Using Comparative Machine Learning Methods to Validate Educational Content

John Kos, Kenneth Eaton, Sareen Zhang, Rahul Dass, Ashok Goel

Georgia Institute of Technology
jkos3@gatech.edu, keaton30@gatech.edu, szhang702@gatech.edu, rdass7@gatech.edu, ashok.goel@cc.gatech.edu

Abstract

Validation as a field of study is important to the development of educational Interactive Learning Environments (ILEs), a type of software that allows dynamic and active engagement with educational material. However, as ILEs become more complex, borrowing from fields such as agent-based modeling, simulation, and "serious games", the educational domain has lagged in adopting the rigorous validation standards typical in these fields. Traditional methods, such as face validation by subject matter experts, are often criticized for their subjectivity and lack of thoroughness for validating pedagogical content or underlying theory. To address this, we present a machine learning-based methodology to validate the content and educational theory of ILEs in the context of complex systems. By demonstrating automated labeling of time-series data from VERA, an ecology focused agent-based modeling and simulation tool, we report a success rate of 92.79% on a manually collected sample. This promising result not only validates VERA but also suggests the broader applicability of our approach to other time series-based ILEs.

Introduction

Interactive Learning Environments (ILEs) are software enable educational tools that enable users to experience hand-on learning of complex subject matter. Further, validation is an important and well studied topic across the education domain. Educational applications and ILEs are often evaluated on a series of axis including learner perspective, system quality, ethical considerations, motivation, accessibility, etc (Lee and Kim 2015; Ozkan and Koseler 2009). Typically, in reference to how the educational content is presented, validation of ILEs focuses on the correctness of the vocabulary used or the procedural rules demonstrated in the ILE. (Ozkan and Koseler 2009; Kim and Lee 2008). However, this may become insufficient or intractable for emerging classes of ILEs.

The introduction of computers into classrooms has allowed for more complex representations of educational content to emerge, such as agent-based modeling, simulation, or "serious games " (Cook and Hatala 2016). Typically, a domain expert systematically assesses whether these tools accurately represent the educational content. This type of validation is referred to as face validation (Arifin and Madey 2015; Hermann 1967). However, with the increasing complexity of ILEs, face validation has proven insufficient, as it is often impractical to verify the content representation manually (Qiao et al. 2018; Niazi, Hussain, and Kolberg 2017). Often times these ILEs are so complex that it is infeasible for a human to verify that the ILE properly represents the content it is trying to teach (Cock et al. 2022). Put simply, face validation is slow and subjective.

Liu et al. describe "theory validity" as the alignment between the way educational content is presented within a learning tool and the actual information being taught (Liu et al. 2011). In simpler terms, it measures how accurately the educational tools represent the underlying concepts they aim to teach. Ensuring this type of validity is particularly challenging in ILEs with complex content representations. Developing an automated method to check for theory validity would allow educators to validate these complex pedagogical tools in a more robust manner.

This brings us to our research questions:

1. **Can we develop a method that can demonstrate theory validity in an automated fashion?**
2. **Can this method avoid the drawbacks of face validation (being slow and subjective)?**

To address this, we propose a novel methodology for demonstrating theory validity in an ecological modeling and simulation system called VERA. The Virtual Experimental Research Assistant or "VERA" employs agent-based modeling and simulation to help users create models of ecosystems and analyze population dynamics over time (An et al. 2018, 2020). These agent-based models are expressed in the Component-Mechanism-Phenomenon (CMP) language, an extension of Structure-Behavior-Function models (Joyner, Goel, and Papin 2014; Goel and Joyner 2015; Goel, Rugaber, and Vattam 2009).

The models represent the behavior of a system by defining its components and relations (An et al. 2020, 2018). This format of modeling allows users to define components such as biotics, abiotics, and habitats, as well as their relationships (An et al. 2018). Each component type has a set of up to 13 adjustable parameters that can be tailored to particular species, such as number of offspring, age of maturity, and age of death. Running the simulation allows the user to see a time-series graph representation of the population

size over time. For a domain expert to face validate VERA, they would need to check that any model a user might create aligns with typical time series population graphs found in ecology. In practice, this requires manually iterating through an innumerable number of model and parameter changes.

Our methodology, further outlined and justified in the Methodology Section, consists of tuning two machine learning models typically used in time-series analysis such that they have a high success when compared to a face validated dataset we collected from learners using VERA. We first develop a Hierarchical Clustering method in order to demonstrate that underlying patterns in time-series creation by VERA match Face Validated labeling. Hierarchical Clustering, by nature of being a bottom up method, ensures that any grouping of data is not influenced by human subjectivity. Next, we develop a curve fitting method to determine if we can automate the slow process of expert labeling. Agreement between these methods function as an answer to our second research question. We then apply these two methods to a test dataset in order to demonstrate that the method can generalize to all datasets created by VERA. This generalizability is proof of our first research question.

Our methodology revealed a 92.79% label agreement between curve fitting and hierarchical clustering on the test dataset. For the rest of this paper, unless otherwise clearly stated, we will refer to this agreement between methods (Curve Fitting and Clustering) as accuracy.

Background

This research aims to develop a method for determining the theory validity of open-domain pedagogical modeling, simulation, and "serious games" using content overlap metrics (Jia et al. 2022). According to Cook and Hatala (Cook and Hatala 2016), despite the increasingly popularity of simulation-based tools, there is a notable lack of any validation studies supporting their use within education. Research into the pedagogical validity of these tools is still in its infancy, with the majority of discussions on validation being limited to the medical field (Bogomolova et al. 2021; McGrath et al. 2018; Kong and Wang 2021). Although there has been research into using agent-based modeling and simulation tools for education, none of these studies discuss validity (Janssen, Lee, and Waring 2014; Day-Black 2015; De la Torre et al. 2021). Conversely, although there is substantial research on the empirical validity of agent-based modeling for scientific research, none has investigated the educational value of these tools (Moon and Bae 2015; Gu and Novak 2009).

For the majority of these tools, especially agent-based modeling and simulation tools like VERA, the most common method of validation, and typically the only method used, is Face Validation (Arifin and Madey 2015; Niazi, Hussain, and Kolberg 2017). In this method, a domain expert ensures that the tool reasonably represents the concepts they aim to demonstrate. This is a highly subjective process (Jia et al. 2022; Niazi, Hussain, and Kolberg 2017).

As shown by Cook (Cook et al. 2014) in an earlier paper concerning what counts as validity evidence, the two most popular forms of content evidence were "group consensus or expert review", and creating an instrument "based on (or modified from) a previously validated instrument." VERA has already undergone "expert review", by domain experts inside and outside our institution. In other words, VERA has been Face Validated. This paper presents a method by which ILEs can be validated against the "previous instrument" of mathematical models. Mathematical models are widely used in education as the theoretical foundation of many disciplines, for example in Economics (Windrum, Fagiolo, and Moneta 2007; Tesfatsion 2006), Ecology, and Epidemiology (Gu and Novak 2009). They serve as a ground truth which we validate against. However, when engaging in systems thinking, mathematical models become exponentially more complex and challenging to understand (Pielou 1981), hence the popularity of ILEs used to illustrate them.

Inspired by studies on the empirical validity of agent-based models, our research uses Model Docking (Arifin and Madey 2015), the process of validating against another instrument. This shows the content overlap between VERA and mathematical ecological models, thereby demonstrating VERA's "theory validity" (Liu et al. 2011). Similar comparisons have been used in Epidemiology (Gu and Novak 2009), and City Commerce Models (Moon and Bae 2015), however, these are attempts at proving the empirical validity of the agent based models instead of showing the education value.

Methodology

Our methodology employs two machine learning techniques: Curve Fitting and Hierarchical Clustering. Curve fitting was chosen as a well-established top-down approach to classifying time series data (Eberhardt, Breiwick, and Demaster 2008). This approach involves selecting a set of ecological mathematical models (curves) and assessing their fit with VERA's time series output. Curve Fitting is often more adaptable than other empirical methods and is effective at suppressing noise (Zeng et al. 2020). Additionally, Curve Fitting has been used extensively for extracting vegetation phenological metrics with different curve types including Logistic (Cao et al. 2015) and Gaussian (Jonsson and Eklundh 2002) curves. By using this approach, we can label each curve and compare that to the results found through the clustering method.

Conversely, clustering is a bottom-up method that groups similar data samples within unlabeled data (Murtagh and Contreras 2017; Kassambara 2017; Luczak 2016). We focus on agglomerative Hierarchical Clustering, which begins with each data sample as its own cluster and iteratively merges the closest pairs of clusters until they form one cluster (Omran, Engelbrecht, and Salman 2007). Although this approach has previously been applied to a variety of time series datasets, which served as inspiration for our use, we aim to apply it within the context of theory validity at the intersection of ecology and education.

Both curve fitting and hierarchical clustering methods are applied to the time series output generated by VERA to cross-validate the results. The curve fitting approach assigns a label to each time-series based on predefined mathematical models. This can obscure the distinction between noise

and significant features of the data. In contrast, the clustering approach treats all portions of the data equally, preserving patterns that may be overlooked by curve fitting. By identifying general cluster boundaries that align with the curve fitting methods, while acknowledging potential outliers within clusters, we can affirm that VERA's simulation output reinforces the population curves taught in the classroom.

Ecological Curves

Prior research on self-directed learning sought to better understand the types of models created by people outside classroom settings, where the models created are not influenced by any assignment (An et al. 2022). To understand model behaviors, we aimed to align the simulation outputs with expected ecological relationships. For example, oscillation could represent a predator-prey relationship where the populations cycle with each other (An et al. 2020).

While these relationships are well-defined ecologically, it was not clear how to align them with the simulation outputs. This serves as the initial motivation for the development of this methodology. While in practical terms, this methodology allows us to classify the simulation outputs of VERA according to mathematical models founds in ecology, in theoretical terms, it offers another avenue of demonstrating theory validity for the system as a whole.

In this study, we classify simulation outputs using seven distinct curve categories derived from ecological mathematical modeling. These categories include Constant (Con), Dying Off (Die), Oscillation (Osc), Exponential Growth (Exp), Capped Growth (Cap), and Gaussian (Gau). Additionally, we introduced an 'outlier' (Out) category for population graphs that do not align with any of the aforementioned curves. These categories serve as the basis for applying machine learning methods to classify the simulation data.

The curve fitting approach was then developed with these meaningful curves in mind, however, many different other types of curves were tested but were found to be unsuccessful, either not being chosen by the curve fitting or only being chosen incorrectly. In Figure 1, we can see all the different curve types, not including constant which would just be a straight line across the top. While this classification list is non-exhaustive of all the curves in ecological modeling, we will see that it is comprehensive enough for testing purposes.

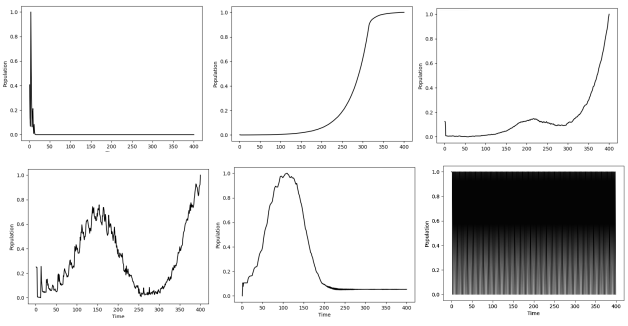

Figure 1: From top left to bottom right (1) dying off, (2) capped growth, (3) exponential, (4) oscillation, (5) gaussian, (6) outlier

Datasets and Data Processing

As part of the manual process to collect our two datasets (training and test data sets), we first cloned each model and then ran simulations to re-create the simulated data. This was necessary because simulation output is currently not stored in the VERA database. The time series generated by the simulation was then downloaded as CSV files.

The training dataset was comprised of self-directed learners, who are learners with no known ties to any institutional partnerships, from January 1st of 2019 to December 31st of 2021. It was made up of 197 VERA projects, which comprised 724 species (curves). For every curve in the training dataset, we used manual expert labeling accounting for around 20 hours worth of effort. We then ran the two methods on this dataset and used it to set the hyperparameters for the system, such that both curve fitting and clustering both individually returned the highest accuracy when compared to the expert labels. For the clustering method, we determined that the optimal distance measure was 30 and the outlier classification threshold was 55%, as detailed later in the paper. For the curve fitting method, we set the error threshold hyperparameter to 5.7, based on the residual sum of squares. Any curve with an error above this threshold was classified as an outlier.

The test dataset is comprised of a random sample of all models developed by learners using the VERA application from January 1st of 2022 to September 1, 2023. In total, the sample consisted of 263 VERA projects and 971 total species (read curves). Further description of the character of the datasets in given in the Results Section.

To account for variations in the population scales in simulation, ranging from 0-25,000 units, we normalized the data to range of zero to one based on the maximum population value in each graph. We standardized all data lengths to 400 values (representing 400 months in simulation) to ensure uniformity across samples. This is the default simulation length in VERA, though users have the flexibility to lengthen or shorten the simulation length. Graphs with fewer than 400 data points were excluded from the sample, as they do not provide sufficient space for complex data patterns to fully develop.

Ethics Review

When users create a VERA account, whether they are students enrolled in academic courses or public users, they are required to provide informed consent that their data, while anonymized, will be used for research and analysis by VERA developers. Since VERA is accessible to the public, learner models stored in its database include contributions from both academic and public users.

We have access to demographic information such as age, gender, and race for student users, which can be acquired through course registers. However, VERA's login infrastructure does not collect or store this demographic data for public users. For this study, we manually collected a random sample of learner models from the VERA database, spanning the years 2019-2023, in a fully anonymized fashion. This sample was used to analyze the educational impact and

theoretical validity of VERA, ensuring that the data remains anonymized and compliant with our standing Institutional Review Board (IRB) protocol. This protocol governs all system log-level actions within VERA, permitting the analysis of user data while ensuring that no Personally Identifiable Information (PII) from external sources is used.

Curve Fitting

Curve fitting was performed by inputting the expected fundamental curves, their corresponding mathematical equations, and the time series data into SciPy's *curve_fit* function. The *curve_fit* function returned parameters for the mathematical equation that minimized the residual sum of squares, ensuring the best fit to the given time series. Each time series was then tested against all curve types to identify the one that minimized error. Figure 2 shows a plot where SciPy attempts to fit different curves to the data function. Here Gaussian would be chosen as the label for the data.

Given the variability in population graph scales and the multidimensionality of the parameter space, four methods were taken to ensure successful curve matching: (1) normalizing the time series, (2) dividing the parameter space and running search for each subsection, (3) providing the curve fitting function with a set of prototypical parameters, and (4) using simple rule-based methods for the simplest curves.

The first three methods constrained the search space, increasing the likelihood of identifying the correct curve label. The final method was applied to each curve as a check to determine if it could be identified before using more data intensive methods. In practice, this meant that the *constant* and *exponential dying* curves were evaluated using rules-based methods. These specific curves are simple to evaluate using rules because they are the most likely to be represented as a constant series of values within the time-series. For example, dying out would indicate that the population goes to and remains at zero. Similarly, for constant, the population does not change. Identifying these curves with simple rules instead of machine learning methods allows us to cut down on computation time.

To detect outliers, error was evaluated by measuring the residual sum of square between the fitted curve and the given population graph. The error threshold hyperparameter was manually tuned on the training set in order return best results when compared to expert labels, which gave a value of 5.7. Any curve above that threshold was then classified as an outlier.

Hierarchical Clustering

To cluster the data, we used Agglomerative Hierarchical Clustering with complete linkage, where clusters were merged based on the farthest distance between points in the different clusters. We removed constant curves types from the training dataset to focus the clustering on curves that have varying features, which are more likely to reveal meaningful patterns and groupings in the data. This is done by creating a set of all values from each curve in Python. If the set length was one, the curve was identified as *constant* and removed from clustering.

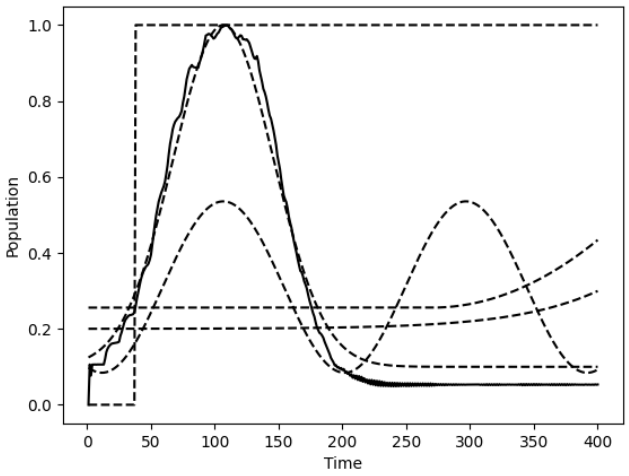

Figure 2: Visualization of Curve Fitting. Solid line is data. Dashed lines are attempts to fit mathematical models.

A dendrogram is created using SciPy and then it is flattened using the SciPy fcluster function. This separates the dendorgram into clusters based on euclidean distance. Within each cluster a prototypical representative, the mediod, was found by determining which curve has the least distance between itself and all of the other members. For cluster classification, clusters were categorized based on the label attached to the representative curve. This label was at first determined using the expert labeling during the tuning of the methods during the training set, and then by the curve fitting for the evaluation of the training and test sets. The label for each member of the cluster is then also evaluated in order to see if it matched the representatives' label in more than 55% of cases. If the 55% threshold was not met, the entire cluster was classified as outlier. This is to avoid cases where the representative does not match the labels of the majority of curves in the cluster. We refer to this last step as the 'voting step'.

In summation, hierarchical cluster consists of three steps, (1) clustering based on a set distance function, (2) labeling of the cluster based on the mediod curve, and (3) a voting step where greater than 55% of curve labels in a cluster must match the representative label or else the cluster is classified as an outlier.

Results

The results section will consist of the following. First, we will give a breakdown of the number of samples per curve category of the datasets used. Next, we will then compare the training data success rates of each method—curve fitting and clustering—against the expert-labeled data. Last, we will demonstrate the agreement between the methods for both the training set and the test set, illustrating that the method (1) finds agreement between curve fittings and clustering and (2) generalizes to new datasets. As we will see, the agreement between the methods confirms our second research question, and the generalization to new datasets confirms our first research question.

Description of the datasets

Table 1 contains a description of the curves found in the datasets separated by curve type, as labeled by the curve fitting methodology. The training dataset was made up of self-directed learners who were working with VERA for their unknown own goals. For this reason, we see a smaller distribution of the curve types that are more complex to create within VERA, exponential, oscillation, and Gaussian. This is likely because these learners lacked any instruction on how to use VERA. Conversely, we see a larger percentage of these types of curves in our test dataset. This is because our test dataset was a random sample of all users on the VERA platform, including students within higher education institutions. These students likely got instruction of some sort and for that reason were more capable of creating more complex curves.

Dataset	Curve Categories							Total
	Exp	Cap	Die	Osc	Gau	Con	Out	
Training Set	6	106	291	14	21	155	131	724
Test Set	29	75	420	69	169	162	47	971

Table 1: Distribution of curves across seven different categories for Training Set and Test Set. The curve categories are Exponential Growth (Exp), etc.

Training compared to expert labels

As part of developing both of these methods, we tuned the hyperparameters such that both methods returned the highest success rate as compared to expert labels. This returned a success rate of around 91% for both methods as shown in Table 2. Given the size of the manually collected sample, and the similar rates of accuracy, we took this as indication that comparison between the methods would be successful.

Compared to Expert Labels	Success Rate
Curve Fitting	91.57%
Hierarchical Clustering	91.44%

Table 2: Accuracy of two machine learning methods compared to expert labels.

Findings from automated labeling

Table 3 describes the success rate of both methods before and after the voting step. The difference between the success rate after voting indicates the proportion of clusters where more than 55% of the curve labels did not match the representative label. Most often, this is a function of cluster size. For example, a cluster of size two where one curve is capped growth and the other is outlier would then be reclassified as outlier, lowering the success percentage based on the misclassification of the capped growth curve.

The decrease in success rate after the voting step for the test set was smaller than the decrease seen for the training set. The accuracy difference for training set was 5.54%, while the difference for the test set was 1.46% This final rate of the test set, 92.79% was higher than either of the success rates (see Table 2) for the methodologies individually when compared to the expert labeling.

Success Rate %	Primary Classification	After Voting Step
Training Set	92.69	87.15
Test Set	94.25	92.79

Table 3: Comparing the success rates of both machine learning methods before and after voting step.

Discussion

Our research questions were as follows: (1) Can we develop a method that can demonstrate theory validity in an automated fashion? and (2) Can this method avoid the drawbacks of face validation (being slow and subjective)?

When individually compared to expert labeling, we see the methods reach a success rate of around 91%. The similar rates of success between the method suggest that both methods might be identifying similar underlying patterns in the simulation outputs of VERA, that were not visible to the domain expert. On the test set, we see a success rate of 92.69% - a rate higher than that found in the comparison to expert labeling - which proves this fact. The higher rate of success demonstrates (1) curve fitting is faster than expert labeling without any decrease in accuracy, and (2) the combination of the two methods ensures a level of objectivity that even an expert is not capable of achieving. This is confirmation of our second research question, as our methodology performs similarly to face validation. The two methods generalizability to the test dataset, demonstrates that this is a reliable and automated method to demonstrate theory validity, confirming our first research question.

In this research, we developed a methodology that could theory validate (Liu et al. 2011) certain classes of complex ILEs where face validation proves to be insufficient due to its slow and subjective nature (Niazi, Hussain, and Kolberg 2017; Jia et al. 2022). Additionally, this research exists to fill the gap in validation of agent-based modeling and simulation literature in an educational context as described by Cook et al. (Cook and Hatala 2016). More specifically, taking inspiration from model docking, (Arifin and Madey 2015), we demonstrate that the agent-based ecological modeling done in VERA has content overlap (Jia et al. 2022) with the mathematical models foundational to the study of ecology. To the best knowledge of the authors, this research is the first of its kind to bridge the gap between pedagogical agent-based modeling (Janssen, Lee, and Waring 2014) and validation methods from empirically focused agent-based modeling (Gu and Novak 2009).

Size of Dataset

We believe that our results would improve with a larger dataset. The method's accuracy improved from the training set to the test set, surpassing the accuracy of expert labeling. The test set is around 25% larger than the training set. The

increase in success rate over expert labeling from the training set to the test set supports this observation. This is further illustrated by the training set's difficulty in clustering exponential curves, whereas the test set contained five smaller clusters. This suggests that the training set lacked sufficient curves, both in total and by type, to form meaningful clusters for every curve type, as evidenced by the distribution of curves between the training and test sets (see Table 1). Despite this, the methodology remained robust and performed well on the test set.

Limitations

The main thrust of this paper is to focus on developing an automated methodology to demonstrate a content overlap, and therefore theory validity, between VERA's simulation outputs and mathematical models in ecology.

This research only considered assessing the utility of the ML methods based on a single set of expert labels. Inter-expert agreement was not considered and could be a possible avenue for improvement. This would allow us to not only evaluate whether curve fitting rates as highly as expert labeling, but also would allow us to quantify the degree of subjectivity in the labeling task for VERA specifically.

Lastly, this research is only focused on ILEs with complex representations founded in mathematical models. It does not apply to ILEs with simple relationships to mathematical models with only one or two parameters (Moore et al. 2014), coding, writing, creative acts, etc.

Future work

The clearest direction for future work is the application of this research to a larger dataset. This dataset would likely be the entirety of the VERA database, made of up over 7000 models. Additionally, a sensitivity analysis of VERA would allow for this analysis to be done across all possible models that can be created in VERA. The sensitivity analysis offers a guarantee of the value of the methodology that a sample of, or even the entire population of, user models does not.

More analysis is needed to further determine the nature of the outlier curves, but as described above there is some indication that a larger dataset would decrease the number of clusters classified as outlier during the voting step. This would happen through two processes. A larger dataset should converge on specific clusters for outlier curves of similar types, while simultaneously classifying less clusters as outlier due to a lack of confluence in cluster labels.

Additionally, patterns in these curves may reveal either more complex known ecological patterns currently unknown to VERA developers or it may reveal something corresponding to the software of VERA itself. Further research into outlier patterns is the best way to differentiate between these two possible causes, and both lead to a greater understanding of the VERA system as a whole.

Conclusion

This work investigates two research questions: (1) Can we develop a method that can demonstrate theory validity in an automated fashion? (2) Can we build this method to circumvent the slow and subjective nature of face validation?

To answer these research question, we developed a methodology that enables automatic validation of theory validity for complex interactive learning environments (ILEs) based on mathematical models, which would be impractical to exhaustively validate manually. The high rate of agreement between clustering and curve fitting, especially when compared to the success rate of the individual methods to expert labeling, gives us confirmation of our second research question. That is, our two methods, when paired together, help circumvent the slow and subjective nature of face validation while still labeling time-series with high accuracy. The overarching methodology's generalizability and success rate of 92.79% on a test set is confirmation of our first research question. This methodology can serve as a useful tool to validate complex ILEs, that are based on mathematical models, in an automated fashion. Further, this methodology validates VERA's conceptual models and simulation, and justifies its continued use in higher education.

Acknowledgments

This research has been supported by NSF Grants #2112532 and #2247790 to the National AI Institute for Adult Learning and Online Education. We thank members of the VERA project in the Design Intelligence Laboratory for their contributions to this work.

References

An, S.; et al. 2018. VERA: popularizing science through AI. In *Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part II 19*, 31–35. Spring International Publishing.

An, S.; et al. 2020. Scientific modeling using large scale knowledge. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21*, 20–24.

An, S.; et al. 2022. Understanding Self-Directed Learning with Sequential Pattern Mining. *Aritifical Intelligence in Education*, 502–505.

Arifin, S. N.; and Madey, G. R. 2015. Verification, validation, and replication methods for agent-based modeling and simulation: lessons learned the hard way! *Concepts and methodologies for modeling and simulation: A tribute to Tuncer Oren*.

Bogomolova, K.; et al. 2021. Development of a virtual three-dimensional assessment scenario for anatomical education. *Anatomical sciences education*, 14.

Cao, R.; et al. 2015. An improved logistic method for detecting spring vegetation phenology in grasslands from MODIS EVI time-series data. *Agricultural and Forest Meteorology*, 200: 9–20.

Cock, J. M.; et al. 2022. Generalisable methods for early prediction in interactive simulations for education. arXiv:2207.01457.

Cook, D. A.; and Hatala, R. 2016. Validation of educational assessments: a primer for simulation and beyond. *Advances in simulation*, 1: 1–12.

Cook, D. A.; et al. 2014. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Advances in Health Sciences Education*, 19.

Day-Black, C. 2015. Gamification: An Innovative Teaching-Learning Strategy for the Digital Nursing Students in a Community Health Nursing Course. *ABNF Journal*, 26.

De la Torre, R.; et al. 2021. The role of simulation and serious games in teaching concepts on circular economy and sustainable energy. *Energies*, 14: 1138.

Eberhardt, L. L.; Breiwick, J. M.; and Demaster, D. P. 2008. Analyzing population growth curves. *Oikos*, 117: 1240–1246.

Goel, A.; and Joyner, D. 2015. Impact of a creativity support tool on student learning about scientific discovery processes. In *Proceedings of the Sixth International Conference on Computational Creativity*.

Goel, A. K.; Rugaber, S.; and Vattam, S. 2009. Structure, behavior, and function of complex systems: The structure, behavior, and function modeling language. *Ai Edam*, 23.

Gu, W.; and Novak, R. J. 2009. Agent-based modelling of mosquito foraging behaviour for malaria control. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 103: 1105–1112.

Hermann, C. F. 1967. Validation problems in games and simulations with special reference to models of international politics. *Behavioral Science*, 12: 216–231.

Janssen, M. A.; Lee, A.; and Waring, T. M. 2014. Experimental platforms for behavioral experiments on social-ecological systems. *Ecology and Society*, 19.

Jia, Q.; et al. 2022. Insta-Reviewer: A Data-Driven Approach for Generating Instant Feedback on Students' Project Reports. *International Educational Data Mining Society*.

Jonsson, P.; and Eklundh, L. 2002. Seasonality extraction by function fitting to time-series of satellite sensor data. *IEEE transactions on Geoscience and Remote Sensing*, 40(8): 1824–1832.

Joyner, D. A.; Goel, A. K.; and Papin, N. M. 2014. MILA–S: generation of agent-based simulations from conceptual models of complex systems. In *Proceedings of the 19th international conference on intelligent user interfaces*, 289–298.

Kassambara, A. 2017. *Practical guide to cluster analysis in R: Unsupervised machine learning*, volume 1. Sthda.

Kim, S. W.; and Lee, M. G. 2008. Validation of an evaluation model for learning management systems. *Journal of Computer Assisted Learning*, 24: 284–294.

Kong, S. C.; and Wang, Y. Q. 2021. Item response analysis of computational thinking practices: Test characteristics and students' learning abilities in visual programming contexts. *Computers in Human Behavior*, 122.

Lee, J. S.; and Kim, S. W. 2015. Validation of a tool evaluating educational apps for smart education. *Journal of Educational Computing Research*, 52: 435–450.

Liu, F. C.; et al. 2011. From Internal Validation to Sensitivity Test: How Grid Computing Facilitates the Construction of an Agent-Based Simulation in Social Sciences. In *In Proceedings of The International Symposium on Grids and Clouds and the Open Grid Forum—PoS (ISGC 2011 and OGF 31).*, 2.

Luczak, M. 2016. Hierarchical clustering of time series data with parametric derivative dynamic time warping. *Expert Systems with Applications*, 62: 116–130.

McGrath, J. L.; et al. 2018. Using virtual reality simulation environments to assess competence for emergency medicine learners. *Academic Emergency Medicine*, 25.

Moon, L. C.; and Bae, J. W. 2015. Comparisons of Validated Agent-Based Model and Calibrated Statistical Model. *Concepts and methodologies for modeling and simulation: A tribute to Tuncer Oren*, 243–256.

Moore, E. B.; et al. 2014. PhET interactive simulations: Transformative tools for teaching chemistry. *Journal of chemical education*, 91: 1191–1197.

Murtagh, F.; and Contreras, P. 2017. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2.

Niazi, M. A.; Hussain, A.; and Kolberg, M. 2017. Verification and validation of agent based simulations using the VOMAS (virtual overlay multi-agent system) approach. arXiv:1708.02361.

Omran, M. G.; Engelbrecht, A. P.; and Salman, A. 2007. An overview of clustering methods. *Intelligent Data Analysis*, 11(6): 583–605.

Ozkan, S.; and Koseler, R. 2009. Multi-dimensional students' evaluation of e-learning systems in the higher education context: An empirical investigation. *Computers and Education*, 53: 1285–1296.

Pielou, E. 1981. The usefulness of ecological models: a stock-taking. *The Quarterly Review of Biology*, 56: 17–31.

Qiao, X.; et al. 2018. Source apportionment of PM2. 5 for 25 Chinese provincial capitals and municipalities using a source-oriented Community Multiscale Air Quality model. *Science of the Total Environment*, 612: 462–471.

Tesfatsion, L. 2006. Agent-based computational economics: A constructive approach to economic theory. *Handbook of computational economics*, 2: 831–880.

Windrum, P.; Fagiolo, G.; and Moneta, A. 2007. Empirical validation of agent-based models: Alternatives and prospects. *Journal of Artificial Societies and Social Simulation*, 10: 8.

Zeng, L.; et al. 2020. A review of vegetation phenological metrics extraction using time-series, multispectral satellite data. *Remote Sensing of Environment*, 237: 111511.