent-based simulation models with the aim of utilizing this information for evaluating face validity of simulation models. We illustrate our approach using the Schelling segregation model.

Keywords: Face validity · Agent-based simulation · Process mining

1 Introduction

Face validity is a key aspect of simulation model design. It involves asking knowledgeable individuals about the system if the model and its behavior are plausible

H. A. Proper et al. (Eds.): EDOC 2023, LNCS 14367, pp. 134–151, 2024.
https://doi.org/10.1007/978-3-031-46587-1_8

[24]. This step is commonly used to evaluate how well the model captures the essential features of the real-world system that it represents [16,25]. Face validity methods rely on human expertise and include expert assessments and structured walk-throughs [15]. This process ensures that the simulation model outcomes are reasonable and plausible within the theoretical framework and implicit knowledge of system experts or stakeholders. Although other types of validation, such as statistical validation, are also important, face validity is relevant because it is a common first step in assessing the simulation model's validity. Nevertheless, after designing their model, practitioners and researchers should rigorously test its performance under a variety of conditions [9].

Assessing face validity in simulation modeling can be challenging [20]. One major challenge is the potential for a discrepancy between the model's assumptions and the real-world system it represents, which can affect the accuracy of the model [16]. The complexity of the modeled system can also make achieving face validity challenging, as can balancing model accuracy with simplicity [9]. Additionally, the subjective nature of face validity and the potential for biases in the validation process must be taken into account [28]. Incorporating feedback from subject matter experts and stakeholders can enhance face validity, but this process can be time-consuming and resource-intensive. For instance, if assessors find it difficult to categorize and examine every available option, achieving comprehensive face validity can become challenging [10].

To this end, the emerging field of process mining can provide a valuable tool for enhancing face validity in simulation models and addressing some of the concerns previously mentioned. Process mining extracts knowledge from event logs of real-life processes [1] and can validate behavior in simulation models against real-world behavior. By comparing simulation output with data extracted from real-world processes, process mining can help identify discrepancies. This allows for adjustments to be made to the simulation model to improve its validity, as discussed in Subsect. 2.2. The wealth of techniques developed within the process mining discipline can also be applied to event logs generated by a simulation model. This enables experts in simulation and process mining to assess resulting process models and corresponding performance insights for face validity. By utilizing process mining, including its visually appealing discovered process models, simulation modelers can provide a novel way to ensure that simulated processes and outcomes are consistent with real-world systems. This can contribute positively to performing face validity assessments.

In simulation models, outliers can significantly impact the model's validity. Outliers, which lie an abnormal distance from other values (in an arbitrary sample), can draw attention to unusual behavior, either raising doubts about or reinforcing the model's credibility. Outliers also play a key role in conducting face validity checks. By identifying and addressing outliers, the accuracy and reliability of a simulation model can be enhanced, leading to more valid conclusions and improved decision-making. Outliers can reveal important and unexpected behaviors that capture a wide range of real-world phenomena. By verifying that certain cases are indeed outliers, the model's face validity can

be strengthened. For instance, in a disease spread simulation model, the identification of an outlier case as a 'super-spreader' can bolster the model's face validity by accurately representing the significant impact of super-spreaders on disease spread in real-world scenarios. Conversely, outliers may also compromise the validity of a simulation model if their existence is doubted or considered unrealistic by experts in the field.

In this paper, we propose an approach that employs process mining techniques to detect outlier behaviors in agent-based simulation (ABS) models, with the goal of enhancing the face validity assessment process. While face validity is generally essential for many types of simulation, it is particularly important for ABS [15]. The distinctive attributes of ABS models, including the representation of heterogeneous agents and the emphasis on emergent behavior, underscore the criticality of emphasizing face validity in this domain. Our study makes two main contributions: (1) we apply process mining techniques to extract and identify outlier behaviors from an ABS model, and (2) we incorporate human expertise to conduct a face validity assessment on the knowledge extracted through process mining, thereby reinforcing human judgement in the evaluation process. We demonstrate the versatility and potential usefulness of our approach using the Schelling model of segregation, a popular ABS model, and show how it can be applied to various scenarios of the ABS model. Our approach is guided by Peffers' Design Science Research Methodology [21], as reflected in the structure of this article.

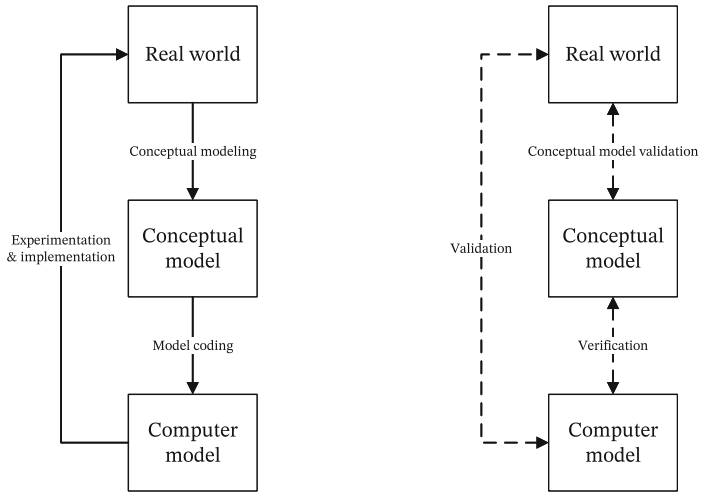
The remainder of this paper is organized as follows. Section 2 describes background on face validity for ABS models and how process mining can enhance the face validity of ABS modeling practices. Section 3 presents our approach. Section 4 demonstrates our approach through an illustrative scenario. Finally, Sect. 5 positions related work and Sect. 6 concludes and provides pointers for further work.

2 Background

In this section, we begin by providing a concise overview of face validity techniques for ABS models. Subsequently, we explore the potential benefits of incorporating process mining techniques into the face validity process of ABS models.

2.1 Face Validity Techniques for Agent-Based Simulation Models

Model validation is the process of assessing the accuracy of a simulation model in representing a real-world system with respect to the study objectives [3] (see also Fig. 1). This involves comparing the model's output to empirical data, experimental results, or expert knowledge to ensure that it realistically represents the system being modeled. Face validity is one specific validation technique, which is considered to be relatively informal and subjective [24]. It involves soliciting feedback from individuals who are knowledgeable about the system being



(a) General modeling process of a simulation model.

(b) Model verification and validation.

Fig. 1. Verification and validation of a simulation model (adapted from [22]).

modeled to determine whether the model and its behavior are reasonable and realistic.

Several face validity techniques have been proposed in the literature, and these techniques are not necessarily mutually exclusive. Table 1 provides an overview of these techniques. Ideally, these techniques should be conducted by independent groups of human experts [15]. While there are other face validity techniques in the field of simulation modeling, our study focuses specifically on techniques related to ABS models. Focusing on ABS models narrows the scope and limitations of our research to a more specific domain.

The methodology for conducting a rigorous face validity assessment of a simulation model is subject to debate and may depend on various factors. However, several general guidelines can be proposed. Firstly, evaluators must possess *sufficient knowledge and expertise* about the system being modeled. Secondly, the model must be *transparent and comprehensible* to the evaluators. It should have explicit explanations of its assumptions, inputs, and outputs. Thirdly, evaluators should be presented with *realistic scenarios* that accurately reflect the system's expected behavior. Finally, face validity assessment should be considered an *iterative process*, with modifications made to the model based on feedback received from the evaluators.

2.2 Achieving Face Validity Through Process Mining Techniques

Process mining provides a data-driven approach to supplementing the face validity process of simulation models. It facilitates a comparison between real-world processes and the behavior of the simulation model (see Fig. 2). By extracting

Table 1. Face validity techniques for ABS models reported in literature.

Name	Reference(s)	Description
Animation	[15,28]	Graphical display of model behavior over time
Output assessment	[15]	Check the plausibility of absolute values, relations, dynamics, and trends of output values in simulation runs
Immersive assessment	[15]	Evaluate the behavior of an isolated simulated agent by observing its perceptions and reactions through its interface. The expert can also assess the behavior of other agents by interacting with the human-controlled agent if the interface allows participation
Turing test	[9,28]	Test if experts distinguish between model-generated and real-world data
Graphical representation	[28]	Visualize the model's output data with graphs
Tracing	[28]	Isolated entities in the model are monitored for their behavior
Internal validity	[28]	Compare the results of multiple replications of a stochastic simulation model using different random seeds. Inconsistent sample points resulting from the random number generators indicate issues with either the programming model or the conceptual model
Historical data validation	[28]	When historical data is available, the model is built using a portion of the data (training sets), and the remaining data (test sets) is used to verify whether the model emulates the behavior of the system
Sensitivity analysis	[28]	Model input and parameters are adjusted to examine their impact on the model's behavior and input, with the expectation that the model will reflect the real system. Sensitive parameters, which significantly affect the model's behavior, must be accurately determined before the model can be utilized
Predictive validation	[28]	Compare the model's predictions to actual system behavior, which can be obtained from operational systems or experiments, including laboratory or field tests

information about the actual execution of processes from event logs generated by both real-life processes and simulation models, process mining techniques can discover process models that graphically represent the flow of activities in a chart (e.g., through nodes, activities, and gateways). The discovered process models and performance insights, such as throughput and waiting times, can provide valuable insights into the execution of activities. For instance, comparing the extracted process model with the expected real-life behavior as determined by experts can enhance the validity of a simulation model by identifying discrepancies. Process mining can provide a systematic approach for analyzing data and identifying outliers or deviations from expected patterns.

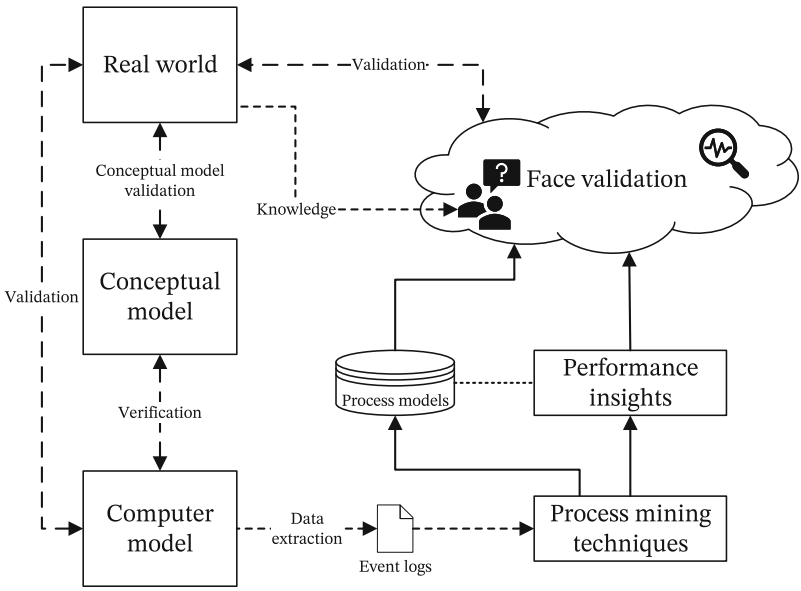


Fig. 2. Visual representation of using process mining techniques to assess the face validity of a simulation model.

To our knowledge, there is limited research on using process mining to enhance the face validity of ABS models. Although some studies may implicitly employ process mining methods, few explicitly discuss their use in the context of face validity or related terms (see Sect. 5) such as plausibility checks. Nonetheless, we believe that process mining techniques can serve as a valuable tool for assessing the face validity of simulation models, as per the guidelines outlined in the preceding subsection. Firstly, assessors of agent-based systems are expected to possess domain knowledge about the system being modeled and its Key Performance Indicators (KPIs). Process mining techniques can provide insights into a wide range of performance indicators that align with the modeled system's KPIs. Secondly, the results obtained through process mining are

based on the analysis of event logs, allowing users to trace the flow of events and understand how the results were obtained. This is important for accurately representing emergent behavior, interaction dynamics, and outlier behaviors in ABS models. Furthermore, visually engaging mined process models can enhance usability and comprehensibility among individuals who lack specialized expertise in the field of process mining but are familiar with the agent system being modeled [6]. Thirdly, evaluators can be presented with specific scenarios of preference or interest (e.g., outlier behaviors) based on event records that describe the actual functioning of the system due to the granularity and sophistication of process mining techniques and many available tools (e.g., for filtering traces) [23]. Finally, process mining can be used iteratively to enhance face validity by regularly comparing the simulation model's behavior with real-world processes. For example, real-time streaming of event data, combined with validity checks at set intervals, allows for monitoring of the model's performance and identification of exceptional behaviors. Injecting new event logs provides additional data points for testing and refining model performance, thereby enhancing accuracy.

3 An Approach for Assessing Face Validity

In this section, we present an approach for assessing face validity of ABS models. We first provide an overview of the approach and then discuss each step of the approach in detail. Our approach is illustrated in Fig. 3 and comprises six steps that leverage execution logs extracted from an ABS model.

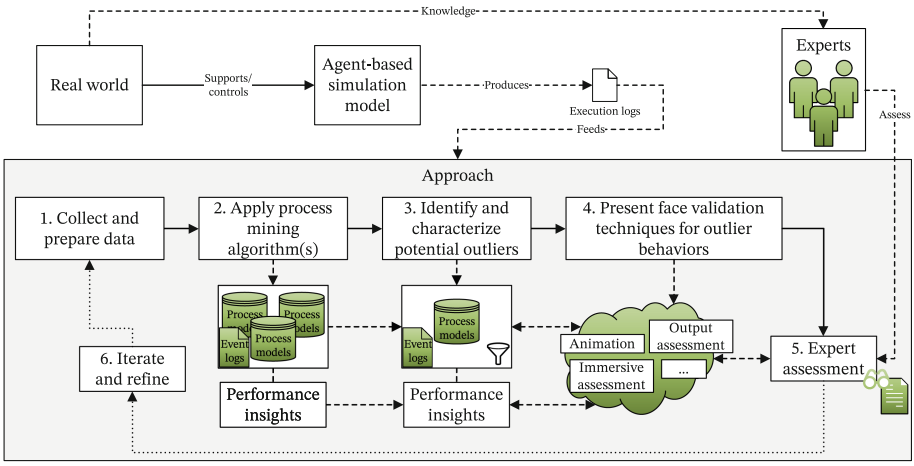


Fig. 3. A face validity assessment approach of ABS models using process mining.

Step 1: Collect and Prepare Data. The initial step involves selecting, cleaning, and aggregating (event) log files from the real-world system being modeled. These files are then prepared for analysis using process mining techniques, which entails locating the raw data, migrating and transforming it into an event log file format, and pre-processing the data (e.g., filtering and cleaning).

Step 2: Apply Process Mining Algorithm(s). The next step involves using process mining techniques to extract information about the actual execution of processes from the event logs. In this step, we select appropriate process mining discovery techniques to mine process models and assess their quality. This assessment is important as multiple process models may be generated, and their representation quality (e.g., fitness and precision) and other quality indicators (e.g., number of event logs) must be considered.

Step 3: Identify and Characterize Potential Outliers. The third step entails ensuring that the obtained process model and insights can be used to identify potential outliers. This can be achieved by displaying different abstractions of the process model through filtering to examine specific aspects of the process. Additionally, the process model can be augmented with various process metrics (i.e., KPIs), such as frequency metrics (e.g., absolute and case frequency, maximum number of repetitions), performance metrics (e.g., total duration, minimum/median/maximum duration), or combined metrics. The identified outliers can then be characterized to understand their deviation from the norm. This involves analyzing their attributes and features to determine their uniqueness or difference, which may involve examining the attributes of the process, the context of the outlier, and potential contributing factors.

Step 4: Present Face Validation Techniques for Outlier Behaviors. The fourth step involves selecting, configuring, and presenting appropriate face validation techniques for use by human assessors, such as visual inspection of the model's behavior and solicitation of feedback from domain experts. Table 1 presents additional examples of face validity techniques. The specific techniques chosen depend on the nature of the ABS and process insights, as well as the availability of relevant data and expertise. Individuals familiar with the system being studied should be instructed on the use of the face validity procedures.

Step 5: Expert Assessment. The fifth step involves having experts assess whether the behaviors presented through the use of process mining techniques are plausible. These experts can identify any discrepancies between the simulation model and the real-world system, or confirm the accuracy of the model's representation of system behaviors.

Step 6: Iterate and Refine. After conducting a face validity check and establishing sufficient credibility in the ABS model, it is common to proceed with

other validation assessments, such as quantitative statistical analysis. However, adjustments may be made to the simulation model to more accurately represent reality and followed by additional face validity assessments. This iterative process involves assessing the simulation model's behavior against real-world processes and making modifications based on feedback from the face validity assessment. This can gradually improve the accuracy and reliability of the ABS model, leading to more valid conclusions and better decision-making. The iteration process can also involve refining previous steps, such as addressing new potential outlier behaviors based on feedback from human assessors and involving different or new experts in conducting face validity assessments.

4 Illustrative Scenario

In this section, we present an illustrative scenario demonstrating the application of our approach using the Schelling model of segregation. We introduce the scenario and detail the key actions taken to apply our approach. Selected outcomes are presented for clarity and conciseness, with a focus on the illustrative nature of the scenario. We conclude by summarizing the lessons learned and the challenges encountered.

4.1 The Schelling Model of Segregation

The Schelling model of segregation is a social simulation model that demonstrates how individual preferences can lead to large-scale social patterns, even with low levels of discrimination or prejudice [26]. The model has been applied in various research fields, has inspired policy-makers and planners to develop strategies for promoting diversity and reducing segregation in urban areas, and has also served as a basis for developing other simulation models exploring social phenomena.

In the classic Schelling segregation model, a grid representing a housing market is filled with individuals who possess a "tolerance threshold" for the percentage of their neighbors that must be of the same race or ethnicity [26]. As the simulation progresses, individuals who are not satisfied with their neighbors relocate to new positions on the grid that meet their tolerance threshold. This can result in the formation of highly segregated neighborhoods as individuals with similar traits congregate. This congregation can occur even when individual preferences are not extreme but rather moderate.

Conducting a face validation assessment through process mining for the Schelling model of segregation is relevant for several reasons. First, process mining techniques can help identify and characterize outliers in the simulation model that may undermine or fortify its validity. By addressing these outliers, the accuracy and reliability of the model can be improved, which can lead to more valid conclusions and better decision-making. Second, the Schelling model of segregation is a widely recognized and influential social simulation model that has

been applied in various fields of research. Ensuring its face validity is important for maintaining its credibility and usefulness as a tool for understanding complex social phenomena and for the wider (agent-based) simulation modeling community.

4.2 Demonstration of the Approach

Step 1: Collect and Prepare Data. For our ABS model, we used the Schelling model implementation described by [7], which we modified using Python 3.6.9 and the AgentPy 0.1.5 library [11]. We limited extraction to the event logs of the scenarios presented in Table 2. These scenarios cover a variety of situations and differentiate among several model parameters. "Ruleset type" refers to the model's configuration where either all agent groups have the same tolerance threshold (homogeneous population) or all but one agent group have the same tolerance threshold (heterogeneous population).

Table 2. Scenarios of the Schelling model considered for event log extraction.

Scenario	Density	Grid size	Ruleset type	Tolerance threshold (%)
1	0.80	20×20	homogeneous	0.55
2	0.70	20×20	homogeneous	0.55
3	0.70	20×20	homogeneous	0.20
4	0.70	20×20	heterogeneous	0.10

Table 3 provides an example of a produced event log. The activities included three types: *moveLocation* (i.e., an agent moves from one location to another), *changeHappy* (an agent's status changes from happy or unhappy to happy), and *changeUnhappy* (an agent's status changes from happy or unhappy to unhappy). For the timestamp, we adopted a similar approach to that described by [7]. We assigned a sequential counter to each step in the model's execution based on the order in which it occurred chronologically.

Table 3. An excerpt of an event log generated.

timestamp	counter	activity	caseID	coordinates	directNeighbors	happinessLevel
2022-01-01 12:31:05	0	changeHappy	23	18,5	2, 14, 22, 87	0.75
2022-01-06 09:00:00	0	moveLocation	55	6,3	13, 16, 56, 61, 81	0.84
2022-01-06 09:00:00	1	changeHappy	13	5,2	8,16,41,55,56,77,83	0.90
2022-01-06 09:00:00	2	changeUnhappy	16	6,2	3,13,55,56,77,81	0.45
2022-01-06 09:00:00	3	changeUnhappy	56	5,3	8,13,16,41,55,61	0.29
2022-01-06 09:00:00	4	changeHappy	61	6,4	6,31,55,56,73,81	0.66
2022-01-06 09:00:00	5	changeHappy	81	7,3	2,16,55,61	0.78
2022-02-01 10:23:00	0	moveLocation	33	11,16	5,25	0.81

For analysis purposes, we concatenated the naming convention of an activity to include both the number of neighbor agents and the number of neighboring agents of a similar group (i.e., *changeHappy_X_Y*, where X = number of neighbors and Y = number of neighbors of a similar group). This naming convention ensured that there was sufficient data to obtain a realistic overview of multiple scenarios while keeping the entire process manageable in size.

Step 2: Apply Process Mining Algorithm(s). When conducting a face validity assessment based on event logs produced by an agent-based system, the choice of process mining discovery algorithm depends on the specific characteristics of the data and the desired outcome. For instance, if the data contains a significant amount of noise or if the process being modeled is less structured, then the Fuzzy Miner might be a more suitable choice [13]. However, if the data is relatively clean and well-structured, then the Heuristic Miner might be more appropriate [12].

We chose the Fuzzy Miner for its efficiency and ease of use with less structured processes. By using significance/correlation metrics to simplify the process model, it provides reliable results for complex data sets [13]. It can also exclude or cluster less important activities. The Fuzzy Miner can animate the event log on top of the created model, providing an understanding of dynamic process behavior, which is desirable for assessing face validity. We selected Disco as our process mining tool due to its usability, fidelity, and performance [14].

Step 3: Identify and Characterize Potential Outliers. We employed process mining techniques to filter the representations and present outlier behaviors. As an illustration, we demonstrate how animation, output, and immersive assessments (as described by [15]) can be conducted. For the identification and characterization of outliers, we analyzed the attributes and features of (potential) outlier behaviors. Further details and visual representations are provided in the following step.

Step 4: Present Face Validation Techniques for Outlier Behaviors. We demonstrate the application of three face validity techniques: animation, output, and immersive assessment. Due to space constraints, we present only a selection of these outcomes. In the following, we provide examples of the information that could be presented to an individual tasked with assessing validity. The next step presents an example of an expert assessment.

Animation Assessment: We created a graphical representation that displays the progress of the graphical model behavior over time in distinct time intervals. All paths and activities were shown, allowing the assessor to visually observe exceptions and identify possible deviations. Figure 4 shows a snapshot of the animation shown for scenario 2. The evaluator can examine various behaviors from the animation and focus on outliers based on attributes such as color and

arrow thickness. This animation is particularly useful for intuitively identifying and highlighting bottlenecks in the process. When numerous cases accumulate on a specific arc, causing congestion, the tool aggregates these cases into larger "bubbles", emphasizing bottlenecks in the process.

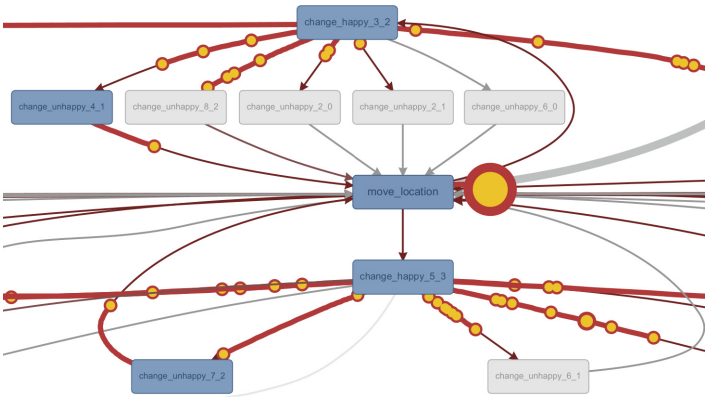


Fig. 4. Snapshot of the graphical animation used for scenario 2.

Output Assessment: We generated a graphical representation of the case duration (i.e., steps conducted by an agent), supplemented by additional statistics such as the number of events, cases, and activities for agents that exhibited movement at least once. Figure 5 illustrates an example of the visual representation employed for scenario 2. This visual representation enables experts to conduct an initial assessment of the plausibility of the ABS model. Using this example, the expert can evaluate extreme cases, such as case durations with the highest number of cases, or identify trends.

Fig. 5. Example representation used for output assessment of scenario 3.

Immersive Assessment: We identified agents that deviated from the average behavior by analyzing cases with an exceptionally high number of associated events, indicating unusual frequency in their movement patterns. A detailed overview of each agent's activities, including timestamps and happiness status,

was provided to the human assessor for evaluation. Since activity names encode information about both the number of neighbors and the number of neighbors belonging to a similar group, this information could provide valuable insights into the validity of the ABS model. Tables 4 and 5 present examples of an agent's trace.

Table 4. An excerpt of traces used for immersive assessment for scenario 1, case 206 with 112 events.

Activity	Time	happinessLevel
...	...	...
change_happy_4_3	01:01:17	false
change_unhappy_7_3	01:01:19	true
move_location	01:01:20	false
move_location	01:01:21	false
change_happy_5_3	01:01:21	false
change_unhappy_2_0	01:01:22	true
...	...	...

Table 5. An excerpt of traces used for immersive assessment for scenario 4, case 248 with 6 events.

Activity	Time	happinessLevel
move_location	01:00:01	false
move_location	01:00:02	false
move_location	01:00:03	false
move_location	01:00:04	false
move_location	01:00:05	false
change_happy_8_3	01:00:05	false

Step 5: Expert Assessment. We provided an expert with a template that included an introduction to the ABS model, its purpose, the assessment method, and instructions for evaluation. The evaluation criteria were listed, and space was provided for the expert to record their observations and provide feedback on the face validity of the ABS model's outcome. The evaluation criteria included: (1) accurate representation of the real-world system being modeled; (2) consistency of agent behavior with their real-world counterparts; and (3) plausibility of the overall simulation model outcome. A human assessor evaluated the face validity based on the artifacts discussed in step 4. The expert evaluated the face validity and scored it as either 'plausible' or 'not plausible'. The assessment results and conclusions regarding the face validity of the ABS model's outcome were summarized, and selected outcomes are reported below.

During both the animation and immersive assessments, an observation was made regarding the perceived unrealistic movement time of agents from one location to another. While this can be attributed to model limitations as described by [7], it is not considered plausible in reality. Another concern, as discovered during the animation assessment, was the simultaneous movement of multiple agents at similar time steps. While this may be attributed to exceptional real-life behaviors or administrative practices in the housing market (e.g., movements designated as "official" at the start of a new day), its repeated occurrence during the animation led the human assessor to deem this model behavior not plausible in reality.

During the output assessment of scenario 3 (Fig. 5), the human assessor observed two distinct peaks. The first peak occurred at the beginning when

16 cases were satisfied after a single move, possibly due to dissatisfaction with a single neighbor. The second peak involved 17 cases that continued to move after 100 time units, representing approximately 6% of the total population. This observation raised concerns about the validity of the ABS model and warrants further investigation, as it was perceived as not plausible in reality.

In addition to the previously mentioned unrealistic timing of events, the immersive assessment yielded mixed results regarding the model's validity as determined by the assessor. For example, the event log trace of case 206 in scenario 1 (as depicted in Table 4) indicates a not plausible number of agent movements, while the observed pattern of transitioning between happy and unhappy states was deemed consistent with real-world practices.

Step 6: Iterate and Refine. Currently, this step is still in progress. Further examination is needed to thoroughly explore the discrepancies. Discussions are also underway about implementing our approach on a real-world dataset and involving policy-makers, real estate agencies, homeowners or renters in the panel of experts to validate the Schelling model.

4.3 Discussion

In this subsection, we briefly discuss the lessons learned and open challenges encountered during our case study. Our findings suggest that the outcomes obtained through process mining techniques were intuitive and easy to follow, making the assessment process relatively efficient. However, we also identified a need for more structured approaches to conducting face validity assessments of ABS models using process mining techniques.

One challenge we encountered was matching the KPIs used in practice with those obtainable through process mining techniques. Further research is needed to address this issue and ensure that KPIs used in face validity assessments are appropriate and well-known to experts. Another open research question concerns addressing the adaptive or changing behavior of agents in ABS models, as also highlighted by [5,6]. Concept drift incorporation into face validity assessments is also important. Further research is needed to develop methods for incorporating dynamic aspects of ABS models, such as their evolving nature, heterogeneity, and complex interactions, into face validity assessments.

We evaluated our proposed method using a segregation model across four scenarios. Although a comprehensive presentation of results for each scenario based on established criteria could provide valuable insights, we have chosen to emphasize only specific findings in this paper due to resource and time constraints as well as limited publication space. It is important to note that our evaluation was limited by our reliance on a single expert opinion for behavior assessment and the application example used in our study may not be representative enough to make more fundamental statements.

5 Related Work

Previous research highlights the importance of face validity in enhancing ABS model credibility. Several methods, including expert panels and stakeholder engagement (see e.g., Table 1), have been proposed for assessing face validity. The benefits of incorporating face validity assessments in ABS model design and implementation for decision-making and policy development are also emphasized.

For example, [27] developed a computational model of collaborative learning health systems using an agent-based approach and demonstrated its initial computational and face validity. The authors demonstrated face validity by examining the effects of varying a single parameter. However, they acknowledge the model's face validity for a wide range of stakeholders is unknown and call for further refinement through collaboration with experts. In other work, [2] proposed a framework for evaluating health care markets through agent-based modeling and presented a face-validity assessment procedure by examining the degree to which empirical studies support key theorized relationships within health care markets and comparing them with relationships generated by their model. Furthermore, [18] presented an agent-based model of a stock market that incorporates common-sense evidence and implements realistic trading strategies based on practitioners' literature. The model was validated using a four-step approach consisting of face validity assessment, sensitivity analysis, calibration, and validation of model outputs.

In the process mining domain, [17] used semi-structured interviews to evaluate the face validity of process mining results, but the specific questions used during the interviews were not reported. In the ABS domain, [8] used numerical simulations to test the face validity of a part of their ABS model. While the authors' findings support the plausibility of the outcomes within the theoretical framework, they advocate for further empirical estimation of model parameters through real-world measurements. Work described in [19] used feedback from a project manager to assess the face validity of their model. In [4], the authors conducted an exploratory study on face validity assessment in ABS models, presenting a proof-of-concept that combines process mining with visualization. Their results provide initial evidence of the effectiveness of this approach, but also highlight the need for further research to gain a finer-grained understanding of agent-level dynamics, such as studying emergent behaviors at specific group levels, including outlier behaviors.

Our work builds upon previous research by applying existing process mining techniques to identify outliers in an ABS model and having a human expert assess these outliers. Previous approaches focused on ensuring that KPIs, such as average cycle time or work-in-progress levels, aligned with observed KPIs, but neglected to investigate outlier behavior. By including this behavior, we provide meaningful insights into the validity of an ABS model. Furthermore, by incorporating process mining techniques into the design and execution of ABS models, we enable face validity assessments even for non-experts. Our proposed approach, combined with the application of the Schelling model of segregation,

represents an initial step towards analyzing complex socio-technical systems typically modeled and simulated using agent-based techniques.

6 Conclusion and Future Work

In this paper, we presented an approach using process mining techniques to detect outliers in ABS models and evaluate their face validity. Our approach leverages human expertise to perform a face validity assessment on the information obtained from the process mining analysis, focusing on outlier behaviors. We demonstrated our approach using the Schelling model of segregation and showed how it can be used to assess face validity. In particular, we demonstrated how animation, output assessment, and immersive assessment can be employed as face validity techniques for ABS models. This study offers valuable insights for enhancing the face validity assessment process of ABS models using process mining and holds broader potential for advancing the field of simulation modeling.

Our study findings are subject to validity threats due to the complexity of ABS models and the need for specialized face validation techniques to accurately capture agent behavior and emergence. We focused on a small set of techniques and demonstrated our approach in a limited experimental environment, limiting the generalizability of our results. Face validity is subjective and reliant on the validator's judgment, making it challenging to standardize or quantify results and potentially leading to inconsistencies in the validation process. Additionally, our inspection of a restricted subset of simulation outputs may fail to capture essential aspects of the simulation's behavior, potentially resulting in an incomplete evaluation of the model's validity. Nevertheless, ABS models have unique characteristics such as complex agent interactions, stochasticity, and emergent behaviors, making them challenging to comprehensively assess by experts.

Future work will involve refining and extending our approach by adapting ABS models to various domains and settings, investigating a broader range of process mining techniques, and including insights from existing literature on face validity within the broader context of simulation modeling and analysis. We plan to conduct a more extensive user evaluation, such as a discussion panel, to assess the practical relevance and generalizability of our findings. Additionally, implementing a more formalized or automated approach could facilitate a more systematic face validity assessment, particularly as our approach primarily outlines what should be done without specifying how it can be achieved (e.g., selecting appropriate process mining techniques). Finally, it would be interesting to apply our proposed method to a simulation model that excels in terms of KPIs but struggles to simulate accurately outlier behavior.

Acknowledgements. We acknowledge the Helmholtz Information & Data Science Academy (HIDA) for providing financial support enabling a short-term research stay at Karlsruhe Institute of Technology (KIT), Germany. We express our gratitude to Ruben Govers for his assistance with the simulation study.

References

1. van der Aalst, W.M.P.: Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-19345-3
2. Alibrahim, A., Wu, S.: Modelling competition in health care markets as a complex adaptive system: an agent-based framework. Health Syst. **9**(3), 212–225 (2020)
3. Balci, O.: Validation, verification, and testing techniques throughout the life cycle of a simulation study. In: Proceedings of Winter Simulation Conference, pp. 215–220. IEEE (1994)
4. Bemthuis, R., Govers, R., Lazarova-Molnar, S.: Using process mining for face validity assessment in agent-based simulation models: an exploratory case study. In: Cooperative Information Systems (in press)
5. Bemthuis, R., Mes, M., Iacob, M.E., Havinga, P.: Using agent-based simulation for emergent behavior detection in cyber-physical systems. In: 2020 Winter Simulation Conference (WSC), pp. 230–241. IEEE (2020)
6. Bemthuis, R.H., Koot, M., Mes, M.R., Bukhsh, F.A., Iacob, M.E., Meratnia, N.: An agent-based process mining architecture for emergent behavior analysis. In: 2019 IEEE 23rd International Enterprise Distributed Object Computing Workshop (EDOCW), pp. 54–64. IEEE (2019)
7. Bemthuis, R.H., Lazarova-Molnar, S.: Discovering agent models using process mining: Initial approach and a case study. In: 2022 IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), pp. 163–172 (2022)
8. Cenani, S., Arentze, T.A., Timmermans, H.J.P.: Agent-based modeling of cognitive learning of dynamic activity-travel patterns. Transp. Lett. **5**(4), 184–200 (2013)
9. Cooley, P., Solano, E.: Agent-based model (ABM) validation considerations. In: Proceedings of the SIMUL 2011, The Third International Conference on Advances in System Simulation, pp. 134–139 (2011)
10. Day, R.S.: Challenges of biological realism and validation in simulation-based medical education. Artif. Intell. Med. **38**(1), 47–66 (2006)
11. Foramitti, J.: AgentPy: A package for agent-based modeling in Python. J. Open Source Softw. **6**(62), 3065 (2021)
12. Gomes, A.F.D., de Lacerda, A.C.W.G., da Silva Fialho, J.R.: Comparative analysis of process mining algorithms in process discover. In: de Paz Santana, J.F., de la Iglesia, D.H., López Rivero, A.J. (eds.) DiTTEt 2021. AISC, vol. 1410, pp. 258–270. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-87687-6_25
13. Günther, C.W., van der Aalst, W.M.P.: Fuzzy mining – adaptive process simplification based on multi-perspective metrics. In: Alonso, G., Dadam, P., Rosemann, M. (eds.) BPM 2007. LNCS, vol. 4714, pp. 328–343. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-75183-0_24
14. Günther, C.W., Rozinat, A.: Disco: discover your processes. BPM (Demos) **940**(1), 40–44 (2012)
15. Klügl, F.: A validation methodology for agent-based simulations. In: Proceedings of the 2008 ACM Symposium on Applied Computing, pp. 39–43 (2008)
16. Law, A.M.: How to build valid and credible simulation models. In: 2019 Winter Simulation Conference (WSC), pp. 1402–1414. IEEE (2019)
17. Leemans, S.J.J., Partington, A., Karnon, J., Wynn, M.T.: Process mining for healthcare decision analytics with micro-costing estimations. Artif. Intell. Med. **135**, 102473 (2023)

18. Llacay, B., Peffer, G.: Using realistic trading strategies in an agent-based stock market model. Comput. Math. Organ. Theory **24**, 308–350 (2018)
19. Martínez-Miranda, J., Pavón, J.: Modeling the influence of trust on work team performance. Simulation **88**(4), 408–436 (2012)
20. Midgley, D., Marks, R., Kunchamwar, D.: Building and assurance of agent-based models: an example and challenge to the field. J. Bus. Res. **60**(8), 884–893 (2007)
21. Peffers, K., Tuunanen, T., Rothenberger, M.A., Chatterjee, S.: A design science research methodology for information systems research. J. Manag. Inf. Syst. **24**(3), 45–77 (2007)
22. Robinson, S.: Simulation model verification and validation: increasing the users' confidence. In: Proceedings of the 29th Conference on Winter Simulation, pp. 53–59 (1997)
23. dos Santos Garcia, C., et al.: Process mining techniques and applications-a systematic mapping study. Expert Syst. Appl. **133**, 260–295 (2019)
24. Sargent, R.G.: Validation and verification of simulation models. In: Proceedings of the 24th Conference on Winter Simulation, pp. 104–114 (1992)
25. Sargent, R.G.: Verification and validation of simulation models. In: Proceedings of the 2010 Winter Simulation Conference, pp. 166–183. IEEE (2010)
26. Schelling, T.C.: Dynamic models of segregation. J. Math. Sociol. **1**(2), 143–186 (1971)
27. Seid, M., Bridgeland, D., Bridgeland, A., Hartley, D.M.: A collaborative learning health system agent-based model: computational and face validity. Learn. Health Syst. **5**(3), e10261 (2021)
28. Xiang, X., Kennedy, R., Madey, G., Cabaniss, S.: Verification and validation of agent-based scientific simulation models. In: Agent-Directed Simulation Conference, pp. 47–55 (2005)