Advances in Cognitive Systems X (20XX) 1-6 Submitted X/20XX; published X/20XX

Contextualized Access to Large-Scale Domain Knowledge for Conceptual Modeling of Agent-Based Systems

Sungeun An[1] SUNGEUN.AN@GATECH.EDU
Jennifer Hammock[2] JHAMMOCK@SI.ORG
Spencer Rugaber[1] SPENCER@CC.GATECH.EDU
Ashok K. Goel[1] GOEL@CC.GATECH.EDU

[1] School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA 30308 USA

[2] National Museum of Natural History, Smithsonian Institution, Washington D.C. 20002 USA

Abstract

Modeling is a core process of cognition and learning in science. However, model construction is cognitively challenging in part because it requires domain knowledge. We describe an experiment in contextualizing access to large-scale domain knowledge for constructing conceptual models of agent-based systems. In particular, VERA is an interactive modeling environment for constructing conceptual models of ecological phenomena that provides access to large-scale knowledge in Encyclopedia of Life (EOL). We found that contextualized access to EOL's large-scale domain knowledge results in the construction of deeper conceptual models. Specifically, we found that access to EOL through VERA empowered self-directed learners to build more complex models and tailor the parameter values of the ecological models to their personal preferences.

1. Introduction

Modeling is a core process of cognition in science (Bradshaw et al., 1983; Darden, 1998; Dunbar & Fugelsang, 2005; Griffith et al., 2000, 2019; Langley et al., 1987; Schwarz & White, 2005; Clement, 2008; Nersessian, 2010). In general, scientists understand complex phenomena by constructing, evaluating, revising, and using hypotheses and models of the given phenomenon. It follows that modeling is also a core process of learning in science. Authentic learning requires that learners learn about scientific thinking and acquire scientific knowledge through a process that imitates that of scientists (Edelson, 1998; Hogan & Thomas, 2001; VanLehn, 2013; White & Frederiksen, 1990).

Models however can be of many types, for example, conceptual models, process models, mathematical models, statistical models, diagrammatic models, etc. The "right" nature of the model depends in part on the nature of the system being modeled and partly on the intended use and user of the model. For systems in which the global behaviors of a system emerge out of numerous local interactions among autonomous agents inhabiting the system, agent-based modeling has become very popular since the advent of modern computing. Economics, ecology, and epidemiology are just a few examples of agent-based domains.

Model construction is cognitively challenging for most learners (Clement, 2008; Goel et al., 2013; Hogan & Thomas, 2001; Schwarz & White, 2005; Sins et al., 2005; White & Frederiksen, 1990). This is in part because learners are learning not only the modeling language and process of modeling but also acquiring the domain knowledge needed for modeling the system. Agent-based modeling in particular requires expertise in computer programming, mathematical equations, and stochastic processes that most learners may not have. Thus, the question becomes: How can we provide learners with interactive tools for easily constructing agent-based models? In the MILA interactive modeling environment, Joyner et al. (2014), middle school students construct conceptual models of an ecological phenomenon and MILA automatically spawns agent-based simulations from the conceptual model to help the learners evaluate their conceptual model.

However, an open question from the MILA project is: How can we provide learners with access to the domain knowledge they need for constructing conceptual models of ecological phenomena and spawning agent-based simulations? While large amounts of knowledge about many domains are now readily accessible on the internet, much of this general-purpose knowledge is not specific to any problem-solving context and thus difficult to apply by many learners. In contrast, the interactive learning environment called VERA provides contextualized access to large-scale knowledge in Encyclopedia of Life (EOL), making knowledge specific to a particular problem-solving context.

In this paper, we describe the knowledge engineering required to integrate VERA with EOL. The choice of the domain in ecology is deliberate: Ecology is a scientific domain of growing interest due to its close relationship to environmental sustainability. Further, ecology has attracted much attention in the cognitive systems community (Bridewell et al., 2008; Leelawong & Biswas, 2008; Salles & Bredeweg, 2003). Leelawong & Biswas (2008), for example, describe Betty's Brain that supports middle school science students in learning ecological knowledge by teaching a virtual agent (named Betty) about how to solve an ecological problem. VERA takes a different approach, providing contextualized access to large-scale ecological knowledge.

To test our hypothesis about contextualized access to large-scale domain knowledge resulting in construction of deeper models, we provided public access to VERA through EOL. It is important to note that in this study, we do not know the learning goals, assessments, or outcomes, or even the demographics of the learners or their precise geographical location. Instead, we have access only to data only on the modeling behaviors of the learners (the log data) and the modeling outcomes (final work products). This is an example of self-directed learning (likely mostly by adults) to address ill-defined real-world problems. Our analysis indicates that contextualized access to domain knowledge helped self-directed learners build more complex models as measured by the number of components in the model as well as more specific models as measured by the setting of the parameter values of the models.

2. Related Work

The domain of ecology admits models of several kinds including both conceptual models and agent-based models of complex phenomena. Conceptual models of ecological phenomena are declarative representations of the components of an ecological system and the relationships among them. While conceptual models may specify the parameters characterizing the various components and relation-

ships among the component parameters, conceptual models are qualitative. The publicly and freely available VERA interactive modeling environment (vera.cc.gatech.edu) (An et al., 2020, 2021a,b, 2022) uses the Component-Mechanism-Phenomenon (CMP) language for expressing models of ecological phenomena (Joyner et al., 2011). CMP is a variation of the Structure-Behavior-Function (SBF) language for conceptual modeling of complex systems in general (Goel et al., 1996, 2009): we designed CMP specifically for modeling biological systems. SBF itself originates in the Functional Representation scheme (Chandrasekaran, 1994; Goel, 2013).

In contrast, agent-based simulations derive the global behavior of an ecological system from local interactions among the different species in the system (Railsback & Grimm, 2019; Odum & Odum, 2000). Agent-based simulations are quantitative and stochastic. They take numerical values for the system parameters as input and derive the temporal evolution of the values of selected parameter values as output. VERA uses the publicly and freely available NetLogo agent-based simulation platform (ccl.northwestern.edu/netlogo) (Wilensky & Resnick, 1999).

Conceptual models and agent-based models each have unique affordances and limitations and can be considered complementary to each other (De Jong & Van Joolingen, 1998; Metcalf et al., 2000; Vattam et al., 2009; Joyner et al., 2014; VanLehn, 2013). Conceptual models typically are easier to understand due to their graphical and conceptual nature; they lend themselves to rapid construction, evaluation, and revision. Agent-based simulations tend to be more precise, detailed, and rigorous than conceptual models. On the other hand, agent-based simulations are difficult to set up for novice learners and are not as suited to rapid construction, use, and revision as conceptual models.

Most complex ecological systems have many parameters. Many interactive learning environments in ecology support model parameterization either by providing pre-defined qualitative components and relationships (van Joolingen et al., 2005; Leelawong & Biswas, 2008) or by suggesting values for components and relationships that best fit an existing data set (Bridewell et al., 2008; Broniec et al., 2021). However, these approaches are not always suitable for ecological modeling in two ways. First, given a large number of biological species in nature (about 10 million species currently in existence) and the large number of ecological phenomena being modeled, it is nearly impossible to provide pre-defined components complete with parameter values for all phenomena of potential interest. Second, ecological population data is often difficult to obtain because it requires long-term observations, and experimentation with real systems is limited (Salles & Bredeweg, 2003). This is in part why citizen science is often used as an ecological research tool for crowdsourcing data collection over large geographic regions (Dickinson et al., 2010; Howe et al., 2006). Consequently, ecological population data to compare and contrast to a model is relatively sparse in ecology compared to many other science domains.

Our research on the VERA interactive modeling environment builds on previous research on inquiry-based modeling in ACT (Goel et al., 2013; Vattam et al., 2009), EMT (Joyner et al., 2011), and MILA (Joyner et al., 2014; Joyner & Goel, 2015). ACT provided middle school science students with an interactive tool to build SBF models of ecological systems as well as an expert's agent-based simulation of the ecological system. EMT used CMP models that adapted SBF modeling to biological systems. MILA automatically translated the CMP models into agent-based simulations. Since the agent-based simulations are generated based on the conceptual models, this preserves the ca-

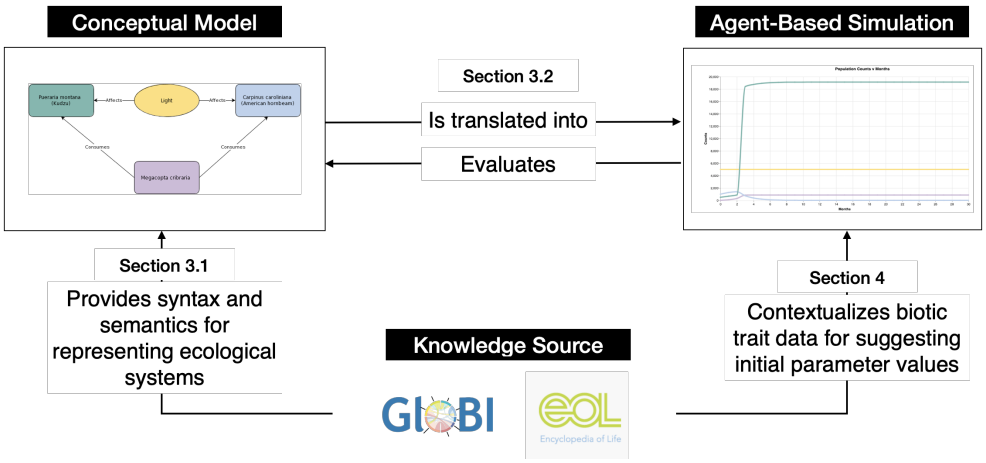

Figure 1. Schematic Overview of Using Domain Knowledge for Conceptual Modeling and Simulation.

pacity for rapid revision and knowledge sharing allowed by the conceptual models while extending them to provide the repeated testing and feedback of the simulations (Joyner et al., 2014). Note that while learners are able to run the agent-based simulation without having to know the modeling language used in the simulations, they still need domain knowledge for model parameterization, selecting reasonable values for the simulation parameters. Novice learners' difficulty with parameterization is well known in the literature (An et al., 2021a,b; Sins et al., 2005; Hogan & Thomas, 2001).

3. VERA

VERA (Virtual Experimentation Research Assistant; vera.cc.gatech.edu) is an online learning environment that enables learners to construct conceptual models of ecological systems and run agent-based simulations of these models (An et al., 2020). By model construction, learners can explore ecological systems and perform "what-if" experiments to either explain an existing ecological system or attempt to predict the outcome of future changes to one.

This section introduces the design of VERA for conceptual modeling and executing the agent-based simulation. As shown in Figure 1, Section 3.1 describes VERA's syntax and semantics for a conceptual model based on the ontology of attributes and ecological interactions used by Global Biotic Interactions (GloBI) (Poelen et al., 2014) and Encyclopedia of Life (EOL) (Parr et al., 2016). Section 3.2 describes an AI compiler that translates the conceptual models into agent-based simulation. The next section (Section 4) describes how VERA retrieves and contextualizes large-scale domain knowledge for suggesting initial parameter values.

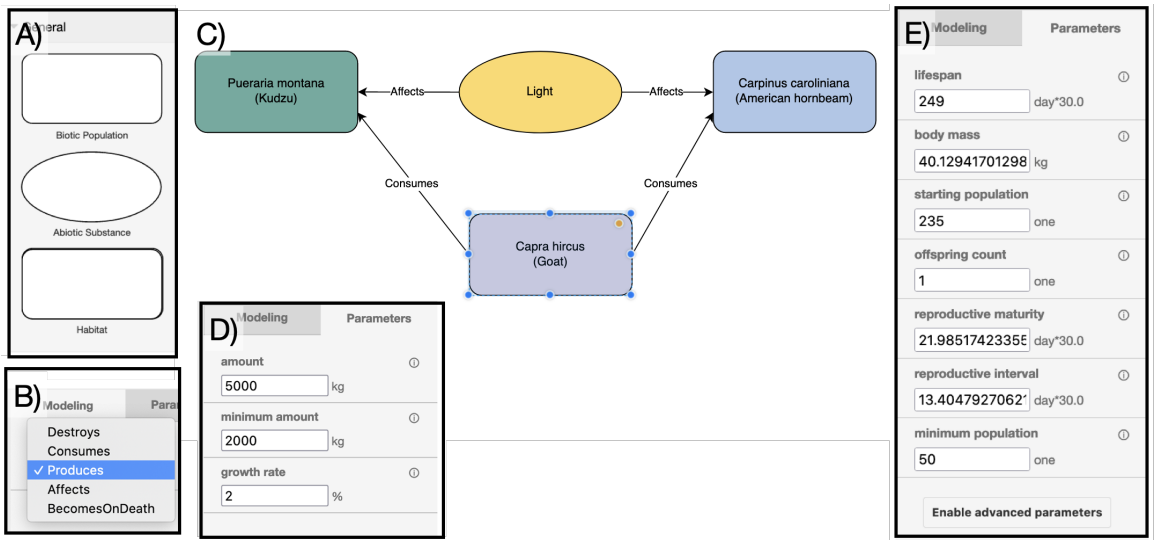

Figure 2. Conceptual Modeling Canvas in VERA. A) Components: Biotic, Abiotic, Habitat. B) Mechanism: Consume, Destroy, Produce, etc. C) A Conceptual Model. D) Abiotic Properties (Light). E) Biotic Properties (Goat).

3.1 Qualitative Conceptual Modeling

As mentioned above, VERA uses the Component-Mechanism-Phenomenon (CMP) language, a variation of the Structure-Behavior-Function (SBF) language, for conceptual modeling of ecological phenomena. A conceptual model consists of interacting components (e.g., biotic, abiotic), relationships (e.g., consume, destroy, etc.), and their parameter values (e.g., initial population, lifespan, etc.). A learner can add a biotic (rectangular) and/or abiotic (ellipse) component (Figure 2A) and draw a directed relationship between two components (Figure 2B). Figure 2C illustrates a causal conceptual model of kudzu (*Pueraria Montana*), goat (*Capra Hircus*), and American hornbeam (*Carpinus Caroliniana*) in the Southern United States.

To describe casual relationships among components in an ecological system, we used the ontology of the interactions from a digital library called Global Biotic Interactions (GloBI) (Poelen et al., 2014). We reduced 22 interaction types to five interaction types by merging them (see Figure 2B). Examples of such component-to-component interactions include *consumes* (one biotic organism consuming another), *produces* (a biotic organism producing an abiotic substance), and *destroys* (an abiotic substance harming a biotic organism).

The biotic properties (simulation parameters) were selected from the Encyclopedia of Life (EOL) (Parr et al., 2016) and adapted for ecological modeling. Examples of the biotic properties are *lifespan, body mass, offspring count, reproductive maturity, reproductive interval, etc.* (see Figure 2E). Table 1 gives a comprehensive list of biotic properties used in VERA, which are used to produce an agent-based simulation.

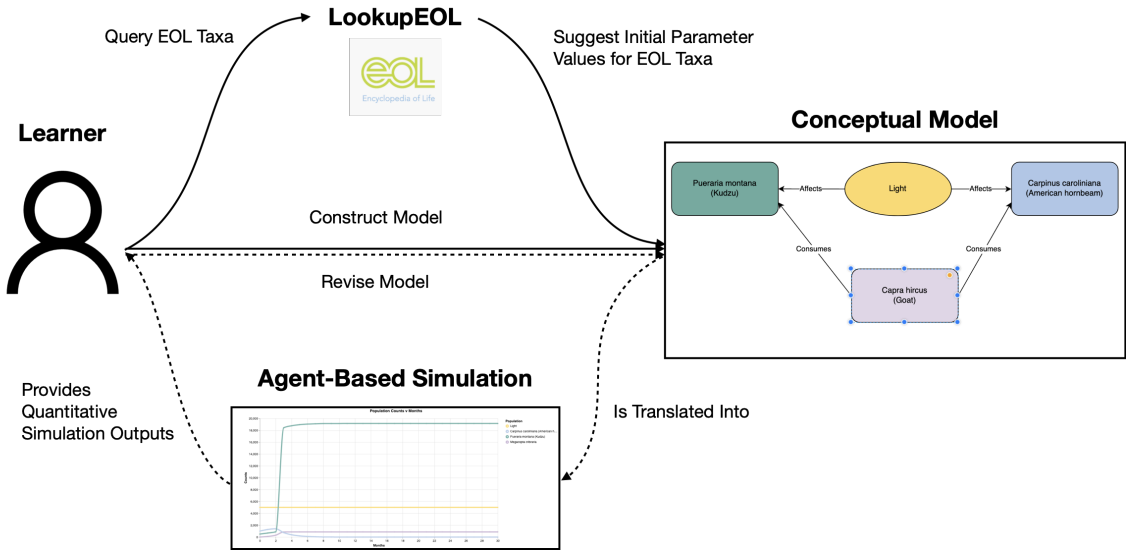

Figure 3. Learner-VERA Interaction Design.

3.2 Bridging the Gap between Conceptual Modeling and Quantitative Simulations

Agent-based models are suitable for simulating complex systems in which the components of the system are determined dynamically and vary with time and circumstances (Lippe et al., 2019). Ecological systems are comprised of many individual agents, and the observable phenomena typically are the outcomes of localized interactions among the agents. Thus, agent-based simulations tend to be well-suited for simulating ecological systems (Wilensky & Resnick, 1999).

Following our earlier work (Joyner et al., 2011), VERA uses an artificial intelligence compiler to automatically translate the patterns in the conceptual models into the primitives of agent-based simulation in NetLogo (Joyner et al., 2014; Wilensky & Resnick, 1999). Figure 4 illustrates the time-series graph of the NetLogo simulation results for the conceptual model in Figure 2C. The running of the simulation enables the learner to observe the evolution of the simulated parameter values over time and iterate through the model-simulate-refine loops. In this way, VERA integrates both qualitative reasoning in the conceptual model and quantitative reasoning in the simulation on one hand, and explanatory reasoning and predictive reasoning on the other.

The mechanisms for translating the semantics of CMP conceptual models into the semantics of the Netlogo agent-based simulations are composed of four steps. First, the compiler extracts the conceptual model from the visual layout via the mxGraph java library. Then it is represented as VeraWeb Domain Model to handle high-level simulation concepts. Next, Domain Simulation Builder decomposes high-level domain-specific behaviors and logic into domain-independent simulation operations. Finally, the resulting simulation abstract syntax tree (AST) and abstract semantic graph (ASG) compile abstract simulation into target simulation native constructs such as NetLogo.

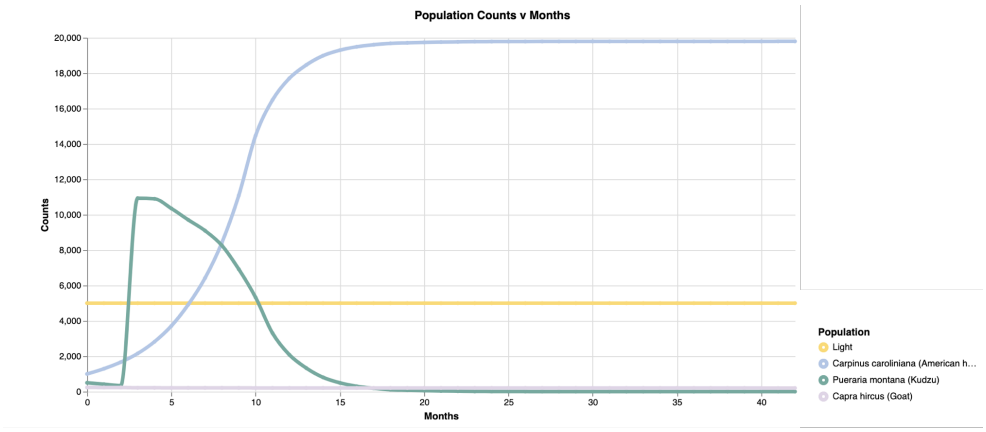

Figure 4. The Simulation Output Graph – x axis: Time (months); y axis: Population.

4. Contextualizing Domain Knowledge

In this section, we describe how VERA retrieves and contextualizes large-scale domain knowledge from Encyclopedia of Life (EOL) for suggesting initial parameter values. Contextualized access to domain knowledge refers to making general-purpose knowledge specific to a problem-solving context. In VERA, the LookupEOL feature makes trait data in EOL specific to agent-based simulation parameter values.

4.1 Domain Knowledge Source, Encyclopedia of Life (EOL)

Smithsonian's Encyclopedia of Life (EOL) is the world's largest aggregated and curated database of species data with almost two million species and eleven million trait data records in the biological domain (eol.org) (Parr et al., 2016). Figure 5(3) shows a screenshot of trait data of the Red-Tailed Hawk in EOL. EOL aggregates trait records from many sources with multiple records from different studies under a variety of conditions. To automate large-scale queries, EOL API services provide on-demand JSON output that has a summarized view of knowledge about EOL taxa, ecological interactions, and organism attributes.

4.2 Contextualizing EOL Data

VERA provides the contextualized biotic trait data from EOL via the "Lookup EOL" feature for suggesting initial parameter values. Figure 5 shows the five steps of adding a biotic component via the Lookup EOL feature. (1) The learner queries a species name (either a scientific or common name) in VERA. (2) Then the system returns a list of species names that match the input via the EOL Search API. (3) Then, the learner selects one species from the list, and the system calls EOL TraitBank API to retrieve specific traits of the species. (4) The VERA's inference engine uses the retrieved traits (5) to preset the simulation parameters.

Table 1. The Selected Properties of Biotic Components and their Trait Data for Conceptual Modeling.

Property	Description	EOL Trait Data
Lifespan	Average lifespan of organisms in this population in months.	"life span" or "total life span"
Reproductive Maturity	Age when organisms in this population are able to begin reproduction in months.	"age at first birth","age at first reproduction","age at maturity","onset of fertility","egg-laying begins"
Reproductive Interval	Frequency with which organisms in this population are able to reproduce in months.	"inter-birth interval"
Offspring Count	Average number of offspring per spawning individual for a reproduction cycle.	"offspring" "litters per year"
Body Mass	Average body mass per organism.	"body mass" If it is not available, attempt to estimate it based on taxonomic ancestry traits ("body length")
Carbon Biomass	Average carbon biomass in an individual organism.	"carbon biomass" Attempt to estimate it based on taxonomic ancestry traits ("plant height", "body mass")
Respiratory Rate	Average basal metabolic rate, measured as respiration (loss) of carbon biomass.	"respiratory rate" If it is not available, attempt to estimate it based on taxonomic ancestry traits ("basal metabolic rate", "body mass")
Photosynthesis Rate	Average addition of carbon biomass from photosynthesis for a square meter of density-based populations (kg/month).	"photosynthetic rate" If it is not available, attempt to estimate it based on taxonomic ancestry traits ("net carbon fixation rate")
Assimilation Efficiency	Efficiency of assimilating carbon biomass via consumption (0.0 - 1.0).	Attempt to estimate it based on taxonomic ancestry traits.

If the trait data is available in EOL, values are directly retrieved from the EOL trait data. For example, in Table 1, lifespan is retrieved from the existing EOL traits such as "life span" or "total life span" (In case of redundancy, the average value is used). Reproductive maturity is retrieved from traits such as "age at first birth","age at first reproduction","age at maturity","onset of fertility","egg laying begins." However, if the necessary trait data is not available in EOL, VERA attempts to estimate the value based on taxonomic ancestor and/or other available traits. For example, if carbon biomass is not available in EOL, carbon biomass is calculated as $C_{reptilia}(.122) \times bodymass$ in case its ancestor is Reptilia in freshwater; $C_{mammalia}(.16) \times bodymass$ in case of Mammalia. If its taxonomic ancestry traits are not available, VERA provides default values that achieve ecological plausibility (e.g., $C_{default}(.1) \times bodymass$).

Table 2. The Taxonomy of Interactions among Components.

Relationship	Property	Description	GloBI Interactions
X Consumes Y	Consumption rate, interaction probability	When X interacts with Y, it will partially or wholly consume Y, with carbon transfer to X from Y.	"eat", "get eaten by", "preys on", "get preyed on by"
X Destroys Y	Destruction rate, interaction probability	When X interacts with Y, it partially or wholly destroys a simulation entity of type Y with no carbon transfer to X.	"kill", "is killed by", "parasitize", "get parasitized by", "get infected by"
X Produces Y	Production rate	X will produce Y with some stochastic timing and amount.	"visits flowers of", "flowers visited by", "pollinate", "get pollinated by", "spread", "get spread by."
X Affects Y	Growth rate, interaction probability	This is a generic growth modifier that allows for growth rates (negative or positive) to modify Y when X interacts with it, where none of the above relationships apply.	"interacts with" (+, -), "related to" (+, -), "parasitize" (-), "get parasitized by" (-), "hosts", "get hosted by."
X Becomes Y on Death	Percent body mass	When X expires, it produces Y.	-

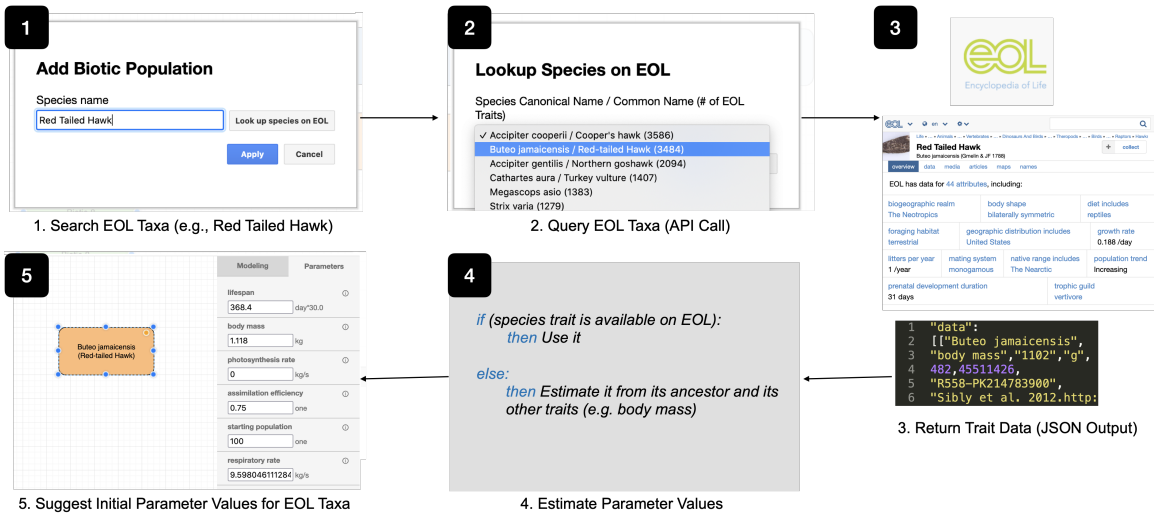

Figure 5. Using EOL TraitBank Data to Set Up Simulation Values.

5. Evaluation

To test our hypothesis about contextualized access to large-scale domain knowledge via the LookupEOL feature resulting in construction of deeper models, we analyzed 326 models generated by 215

learners from 2018 to 2021. In this analysis, the learning goals, as well as the demographics of the learners or even their precise geographical location are unknown; only the modeling behaviors (log data) and outcomes (final work products) are observable.

5.1 Data and Measures

Our unit of analysis is the model, and our dataset consists of models (N=326) built by users (N=215). Drawing from the database of all learners, we used three inclusion criteria. First, to overcome the limitation of non-experimental observational data, we considered only learners who have made at least two models made with and without the LookupEOL feature. All the learners in VERA had access to the LookupEOL feature, but they did not always use it for all of their models. We take advantage of this fact and we selected learners who have both models made using LookupEOL and models made without LookupEOL and compared the model outcomes with and without the feature. Although access to the feature in our study was not assigned randomly, our study eliminates the possibility that learners who used EOL are simply more motivated and more engaged. These inclusion criteria resulted in a dataset that included 26 users.

Second, once we had identified these learners, we collected data on 79 *de novo* (original) models these users had created. We define *de novo* models as those that were created from a blank-slate editor and were not copied from an existing exemplar model. We focused on these models as we wanted to examine the effects of using the LookupEOL feature in model construction (e.g., adding components), but the copied models already include some components with parameter values already assigned. Lastly, we only considered models that include more than one component because we felt that a model with only one component might not have indicated meaningful engagement or learning. Consequently, the dataset used in our analysis included a total of 79 models created by 26 learners.

For each model, we included a binary variable that indicated whether the model was created using the LookupEOL feature or not (*Used EOL*). If the model had included more than one component that is created by the LookupEOL feature, this variable was coded 1, otherwise 0. We also collected a measure for capturing the experience of the learners at the time they created the model in question (*Model Count*). *Model Count* is the cumulative number of models the learner had created at the time they created the model in question.

Our dependent variables sought to measure the quality of the models. Measuring model quality quantitatively is challenging and exacerbated by the fact that there is a large variety in the phenomena being modeled as well as in the goals and behaviors of the learners with no standardized forms of evaluation. Building on previous work, we operationalized the quality of the models in terms of model complexity and model specificity (An et al., 2020; Liem et al., 2013; Pruett & Weigel, 2020; Scaffidi & Chambers, 2012). Model complexity measures the total number of interacting components in a model (e.g., the total number of nodes and edges) (An et al., 2020; Pruett & Weigel, 2020; Scaffidi & Chambers, 2012). Model specificity measures the proportion of components whose parameter values are tailored from the default parameter values.

The first dependent measure is the total complexity including components and relationships (*Total Count*). Because the LookupEOL feature is specifically associated with adding a "Biotic" component, we also considered the number of biotic components (*Biotic Count*) as our dependent

measure. Lastly, *Changed Parameter*?$_{biocom}$ is a binary variable that indicated whether the biotic component includes at least one biotic parameter value (e.g., lifespan, body mass) that is different from the default value. *Changed Parameter*?$_{model}$ is the proportion of *Changed Parameter*?$_{biocom}$ in a model. For example, if a model included two biotic components and one of them contained a parameter value different from the default value while the other biotic component did not, *Changed Parameter*?$_{biocom}$ is coded 1 and 0, respectively, and *Changed Parameter*?$_{model}$ is coded as 0.5.

5.2 Analytic Method

In our analysis, the independent variable is (*Used EOL*). Although we can run a simple correlation between *Used EOL* and model quality measures (*Total Count*, *Biotic Count*, *Changed Parameter*?$_{model}$), evidence of this relationship does not necessarily imply causation. For example, there is a possibility that users will make deeper models as they gain more experience. We controlled for the experience of the learner in terms of the number of previously created models (*Model Count*).

As our dependent variables are numeric, we used Generalized Linear Models (GLM) because they support both categorical and continuous data with various distributions (McCullagh & Nelder, 2019). For our independent variable, *EOLDummyCode* was created and has the numerical value of 1 for *Used EOL*="Yes" and 0 otherwise. The formula for our model that estimates the causal effect of a learner using the EOL on *Biotic Count* is:

$$\widehat{BioticCount} = \beta_0 + \beta_1 \text{EOLDummyCode} + \beta_2 \text{ModelCount} \quad (1)$$

where the predicted value of $\widehat{BioticCount}$ equals a constant or intercept β_0 plus weights or slopes (β_1 and β_2) times two predictor variables (*EOLDummyCode* and *Model Count*). The result of fitting this model was our estimate. The approach described above was the same for other dependent variables (*Total Count* and *Changed Parameter*$_{model}$).

5.3 Results

As a first step, we simply compared all models created by learners using EOL to all the models created without EOL. The results of a 2-sample test for equality of *Biotic Count* on these data. For all dependent variables, we find that there is a statistically significant difference between the EOL model and non-EOL models (Biotic Count: $t = 3.8054$, $p < 0.0005$; Total Count: $t = 3.3624$, $p < 0.005$; Changed Parameter?$_{model}$: $t = 4.3961$, $p < 0.0001$). As explained above, because this difference might be a function of the increase in the experience of learners, we proceeded to fit our regression models.

Table 3 gives the results of the GLM that predicts *Biotic Count, Total Count, Changed Parameter*$_{model}$ from *Model Count* and the dummy code for *Used EOL*. With the (0, 1) coding scheme of *Used EOL* (0: Without the LookupEOL feature, 1: With the LookupEOL feature), the coefficient represents the difference between each level mean. For *Biotic Count*, the parameter estimate for *Used EOL* is positive ($\beta = 0.7581, p < .05$), suggesting that the difference between *Biotic Count* of the models made using EOL and without using EOL by two users with the same level of experience (*Model Count*) is 0.7581. In other words, the model made using EOL included 0.7581 more biotic components on average. The estimate is relatively small, but *Total Count* differed by over 3 ($\beta = 3.6464, p = .023$).

Table 3. Results of the GLM predicting Biotic Count, Total, Changed Parameter$_{model}$ from Model Count and EOLDummyCode.

Variable	Estimate	Std.Err	z	p
Used EOL	0.7581	.295	2.568	.010
Model Count	0.0092	0.019	0.491	0.624
const	2.3145	0.164	14.1257	.000
No. Observations		63		
Log-Likelihood		-87.047		
Deviance		58.478		
Pearson chi2		58.5		

(a) Biotic Count

Variable	Estimate	Std.Err	z	p
Used EOL	3.6464	1.609	2.266	.023
Model Count	0.0289	0.102	0.283	0.777
const	5.8961	0.893	6.601	.00
No. Observations		63		
Log-Likelihood		-193.88		
Deviance		1737.6		
Pearson chi2		1.74e+03		

(b) Total Count

Variable	Estimate	Std.Err	z	p
Used EOL	0.4146	0.105	3.935	.000
Model Count	0.0138	0.007	2.064	0.039
const	0.4498	0.058	7.690	.000
No. Observations		63		
Log-Likelihood		-22.147		
Deviance		7.4508		
Pearson chi2		7.45		

(c) Changed Parameter$_{model}$

This means that the model using EOL not only had more biotic components but also had more abiotic components and associated relationships. For *Changed Parameter*$_{model}$, we also obtained a positive estimate ($\beta = 0.4146, p < .0001$). We find that the proportion of parameterized biotic components in the model built with EOL was 0.4773 higher than that of the model built without using EOL.

To aid in interpreting these results, Figure 6 shows the model-predicted *Biotic Count, Total Count, Changed Parameter*$_{model}$ for a series of prototypical models. The first two panels show that a prototypical model built using the LookupEOL feature by a user who has shared 3 models (the median value of *Model Count*) had 3.14 biotic components, a model complexity of 9.80, while the number of biotic components and model complexity for a model by an otherwise equivalent user without using EOL was 2.02 and 4.80. The third panel shows that 91.39% (0.91) of the biotic components in the prototypical model built using the LookupEOL feature had personalized param-

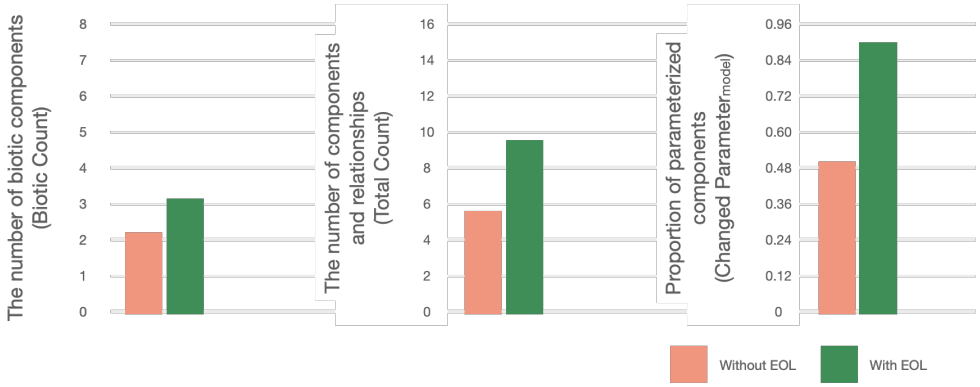

Figure 6. Model-predicted *Biotic Count, Total Count, Changed Parameter*$_{model}$. The values of *Model* Count are held at 3 throughout.

eters rather than default parameters, while 43.66% (0.43) of the biotic components had personalized parameters in the model built without using the LookupEOL feature.

Table 4 shows the names of the biotic components extracted from the models. The component names were classified based on biological taxonomic rank (e.g. species, genus, family, etc.) to indicate taxonomic specificity. We compared the component names of the models that used the LookupEOL feature and the models that didn't use the LookupEOL feature to see which was more specifically expressed. *Species* is more specific than *genus*; *Genus* is more specific than *family* (Intermediate rankings are not shown). As shown in Table 4, the EOL biotic components used *Species* the most (EOL= 40) in a narrower sense (e.g., *Paracentrotus lividus*), whereas the non-EOL biotic components used *Genus* or *Family* (Non-EOL= 23) the most in a more general sense (e.g., Sheep). In addition, the non-EOL components were used to represent non-biological entities (Non-EOL= 13) such as GPUs, Renewable energy initiatives, etc. The default names that appear when the component name is not specified were also more common in the Non-EOL components (Non-EOL= 4; EOL= 1).

6. Discussion

We described the process of contextualizing access to large-scale domain knowledge and evaluated the effect of using contextualized domain knowledge on model quality. We quantitatively measured three aspects of model quality (*Biotic Count, Total Count, Changed Parameter*$_{model}$). We found the models using the LookupEOL feature had higher complexity (Total Count), more numerous biotic components (Biotic Count), and a higher proportion of tailored parameter values (Changed Parameter$_{model}$) than those without using the LookupEOL feature. Our qualitative analysis comparing biotic names between the models that used and did not use the LookupEOL features additionally revealed that the models using the EOL feature used more specific names for the biotic components, including the components that did not use the LookupEOL feature.

Table 4. Biotic Component Names Extracted from Models and Counts.

Classification	Examples	Count	
		Non-EOL (N=60)	EOL (N=47)
Species	Paracentrotus lividus (common sea urchin), Enhydra lutris (Sea Otter), Homo sapiens (Human), Balaenoptera musculus (Blue Whale), Euphausia superba (Antarctic Krill), Quercus virginiana (live oak), ...	20	40
Genus or Family	Sheep, Duck, Locust, Grass, Roses, Flowers, ...	23	3
Others	GPUs, Renewable energy initiatives, Reddits, Casual Bicycle Riders, Recreational Hunters, Food Source, ...	13	3
Default Names	Biotic-1, Biotic-2, ...	4	1

From these results, we derived two findings. First, the contextualized domain knowledge enabled learners to represent the ecological systems "in more detail" with more components and with more specific names. The models that used the EOL feature expressed ecological phenomena with more interacting components ($\beta = 3.6464$, $p < 0.05$) (see Table 3). This includes adding more predators, prey, competitors, and/or other abiotic components associated with the phenomenon. This may be because the lookupEOL feature makes it easier to add biotic components than the manual method; the lookupEOL feature may have motivated the learners to add more components to their models. In addition to using a greater number of components to express a phenomenon, the component itself used more detailed species names (85.10% of the components used the species names, see Table 4) rather than more generic names such as genus or family. When the LookupEOL feature was not used, more general names were used based on the learners' knowledge (e.g., plant). With the LookupEOL feature, more detailed representations were possible because species could be searched and selected (e.g., *Pueraria* (Kudzu)).

Second, the contextualized domain knowledge enabled learners to build models with more "tailored" parameter values. One of the core aims of constructionist learning is to empower the learner to use knowledge in personally meaningful ways (Papert, 2020). Using the LookupEOL feature, learners were able to find a species matching their interests by searching for it and using it directly for their models rather than just using the species given by the system via examples or those commonly known from the textbook. One of the VERA's goals is to offer our non-expert learners the ability to model and explore various phenomena that have real personal importance to the learner. Our results support this goal as the parameters in the models that used the LookupEOL feature were more tailored to the learners ($\beta = 0.4146$, $p < 0.0001$) as opposed to those universal to all learners irrespective of their interests. This suggests that self-directed learners, such as citizen scientists, can explore the ecological systems in their region by modeling local species, which is more meaningful to the learners.

7. Conclusion

Modeling requires domain knowledge. However, providing access to contextualized knowledge requires intricate knowledge engineering: developing a detailed, seamless process for model construction, evaluation, and revision; designing a principled ontology for capturing the contents of the model; conducting a precise analysis of all problem-solving contexts that arise in the modeling process; mapping the problem-solving contexts to specific chunks of knowledge in a large-scale knowledge source; drawing inferences from the available knowledge to meet the requirements of the problem-solving contexts, and so on. VERA is an interactive modeling environment for constructing conceptual models of ecological phenomena that provides access to large-scale biodiversity knowledge in Encyclopedia of Life. In this paper, we described the knowledge engineering required in integrating VERA with EOL.

We also described a study on the use of VERA by self-directed learners. We hypothesized that contextualized access to domain knowledge would help learners build deeper models. In this study, contextualized access to domain knowledge was operationalized as the lookupEOL feature. We compared the models with and without the use of the LookupEOL feature. Since the learners in the study were wholly self-directed, we have data only on their modeling behaviors and the models they constructed, not on their learning goals, assessments, or outcomes. We found that the use of the LookupEOL feature was associated with three aspects of model quality. In conclusion, contextualizing access to domain knowledge through VERA might have helped self-directed learners to represent the ecological systems in more detail with more components and with more specific names and tailor the parameter values of the ecological models to their personal preferences.

This study is a correlational study that does not show a direct causal link between the use of contextualized access to domain knowledge and the model quality. Thus, future work should conduct a controlled experiment by manipulating an independent variable (the use of the LookupEOL feature) as well as conduct an interview with the learners to gain insights on how the LookupEOL helped them build models. Further, future work should explore contextualizing access to other types of domain knowledge, such as feeding relationships, habitat environments, and so on.

Acknowledgements

This research was supported by US NSF BigData grant #1636848 and US NSF National AI Institutes grant #2112532.

References

An, S., Bates, R., Hammock, J., Rugaber, S., Weigel, E., & Goel, A. (2020). Scientific modeling using large scale knowledge. *International Conference on Artificial Intelligence in Education* (pp. 20–24). Springer.

An, S., Broniec, W., Rugaber, S., Weigel, E., Hammock, J., & Goel, A. (2021a). Recognizing novice learner's modeling behaviors. *International Conference on Intelligent Tutoring Systems* (pp. 189–200). Springer.

An, S., Rugaber, S., Hammock, J., & Goel, A. (2022). Understanding self-directed learning with sequential pattern mining. *International Conference on Artificial Intelligence in Education*.

Springer.

An, S., Rugaber, S., Weigel, E., & Goel, A. (2021b). Cognitive strategies for parameter estimation in model exploration. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Bradshaw, G. F., Langley, P. W., & Simon, H. A. (1983). Studying scientific discovery by computer simulation. *Science*, *222*, 971–975.

Bridewell, W., Langley, P., Todorovski, L., & Džeroski, S. (2008). Inductive process modeling. *Machine learning*, *71*, 1–32.

Broniec, W., An, S., Rugaber, S., & Goel, A. K. (2021). Guiding parameter estimation of agent-based modeling through knowledge-based function approximation. *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*.

Chandrasekaran, B. (1994). Functional representation: A brief historical perspective. *Applied Artificial Intelligence an International Journal*, *8*, 173–197.

Clement, J. J. (2008). *Creative model construction in scientists and students*. Springer.

Darden, L. (1998). *Anomaly-driven theory redesign: computational philosophy of science experiments*. na.

De Jong, T., & Van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of educational research*, *68*, 179–201.

Dickinson, J. L., Zuckerberg, B., & Bonter, D. N. (2010). Citizen science as an ecological research tool: challenges and benefits. *Annual review of ecology, evolution, and systematics*, (pp. 149–172).

Dunbar, K., & Fugelsang, J. (2005). Scientific thinking and reasoning. *The Cambridge handbook of thinking and reasoning*, (pp. 705–725).

Edelson, D. C. (1998). Realising authentic science learning through the adaptation of scientific practice. *International handbook of science education*, *1*, 317–331.

Goel, A. K. (2013). A 30-year case study and 15 principles: implications of an artificial intelligence methodology for functional modeling. *AI EDAM*, *27*, 203–215.

Goel, A. K., Rugaber, S., Joyner, D. A., Vattam, S. S., Hmelo-Silver, C. E., Jordan, R., Sinha, S., Honwad, S., & Eberbach, C. (2013). Learning functional models of aquaria: The act project on ecosystem learning in middle school science. In *International handbook of metacognition and learning technologies*, 545–559. Springer.

Goel, A. K., Rugaber, S., & Vattam, S. (2009). Structure, behavior, and function of complex systems: The structure, behavior, and function modeling language. *Ai Edam*, *23*, 23–35.

Goel, A. K., Gómez de Silva Garza, A., Grué, N., Murdock, J. W., Recker, M. M., & Govindaraj, T. (1996). Towards design learning environments—i: Exploring how devices work. *International conference on intelligent tutoring systems* (pp. 493–501). Springer.

Griffith, T., Nersessian, N., & Goel, A. (2000). Function-follows-form: generative modeling in scientific reasoning. *Proc. 22nd Cognitive Science Conference* (pp. 196–201).

Griffith, T. W., Nersessian, N. J., & Goel, A. (2019). The role of generic models in conceptual change1. *Proceedings of the eighteenth annual conference of the cognitive science society* (pp. 312–317). Routledge.

Hogan, K., & Thomas, D. (2001). Cognitive comparisons of students' systems modeling in ecology. *Journal of Science Education and Technology*, *10*, 319–345.

Howe, J., et al. (2006). The rise of crowdsourcing. *Wired magazine*, *14*, 1–4.

van Joolingen, W. R., de Jong, T., Lazonder, A. W., Savelsbergh, E. R., & Manlove, S. (2005). Co-lab: research and development of an online learning environment for collaborative scientific discovery learning. *Computers in human behavior*, *21*, 671–688.

Joyner, D. A., & Goel, A. K. (2015). Organizing metacognitive tutoring around functional roles of teachers. *CogSci*.

Joyner, D. A., Goel, A. K., & Papin, N. M. (2014). Mila–s: generation of agent-based simulations from conceptual models of complex systems. *Proceedings of the 19th international conference on intelligent user interfaces* (pp. 289–298).

Joyner, D. A., Goel, A. K., Rugaber, S., Hmelo-Silver, C., & Jordan, R. (2011). Evolution of an integrated technology for supporting learning about complex systems. *2011 IEEE 11th International Conference on Advanced Learning Technologies* (pp. 257–259). IEEE.

Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes*. MIT press.

Leelawong, K., & Biswas, G. (2008). Designing learning by teaching agents: The betty's brain system. *International Journal of Artificial Intelligence in Education*, *18*, 181–208.

Liem, J., et al. (2013). *Supporting conceptual modelling of dynamic systems: A knowledge engineering perspective on qualitative reasoning*. Universiteit van Amsterdam [Host].

Lippe, M., et al. (2019). Using agent-based modelling to simulate social-ecological systems across scales. *GeoInformatica*, *23*, 269–298.

McCullagh, P., & Nelder, J. A. (2019). *Generalized linear models*. Routledge.

Metcalf, S. J., Krajcik, J., & Soloway, E. (2000). Model-it: A design retrospective. *Innovations in science and mathematics education*, (pp. 77–115).

Nersessian, N. J. (2010). *Creating scientific concepts*. MIT press.

Odum, H. T., & Odum, E. C. (2000). *Modeling for all scales: an introduction to system simulation*. Elsevier.

Papert, S. A. (2020). *Mindstorms: Children, computers, and powerful ideas*. Basic books.

Parr, C. S., Schulz, K. S., Hammock, J., Wilson, N., Leary, P., Rice, J., & Corrigan Jr, R. J. (2016). Traitbank: Practical semantics for organism attribute data. *Semantic Web*, *7*, 577–588.

Poelen, J. H., Simons, J. D., & Mungall, C. J. (2014). Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecological Informatics*, *24*, 148–159.

Pruett, J. L., & Weigel, E. G. (2020). Concept map assessment reveals short-term community-engaged fieldwork enhances sustainability knowledge. *CBE—Life Sciences Education*, *19*, ar38.

Railsback, S. F., & Grimm, V. (2019). *Agent-based and individual-based modeling: a practical introduction*. Princeton university press.

Salles, P., & Bredeweg, B. (2003). Qualitative reasoning about population and community ecology. *AI Magazine*, *24*, 77–77.

Scaffidi, C., & Chambers, C. (2012). Skill progression demonstrated by users in the scratch animation environment. *International Journal of Human-Computer Interaction*, *28*, 383–398.

Schwarz, C. V., & White, B. Y. (2005). Metamodeling knowledge: Developing students' understanding of scientific modeling. *Cognition and instruction*, *23*, 165–205.

Sins, P. H., Savelsbergh, E. R., & van Joolingen, W. R. (2005). The difficult process of scientific modelling: An analysis of novices' reasoning during computer-based modelling. *International Journal of Science Education*, *27*, 1695–1721.

VanLehn, K. (2013). Model construction as a learning activity: A design space and review. *Interactive Learning Environments*, *21*, 371–413.

Vattam, S., Goel, A. K., Rugaber, S., Hmelo-Silver, C., & Jordan, R. (2009). From conceptual models to agent-based simulations: Why and how. *Artificial Intelligence in Education* (pp. 593–595). IOS Press.

White, B. Y., & Frederiksen, J. R. (1990). Causal model progressions as a foundation for intelligent learning environments. *Artificial intelligence*, *42*, 99–157.

Wilensky, U., & Resnick, M. (1999). Thinking in levels: A dynamic systems approach to making sense of the world. *Journal of Science Education and technology*, *8*, 3–19.