Measuring Systems Thinking Using Stealth Assessment

Ing. Andrea Ramirez-Salgado, University of Florida

Andrea is a doctoral student at the University of Florida specializing in Educational Technology within the Curriculum and Instruction program. She has a master's degree in Education and ICT and a bachelor's degree in Software Systems Engineering. Andrea has been teaching undergraduate and graduate courses for the past thirteen years covering topics such as algorithms, process engineering, instructional design, and applications of technology in education. Her research interests include understanding the implications of hands-on software and hardware learning approaches for developing engineering identity and fostering engineering persistence in students. Additionally, she is part of a research team funded by the National Science Foundation that aims to design and develop gamified activities to teach hardware principles at high school and undergraduate levels with a focus on equity principles.

Eric Wright, University of Florida

PhD student in quantitative research methodology in education at the University of Florida, former eLearning instructional designer and developer. Primary interest is in critical theory approaches to quantitative research.

Measuring Systems Thinking Using Stealth Assessment

Abstract

As technology advances and databases grow larger, people require high-level skills to process information effectively [1]. To address complex problems while maintaining a comprehensive view of the situation, one valuable competency is Systems Thinking (ST). ST is a systematic approach that allows individuals to navigate different levels of a system without losing sight of the big picture [2]. For instance, software development involves numerous components, including user needs, environments, change management, performance metrics, budget, workflows, and more. A systems thinker must understand the causal relationships between these components to provide a comprehensive and optimal solution. They use mental models to identify interdependencies between inputs, processes, transactions, automation needs, and desired outputs. Successful systems thinkers offer solutions that address the root causes of problems rather than simply treating symptoms [2]. ST is a vital skill in engineering, but it also applies to environmental and ecological issues, socio-economic problems, medical cases, nursing, and geography education [2]–[6].

Assessing ST typically involves using self-reported measures of systems-thinker characteristics [7], behavior-based assessments [8], and affective domain learning [9]. While these traditional methods are quick and easy, they do not provide ongoing, formative assessments that can guide teaching and learning [6]. To address this gap, new approaches like stealth assessment are emerging. Stealth assessment involves diagnosing ST performance based on evidence from students' interactions with multimedia and using Evidence-Centered Design (ECD) frameworks [10] to create optimal ST achievement conditions. This work-in-progress study proposes the use of a video game designed under ECD and stealth assessment principles to teach ST through simulations and problem-solving strategies.

A further validation study aims to evaluate the game's effectiveness in measuring ST achievement in real-life situations beyond Software Engineering. The study will focus on middle school students and will consist of two phases. In the first phase, participants will play the game individually and describe their thought processes to identify any necessary changes to the assessment. The second phase will involve enrolling 1,000 seventh-grade students in selected schools to play the game during several class periods. The study will collect and store game logs, demographic information, and performance-based measures to analyze the effectiveness of the game.

Introduction

Systems Thinking is the ability to understand how components within a system are connected and interact with one another. It involves recognizing systems and their relationships, understanding feedback and system behavior, creating models to simulate systems, and applying those models to manage change dynamics. Evidence-Centered Design (ECD) is a framework used to assess learners' competency performance and determine their instructional needs. One approach to automated scoring within the ECD framework is stealth assessment, which involves embedding assessments within a computer-based game that provides a realistic context for system thinkers to solve problems. In this framework, educators can utilize ongoing, evidence-

based assessments to adjust instruction to support students' growth in Systems Thinking. These two concepts are fundamental theoretical underpinnings of our study and are detailed in the following sections.

Systems Thinking

ST has numerous definitions depending on the theorist and context [4], [11]. However, in general, ST is about seeing the larger picture of a system that is informed by its components, connections, and processes but is also more than the sum of those parts [11].

Stave and Hopper [11] sought to identify the key components of ST by conducting interviews with systems educators and reviewing literature from the broader field of systems dynamics. They found that ST definitions typically involve five to seven components. These components include (1) recognizing systems, their components, their connections, and how those components and connections come together to create something more; (2) recognizing relationships between components in terms of cause, effect, and feedback loops where cause and effect are not necessarily unidirectional; (3) understanding the relationship between feedback and system behavior as a dynamic feature of systems themselves; (4) differentiating between different types of variables and relationships that form a system in terms of how some classes of variables/relationships may behave differently from others; (5) using conceptual models informed by 1–4 to explain systems and specific behaviors of systems; (6) creating models for simulating systems, particularly mathematically, although some consider this to be outside of the core concept of ST; and (7) applying those simulation models for purposes such as learning more about how a system operates, testing hypotheses, and developing plans/policies for creating change in a system. Dugan et al. [4] conducted a systematic review of ST assessments in the engineering field and arrived at similar conclusions as [11]. However, they also noted additional ST components, such as considering stakeholders and being able to describe systems at multiple levels.

In addition to the components previously mentioned, we consider the following to be critical components of ST proficiency: (1) the ability to identify unfamiliar situations [12] (2) the ability to avoid becoming overly focused on details [13]; and (3) the ability to draw connections and identify similarities between different systems to apply lessons learned from one system to better understand another [13]. These ST components can work together to provide systems thinkers with a deeper understanding of a system at multiple levels, particularly the system as a whole. This ability to understand complex systems holistically can be immensely valuable in solving intricate problems across a wide range of contexts.

Stealth Assessment and Evidence-Centered Design (ECD)

ECD is a framework that helps determine what learners know, can do, and believe to make decisions about their instructional needs [6]. The ECD framework utilizes theoretical models to make inferences about competency performance, which include the Competency Model (CM), Evidence Model (EM), Task Model (TM), and Assembly Model (AM) [6]. The CM is a conceptual framework created from literature and experts' insights, which provides claims or statements about the students' competency based on their assessment performance [6]. The EM is used to determine how observable student behaviors and actions indicate their competency

performance levels [6]. The TM provides specifications for the presentation materials, which allow for the accumulation of evidence and the creation of log data that will inform the scoring process [6]. Lastly, the AM outlines the order of tasks and levels and the beginning and end of the game [6].

One approach to obtaining an automated scoring process within the ECD framework is stealth assessment [6]. The automation and machine-based reasoning feature of stealth assessment facilitate inferences about achievement in a non-testing or invisible environment under validity and consistency assessment considerations. In addition, this study considers stealth assessment under a computer-based game that provides the player/student with a more realistic context that aligns with the complexities that system thinkers face when solving problems. A game is a good condition for the characteristics of this study because as "the player interacts with the game, stealth assessment (which is embedded deeply within the game) analyses patterns of actions using the game's log file to estimate the player's competencies and make claims about them" [10].

Research Question

The main objective of our study is to evaluate the capabilities of our game in stealthy assessing the ST skill proficiency of middle school players at various points during gameplay. Moreover, we plan to leverage the insights gained from the game to deliver tailored ST instruction in non-Software Engineering courses. Thus, our research question is: How reliable and valid is our video game designed under stealth assessment and ECD principles in determining students' proficiency in ST skills, so that we can provide personalized educational support to learners?

Methods

This section will provide an overview of the proposed game environment, including the three models from the ECD framework used to assess ST stealthily and formatively. Furthermore, we provide details of a future study aiming to validate ST's stealth assessment within the game.

Game

To take a holistic approach to problem-solving using ST, it's necessary to analyze complex situations involving stakeholder interests, ecological variables, systems boundaries, and internal systems feedback processes. One key factor in this is understanding the role of feedback processes, which can either reinforce or balance the behavior of a system. However, it's important to note that reinforcing feedback loops are not always beneficial for a system, as they can lead to an overabundance of certain elements [14]. For example, in a prey-predator relationship, a stronger predator population can decrease the prey population, ultimately unbalancing the system. In designing a game to assess ST, it's essential to take into account these variables and their potential impact on the system being studied.

Qin et al. [15] argued that strategy games are suitable when the player needs to comprehend the situation before managing resources and making plans. Therefore, we propose that the most appropriate game genre for assessing ST is a blend of a strategy game and a serious simulation of a system or environment. By combining these two genres, our game will offer interactivity and immersion while presenting various difficulty levels [15]. The next section will provide a

summary of the game narrative, which considers different aspects of ST situations and recommended practices for computer-based games.

The goal of our game is to teach ST skills to the player, and as such, results from our stealth assessment can be used to support the teaching and learning taking place. To this end, the player has the goal of maintaining/restoring ecological balance in park systems. The player serves as an ecological consultant working for park systems, both documenting the ecological and park systems and proposing solutions to related problems. The player will start with simple documentation at smaller, less complex parks and gradually move on to problem identification and solving at larger national parks with more complex relationships. They will be able to move around the park, observe the system, and talk with experts/stakeholders to obtain more information. Using the information they collect, they will construct causal reasoning diagrams within the game to record observations about animal and plant species and other factors such as rangers, visitors, nearby industry, and politics that can impact the park ecology. Solutions from earlier levels will be relatable but not identical to later levels. The player will also be able to see the results of proposed changes after they're implemented. The changes are reflected in a game dashboard that displays information about the different variables that affect the systems' behavior and balance.

The player has a monthly budget, an initial inventory of species, and an estimated number of visitors as an entry point. All this information is available in their dashboard. They can use their budget to commission surveys of species or other park features and/or hire rangers for specific tasks such as removing an invasive plant species or creating park programs. The player will receive a bonus when certain goals are met. New goals and problems may be introduced as each park level proceeds, and each level is won when all goals have been completed and the system is in balance for a specified amount of time.

ECD components for Stealth Assessment of ST in the game

Competency Model

As previously stated, the CM refers to the set of skills and abilities that can be evaluated for a student or player after they have completed the game. In line with previous research, we have proposed a competency model, which is depicted in Fig. 1. The model includes three key competencies that stem from the ST node: (1) attending to elements and interconnections [4], [11], (2) modeling the system [6], and (3) using conceptual models [11]. The first competency is a prerequisite for the second, as understanding the system as a whole and in relation to its parts is necessary before a model can be created. Similarly, the second competency is a prerequisite for the third, as a model of the system is required before conceptual models can be used to describe the system. These competencies have further sub-competencies, illustrated in the diagram’s third level. It is crucial to identify feedback processes in order to understand how changes in feedback can produce different behaviors within the system. Furthermore, an understanding of the complexities of feedback and behaviors is necessary before causal reasoning diagrams can be created, which visually explain how positive and negative feedback affect the system.

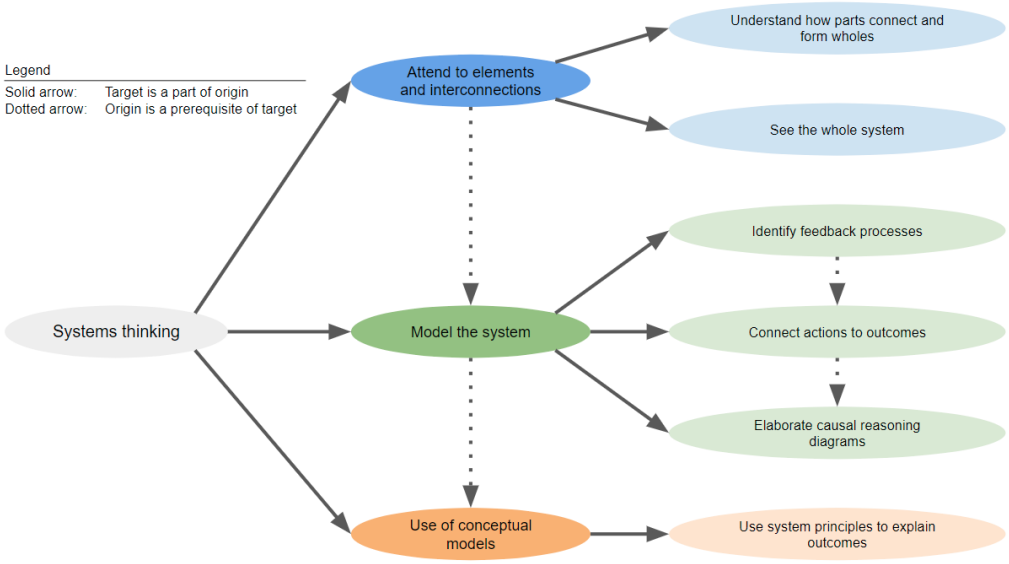


Figure 1. Competency Model

We created tasks that set the standard for individuals approaching a high level of competency for each third-level competency. These tasks are listed below. It's possible to achieve lower competency levels by reducing the complexity of the tasks associated with each third-level competency.

1. See the whole system
   - Can give an understandable name to the system according to its function.
   - Can explain the system as a whole without getting stuck on details.
   - Can explain the system synergy: describe properties of the system that the components alone do not explain [11].
2. Understand how parts connect and form wholes
   - Is able to describe each part, and each part must be described in relation to other parts.
   - Is able to describe the different levels of the system (subsystems).
   - Is able to describe the variables and problems in the system.
   - Is able to identify the flow between variables.
3. Identify feedback processes
   - Is able to describe the system boundaries.
   - Is able to recognize causal chains [11].
   - Is able to describe relationship polarity[11].
   - Is able to describe the polarity of a feedback loop [11].
4. Connect actions to outcomes
   - Is able to create simulation models of the system.

- Is able to describe what will happen to other components when one component changes [11].
- Is able to understand how problems arise due to interactions between components [11].
- Is able to use simulation models to test hypotheses and develop policies.
- Is able to describe the causal structure resulting in an observation [11].

5. Elaborate causal reasoning diagrams
   - Is able to select an appropriate tool to elaborate a causal reasoning diagram.
   - Is able to produce a complex causal loop diagram that describes the system behavior.
   - The causal diagram includes polarity [11].
6. Use system principles to explain outcomes
   - Is able to use a conceptual model to develop solutions [11].
   - Is able to describe why an action will solve a problem [11].
   - Is able to identify similarities between different systems.

Evidence Model

The EM is the set of indicators or observables that reveal the competency proficiency levels of the students/players, and it connects the in-game evidence with the CM variables [16]. The indicators for each of the third levels competencies that are directly-measured via information from the game are listed in Table 1. These primarily come from players' use of the in-game causal diagram with others explicitly related to conversations.

Table 1. Evidence model indicators

Competency	Indicator
Understand how parts connect and form wholes	Add a node to the causal diagram.
	Add a directional relationship (arrow) between two nodes in the causal diagram.
See the whole system	Select with whom to spend time speaking and not speaking.
	Parts of the system (nodes and relationships) asked about.
Identify feedback processes	Identify a direct feedback loop (e.g., A to B to A).
	Identify an indirect feedback loop (e.g., A to B to C to A).
Connect actions to outcomes	Identify possible variables that positively and negatively affect the system in the dashboard.
	Confirm variables that positively and negatively affect the system in the dashboard.
Elaborate causal reasoning diagrams	Nodes and relationships in whole diagram.
	Amount of money collected.

	Number of donations received.
Use system principles to explain outcomes	Select components and relationships from the causal diagram to communicate the system characteristics to donors.

Statistical Model

Our scoring system uses a Bayesian network (BN), which is a probabilistic model that graphically shows the relationship between players' performance and score indicators. We can set up the BN to match the competency and evidence models, making it easy for experts to provide input into the initial model. The BN's distributions are simplified into low, medium, and high categories, which further facilitate expert input. As more players interact with the game, the BN learns from the data, leading to continuous refinement of the model and correction of initial expectations if necessary.

While item response theory can also use Bayesian estimation to make use of priors and updates to those priors, it is more challenging for non-statistical experts to use. On the other hand, summary scores cannot offer this capability at all.

The BN's flexibility is especially useful for our assessment because different game levels provide different amounts of evidence for ST at different ability levels. For instance, the first game level offers limited information and provides more data at lower levels of ST, while later game levels provide more data at higher levels of ST. Determining exact scoring ahead of time using summary scores would be exceedingly difficult and error-prone due to complexity. BNs allow experts to create game-level-specific models that combine into a larger model for estimating ST easily, and if any mistakes are made, the model can be easily updated given evidence from players.

Task Model

The task model facilitates the design of student/player interactions with game elements and specifies how game data will be collected for scoring. The game's presentation materials and product specifications are categorized into six tasks, and a detailed list of features for each task is available in Appendix A. Furthermore, medium-level prototypes have been included in Appendix B, which underwent validation by a group of 15 individuals who provided crucial feedback to enhance the game narrative, multimedia components, and overall engagement.

Presentation Materials

Each level commences with a scenario featuring a minimum of one challenge for the player to tackle, which is communicated through a boss character unique to that level. After successfully resolving a challenge, the same character may assign additional related challenges until all are completed. Additionally, the player will be guided through the interface options and required actions during gameplay for the first level.

The player can select locations to travel to within the park scenario using their map (Task 6). At each location, they can move around in a 3D environment to explore, talk to people and clothed animals in the park by clicking on them, and catalog animals, plants, and park features in the game's causal diagram by clicking on them. At any time, the player can also press the Tab key to open an interface that includes buttons for the park map; saving the game; a phone contact list to call a level-determined set of stakeholders, subject matter experts, and potential donors; the causal diagram feature for viewing and taking notes about components and relationships in the park system; and for accessing the Build Actions (Task 3), Programs (Task 4), and Nature Management (Task 5) menus for purchasing facilities and equipment, running park programs, and adding plants and animals to the park. The specific build actions, programs, plants, and animals will be unique to the scenario and may or may not depend on encountering them in conversation first. They should automatically be added to the causal diagram when the player adds them to the park. Players start with a certain amount of money they can use but can also request more from donors. Actions taken should also affect and update the causal diagram regarding the health of related nodes, whether those nodes and relationships are known or unknown and correctly or incorrectly specified.

The causal diagram in Task 2 is made up of movable nodes/components. Each node should have a name, an image, and a health bar on the right that shows its current health in the system (which changes as the player takes action). Double-clicking on a node allows the player to view educational information about that node that they have obtained through conversations. The player can click and drag empty areas of the diagram to move it around, and zooming in and out can be done using two zoom options in the lower right or by scrolling with the mouse wheel. The diagram also has several buttons for adding nodes that are not currently in the diagram, adding confirmed relationships between nodes, adding hypothesized but unconfirmed relationships between nodes, and removing nodes. When a relationship is added, a circle with a question mark will appear next to the relationship arrow, which the player can click to choose a plus or minus sign to indicate the effect of the relationship. The plus or minus icons can be clicked again to change the direction of the effect.

During Task 1, the player will engage in conversations where they will see what the other person or animal says in text form, followed by numbered conversation options they can choose. The top right of the conversation pop-up will display a phone icon indicating a phone conversation, along with the name and title of the person or animal the player is talking to. In the bottom right of the pop-up, there is a causal diagram icon that the player can click to open the causal diagram. When the player opens the causal diagram during the conversation, they can select and drag one or more nodes to the conversation pop-up to inquire about the node or its relationships. This same interaction can be used to persuade donors to care about a problem and contribute funds to the player. Conversations may also include mentions of system components that should be underlined for information gathering purposes. The player can click on these underlined components to add them to their causal diagram and view them, or if the components have already been added, view them in the diagram without leaving the conversation. This may be helpful for examining a component in the diagram and manually adding any relationships mentioned in the conversation. Additionally, new contacts may be mentioned in the conversation as either a location in the park where they can be found or as a phone contact, and they will be automatically added to the player's contact list.

To accomplish this, 3D-animated park environments are required, complete with people, animals, plants, park amenities, and non-enterable facilities, some of which may be wearing clothing. For the causal diagram nodes, still images are necessary. Audio is also essential, including nature sounds relevant to the environment in which the scenario takes place. Footstep audio should be included as well. The most significant burden is the need for text, which is necessary for branching conversations and conversations that may only be accessible after speaking with others or asking a particular person about something in the causal diagram. When the player has no knowledge about a topic, a standard format can be used to convey this information.

Work Product Specifications

Combining the game log files and the causal diagrams will elicit the learners/players' performance related to the CM we defined for the ST skill. The data segmentation produced by the logs will be based on interactional boundaries [16].The player will interact with persons, animals, and objects while solving the challenges for each level, with no time restrictions. These interactions will allow the player to understand important parts of the system and add nodes and relationships to the causal diagram. The interactions are text-based conversations or multiple-choice questions, and the player creates the causal diagram using the game interface options for this part. All of this will be recorded in the logs. This accumulation of observable outcomes will allow a summary scoring process and, therefore, evidence compilation [17]. The following paragraphs will provide detail about the technical implications for the data that will be recorded.

The conversations the player has with persons, animals, and objects (Task 1) will be analyzed to score their competency in 1) seeing the whole system instead of focusing on parts and 2) explaining the system requirements using conceptual models to explain observations. In particular, to score the seeing the whole system sub facet, we will track the type of person the player decides to contact to gather for help and the time they spend in the conversation in the form of the number of conversation options they choose. For instance, if the challenge is about a prey-predator imbalance, and the player contacts a zoologist, that is a good indicator of a Systems Thinker. On the other side, to score the use of conceptual models, we will analyze the players' clicks and use of the causal diagram when having conversations. Using the diagram to gather information about the park and to argue about essential parts of the system will depict the level of competence of this sub facet.

The causal diagram is a crucial source of scoring data (Task 2). The player will add and remove nodes and relationships in the diagram while interacting with the different elements of the game. When completing a set of challenges within the same level, the game will assess the accuracy of the causal diagram. This assessment will be based on comparisons with a model diagram previously created by ST experts to solve the challenges the player is solving at one particular point. The accuracy of the identified parts, connections between parts, and feedback relationships will help to score the players' level of competence in attending to elements and interconnections and modeling the system.

Finally, the player's money collected at every level is an indicator of their performance. Every time a player interacts with a donor, the game assesses the causal diagram accuracy—compared to the ideal one created by the experts—to decide if the player deserves a raise in their money as

a donation. If the player can solve the four challenges in each level and still have money, this is an indicator of an overall good Systems Thinker. The amount of money available at the end of each level, accompanied by the number of times the player received a donation, is scoring data about their ability to elaborate causal diagrams.

Difficulty Rubric

Rather than assess game difficulty by task or by the individual challenges within a park level, we assess it by park level for the purpose of ordering the levels in the game. This is to make sure levels are of increasing difficulty and because the difficulties of the different task types are not easily placed on the same scale. Furthermore, the challenges in a level are generally designed to build on each other while not necessarily having every task type in a single challenge despite each level containing every task type.

The rubric is available in Appendix C. To calculate level difficulty, add up the numbers in parentheses that correspond to the underlying system in a level. Note that we often refer to minimal causal diagram solutions in the rubric, and these are determined by experts. Different tasks in a level may require different but related minimal solutions, so we also often consider all the minimal solutions combined rather than individual solutions. And the nodes and relationships outside of the minimal solutions are considered nuisances but can increase difficulty because they serve as distractors.

Assembly Model

As mentioned before, the AM provides a structure for the game and orchestrates the CM, EM, and TM to provide the foundations for the assessment [17]. The baseline environment of our game is a park. Each level is a different, more complex park than the previous one. The game includes five of these parks: a small city park, a simple national park, a medium national park, a complex national park, and a Triassic era park. All park levels have the same buttons, interactions, and general functionalities, but the complexity of the challenges changes from one to another. For instance, in a small city park, complexities can be related to lack of budget, vandalism, trash, and dog feces. On the other side, in a large complex national park, complexities can be associated with mining near the park, climate change, overcrowding, etc. The player completes the game when finishing the last of the parks. Players must complete earlier levels to advance to later levels, but they can always return to previous levels if they like. Players can navigate the park in a 3D format at each level, converse with people/animals/things, access a list of contacts, make arguments to potential donors, interact with the causal diagram, and add items and animals to the park infrastructure. The interactions within a level are not linear, so while there is no specific predefined flow between tasks, the general player experience will be as shown in Fig. 2.

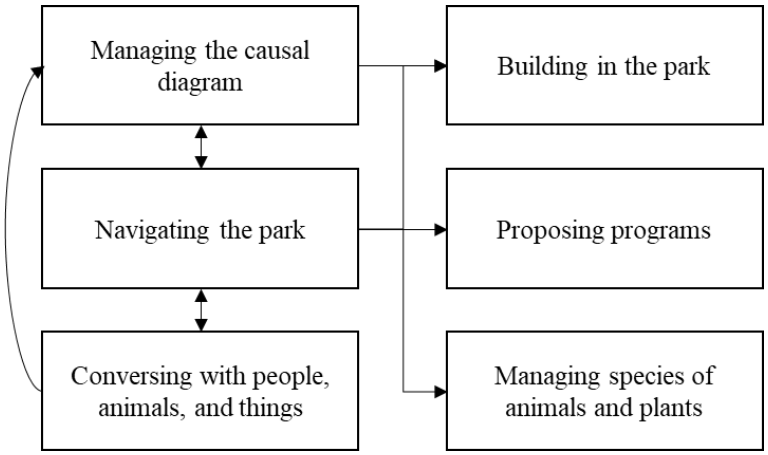


Figure 2. Task Flow

Validation Study

To conduct our validation study, we will use a version of the game that includes a demographic questionnaire at the start. The questionnaire will ask students for information on their gender, race, and ethnicity, which we will use to analyze fairness. We will collect and store data from the game logs, demographic information, and performance-based measures results on our servers. The data will not contain any identifying information except for hashed IDs generated from players' accounts. This will enable us to associate data from multiple game sessions with the same player.

As our intended audience is middle school students and ST is often associated with science, we will recruit students in seventh-grade science classes for our study participants. They will be involved in the study in two phases.

In the study's first phase, we will conduct response process analyses by recruiting seventh-grade students who are not enrolled in the schools used in later phases. We will select ten participants at a time to individually play the game while describing their thought processes out loud. The gameplay and verbalizations will be recorded and analyzed to identify potential changes to the assessment. If significant changes are made, the response process analysis will be repeated with a new group of students until no further major changes are necessary.

During the main phase, we plan to enroll a minimum of 1,000 seventh-grade students in purposefully-selected schools to play the game during several class periods. The duration of the game will be determined based on input from response process analyses, and the number of class periods required to complete it will vary by school. Although students will play the game individually, they will play together in the same class. Students whose parents do not allow them to participate will still be able to play the game, but their data will not be collected.

To evaluate early data before the study is complete, we will conduct preliminary analyses when around 200 students have finished playing the game. Schools participating at different times and

rates allow for this evaluation. These analyses will identify any issues with the assessment, including checking for anything unusual in the Bayesian network and determining whether different subgroups perform or are evaluated differently. The study will be paused if severe problems are discovered that require changes to the game or assessment. Data collected up to that point will likely be ignored, except for parts of the Bayesian network that may be relevant. The main phase will then be restarted, and additional students will be recruited to replace those who already participated.

Data analysis

We will evaluate the correlation between the external performance-based measure and the stealth assessment score separately for each of the three measurement occasions. Multilevel modeling will be used in this analysis to account for the random effect of schools. This will provide evidence of external validity. We will then assess whether different subgroups who are at the same level on the external measure perform similarly in our stealth assessment. Students' gender, race, ethnicity, and their interactions will be added to these models with stealth assessment scores as the outcome and the external measure scores as a predictor. Any subgroups that are too small to include and report without risking participant identification will be excluded from this analysis. This will provide evidence of fairness. Reliability will be assessed using Cronbach's alpha, which will provide evidence of internal consistency. We will also evaluate the multilevel linear relationship between time and stealth assessment score to check whether ST score increases over time as ST skills are learned. Finally, the response process analysis data will be analyzed qualitatively.

Summary of Validity Evidence

To summarize the validity evidence we are gathering, we have evidence from (1) the assessment development process in the form of input and review by ST experts; (2) our competency, evidence, and task models; (3) response process analyses; (4) the associations between our stealth assessment and the external measure; (5) the association between playing the game and ST score over time; (6) Cronbach's alpha for internal consistency; and (7) analysis of subgroup differences related to the intersections of race, gender, and ethnicity.

Future implications

The implications of this study are twofold, methodologically and practically. From a practical point of view, our game design was well accepted as engaging and pedagogical. If adopted, it could be a unique and effective way to assess and develop ST skills, which are essential for success in the 21st century. Methodologically, our proposal is a valuable addition to the ECD framework and stealth assessment initiatives. It demonstrates how a complex skill such as Systems Thinking can be measured using alternative techniques.

Finally, in order to address diversity, equity, and inclusion in our game, it is important to consider the various ways in which individuals approach and analyze problems. This includes recognizing and valuing diverse perspectives, experiences, and problem-solving strategies. Future game considerations should incorporate scenarios and problems representing diverse cultures, backgrounds, and life experiences. Combining an understanding of how structural

inequalities and power dynamics affect complex systems into the game's assessment is crucial. This is particularly important because research has shown that historically marginalized groups tend to score lower in game-based learning environments [18].

References

[1] M. Castells and C. Blackwell, "The information age: economy, society and culture. Volume 1. The rise of the network society," *Environment and Planning B: Planning and Design*, vol. 25, pp. 631–636, 1998.

[2] H. V. Haraldsson, *Introduction to system thinking and causal loop diagrams*. Department of chemical engineering, Lund University Lund, Sweden, 2004.

[3] M. A. Dolansky, S. M. Moore, P. A. Palmieri, and M. K. Singh, "Development and validation of the systems thinking scale," *Journal of General Internal Medicine*, vol. 35, pp. 2314–2320, 2020.

[4] K. E. Dugan, E. A. Mosyjowski, S. R. Daly, and L. R. Lattuca, "Systems thinking assessments in engineering: A systematic literature review," *Systems Research and Behavioral Science*, vol. 39, no. 4, pp. 840–866, 2022.

[5] A. Rempfler and R. Uphues, "System Competence In Geography Education: Development Of Competence Models, Diagnosing Pupils'achievement," *European Journal of Geography*, vol. 3, no. 1, 2012.

[6] V. J. Shute, "Simply assessment," *International Journal of Learning and Media*, vol. 1, no. 2, pp. 1–11, 2009.

[7] R. M. Jaradat, "An instrument to assess individual capacity for system thinking," 2014.

[8] N. Rustaman, H. Firman, and B. Tjasyono, "Development and validation of climate change system thinking instrument (CCSTI) for measuring system thinking on climate change content," presented at the Journal of Physics: Conference Series, IOP Publishing, 2018, p. 012046.

[9] F. Camelia, T. L. Ferris, and D. H. Cropley, "Development and initial validation of an instrument to measure students' learning about systems thinking: The affective domain," *IEEE Systems Journal*, vol. 12, no. 1, pp. 115–124, 2015.

[10] V. J. Shute and S. Rahimi, "Review of computer-based assessment for learning in elementary and secondary education," *Journal of Computer Assisted Learning*, vol. 33, no. 1, pp. 1–19, 2017.

[11] K. Stave and M. Hopper, "What constitutes systems thinking? A proposed taxonomy," presented at the 25th international conference of the system dynamics Society, 2007.

[12] V. J. Shute, C. Sun, and J. Asbell-Clarke, "Demystifying computational thinking," *Educational Research Review*, vol. 22, pp. 142–158, Nov. 2017, doi: 10.1016/j.edurev.2017.09.003.

[13] M. Frank, "Engineering systems thinking: Cognitive competencies of successful systems engineers," *Procedia Computer Science*, vol. 8, pp. 273–278, 2012.

[14] R. D. Arnold and J. P. Wade, "A complete set of systems thinking skills," *Insight*, vol. 20, no. 3, pp. 9–17, 2017.

[15] H. Qin, P.-L. Patrick Rau, and G. Salvendy, "Measuring player immersion in the computer game narrative," *Intl. Journal of Human–Computer Interaction*, vol. 25, no. 2, pp. 107–133, 2009.

[16] A. A. Rupp, M. Gushta, R. J. Mislevy, and D. W. Shaffer, "Evidence-centered design of epistemic games: Measurement principles for complex learning environments," *The Journal of Technology, Learning and Assessment*, vol. 8, no. 4, 2010.
[17] R. J. Mislevy, R. G. Almond, and J. F. Lukas, "A brief introduction to evidence-centered design," *ETS Research Report Series*, vol. 2003, no. 1, pp. i–29, 2003.
[18] C. Harteveld, N. Javvaji, T. Machado, Y. V. Zastavker, V. Bennett, and T. Abdoun, "Gaming4All: Reflecting on Diversity, Equity, and Inclusion for Game-Based Engineering Education," in *2020 IEEE Frontiers in Education Conference (FIE)*, Uppsala, Sweden: IEEE, Oct. 2020, pp. 1–9. doi: 10.1109/FIE44824.2020.9274176.

Appendix A

Task Model

A bulleted list of tasks is provided below:

- **Task 1:** Conversing with people/animals/things

1. Feature 1: A contact list to select contacts to call and initiate a conversation. Note that conversation can also be initiated by clicking on people/animals/things while navigating the park.
    1. Feature 1.1: Button to display the contact list.
    2. Feature 1.2: A pop-up that lists clickable contact names on the left with their associated roles such as botanist, donor, park ranger, etc. on the right. Clicking a name starts the conversation

2. Feature 2: Pop-up dialog.
    1. Feature 2.1: Button at the lower right of the prompt that will open the causal diagram.
    2. Feature 2.2: Text that shows what the person/animal/thing says.
    3. Feature 2.3: Text that shows the response after the player drags any node or set of nodes to the pop-up dialog to ask a question at any point in the conversation.

4. Feature 2.3: Words that reference nodes are underlined. Clicking underlined words will add the associated node to the causal diagram and show the node in the diagram.
5. Feature 2.4: Numbered text response options that the player can click to choose what to say to the person/animal and advance the conversation. May also include a response option to return to the previous topic.
6. Feature 2.5: A response option for arguing from the causal diagram. This opens the causal diagram and allows the player to select and drag nodes and relationships that form an argument to a donor.
7. Feature 2.6: The person/animal/thing's name and role displayed at the top of the dialog.

3. Feature 3: Media representation of a person, a thing, or animal that depicts who/what the player is interacting with.
4. Feature 4: Button to close interaction.

- **Task 2:** Managing the causal diagram
  1. Feature 1: Button to add a rectangular node from a list of known nodes.
  2. Feature 2: Button to add bold arrows for confirmed relationships.
  3. Feature 3: Button to add normal arrows for unconfirmed relationships.
  4. Feature 4: Button to select an eraser to delete a node or relationship.
  5. Feature 5: Polarity icon next to every relationship that can be clicked to set polarity to + or -. Defaults to a question mark.
  6. Feature 6: Buttons to zoom in and out the diagram.

7. Feature 7: Empty spaces in the diagram can be clicked and dragged to pan the diagram.
8. Feature 8: Each node contains a text name, an icon representing the node whenever possible, and a health bar on the right side that shows how well the node is doing in the underlying system.
9. Feature 9: Each node can be double-clicked to display a pop-up with all information about the node that has been seen by the player in conversations.

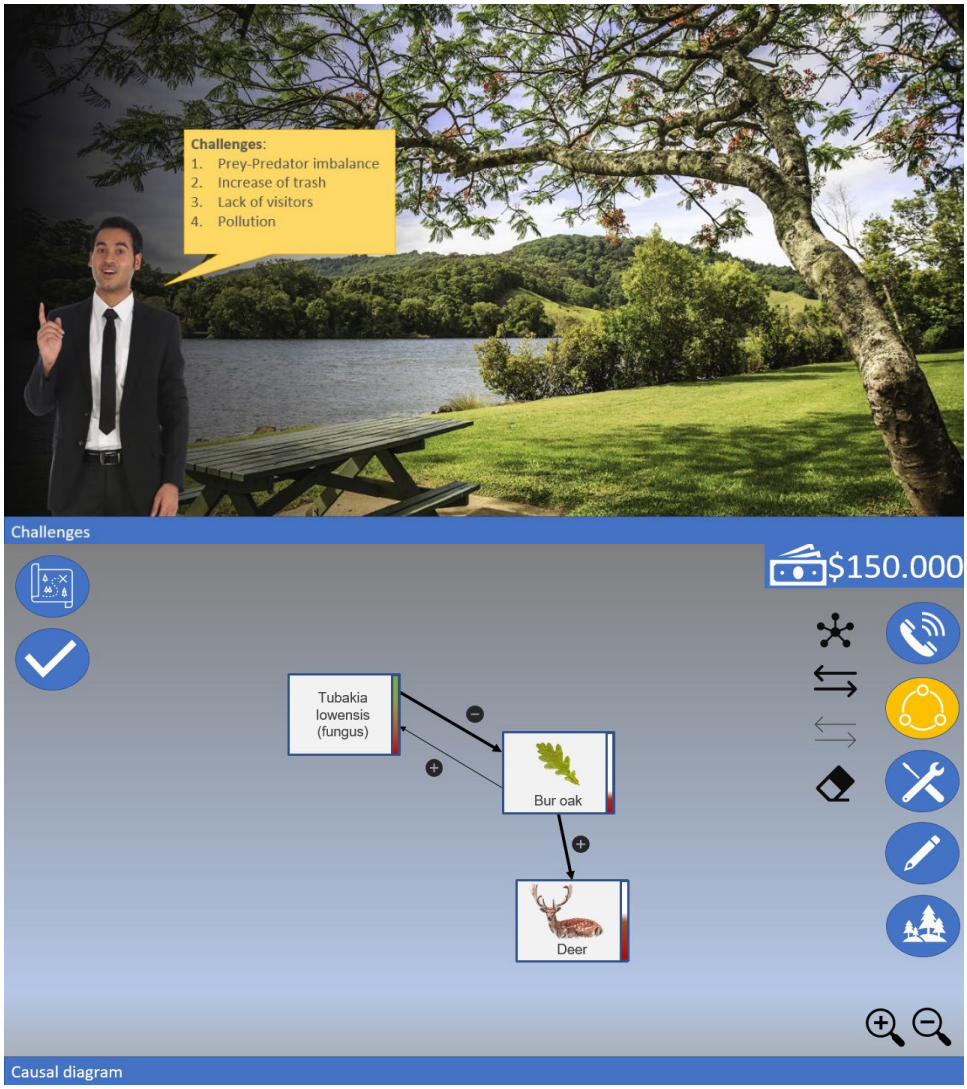
- **Task 3:** Building in the park
    1. Feature 1: Media representations in a table that the player can drag and drop into the park environment. Examples: restrooms, trails, showers, picnic spaces, trash cans, signals, fences, benches, etc. Every time a player add an item to the park, the item is added as a node in the causal diagram.
    2. Feature 2: Button to delete items.
- **Task 4:** Proposing programs
    1. Feature 1: Button to create an educational program that could bring 25 students from a local school. The educational program is a pre-setted narrative that the player drags and drops into the game environment. Programs added to the park as added a node in the causal diagram.
    2. Feature 2: Button to create a community program that could bring 50 people from the county. The community program is a pre-setted narrative that players drag and drop to the game environment. Programs added to the park as added a node in the causal diagram.
- **Task 5:** Managing species of animals and plants

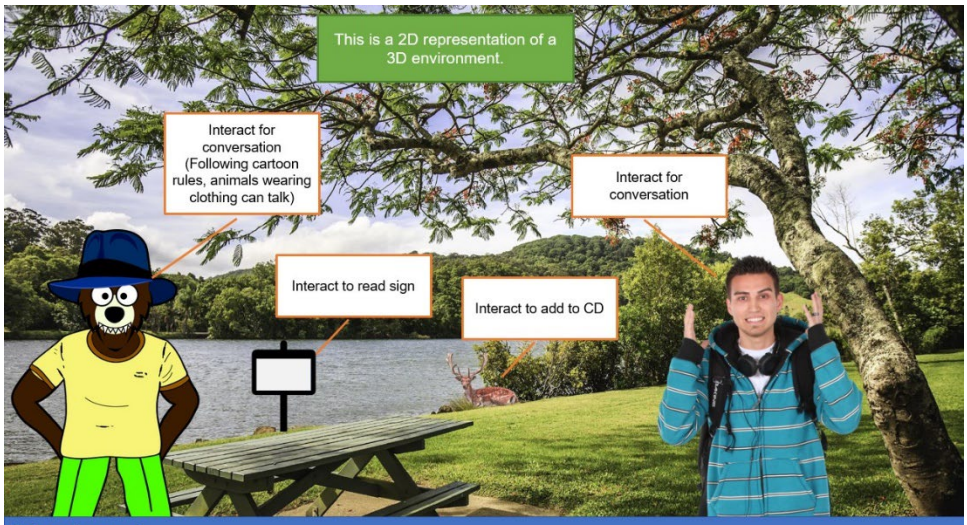
1. Feature 1: Button to eliminate a particular amount of species by selecting them and confirming the deletion.
2. Feature 2: Media representations in a table of available species that players can drag and drop to the park. Species added to the park are added as nodes in the causal diagram if not already present.

- **Task 6:** Navigating the park
    1. Feature 1: A 3D freely-navigable park environment with features determined by the current state of the underlying system.
    2. Feature 2: Species and buildings/facilities that can be clicked to be added to the causal diagram.
    3. Feature 3: Signs, people, and anthropomorphized animals wearing clothing that can be clicked to start conversations.
    4. Feature 4: A map of the park.
        1. Feature 4.1: A button to display the map.
        2. Feature 4.2: A button to close the map.
        3. Feature 4.3: An icon for the current player position.
        4. Feature 4.4: Locations noted with icons and text that players can click to quickly travel to them.

Appendix B

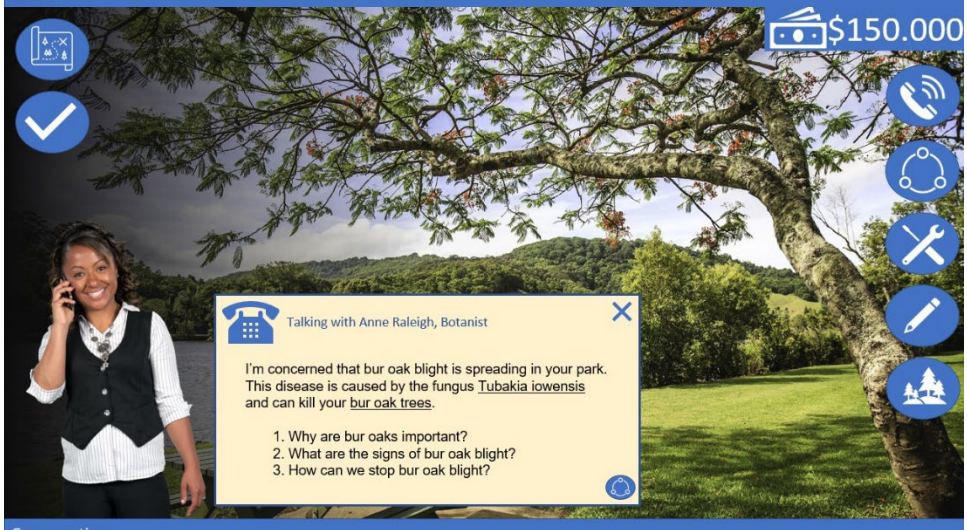
Game Prototypes



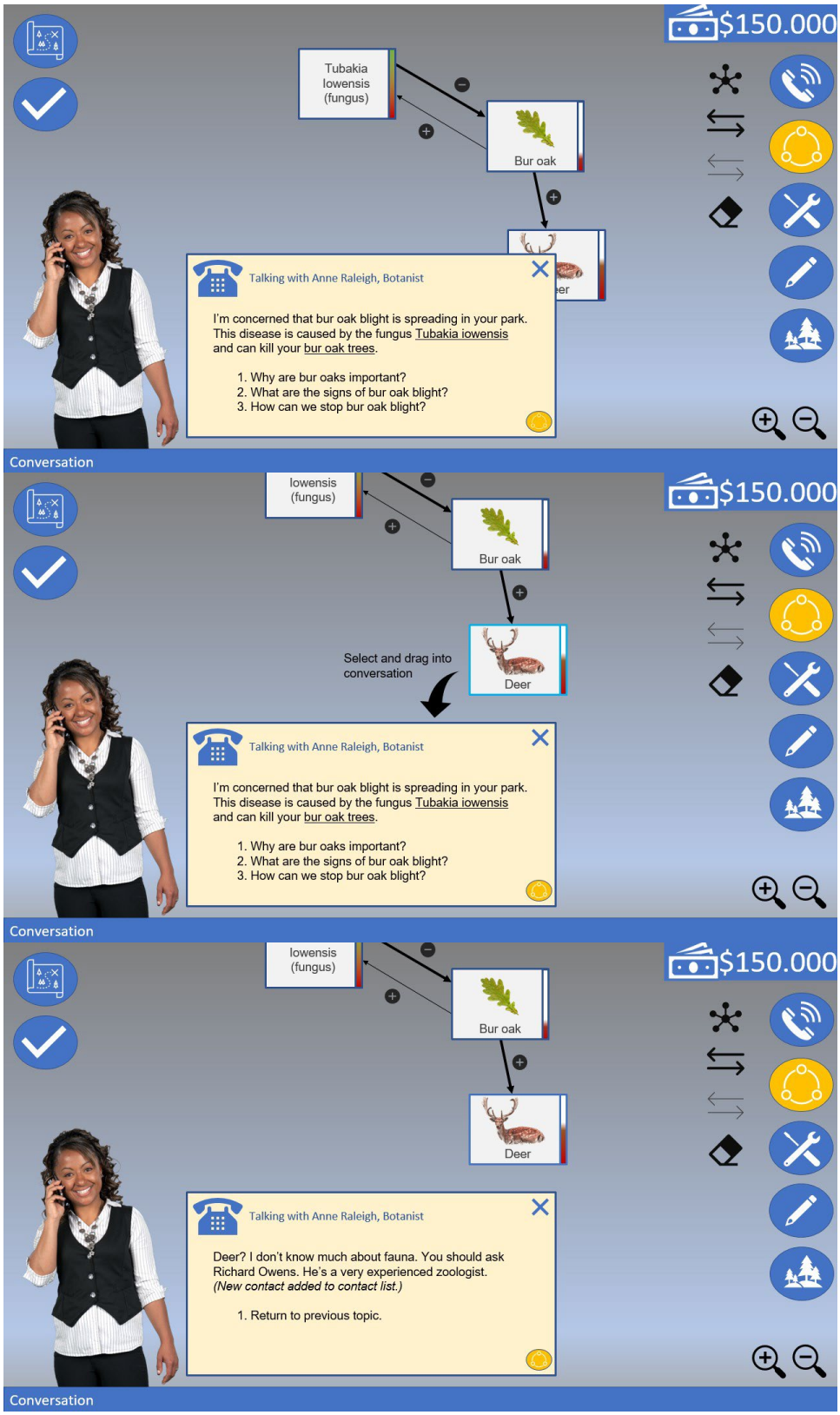


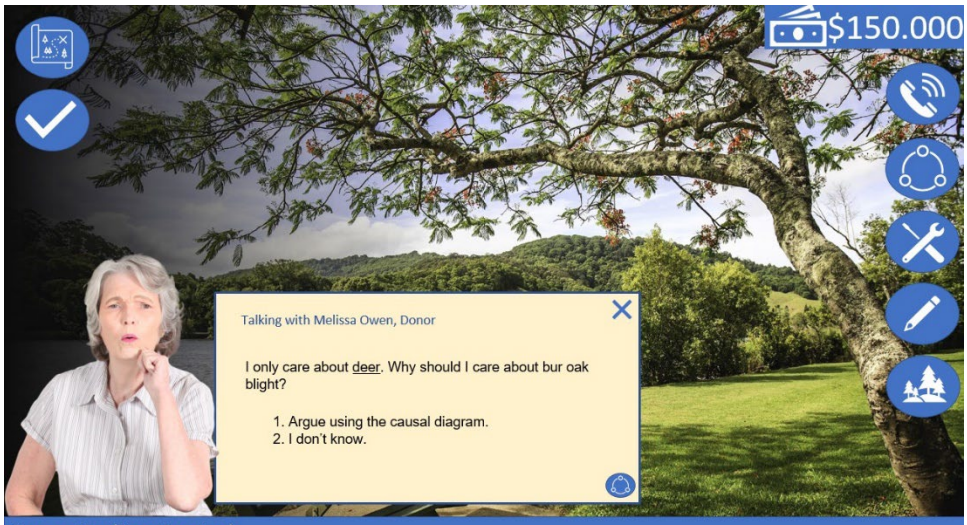
Park environment

Menu

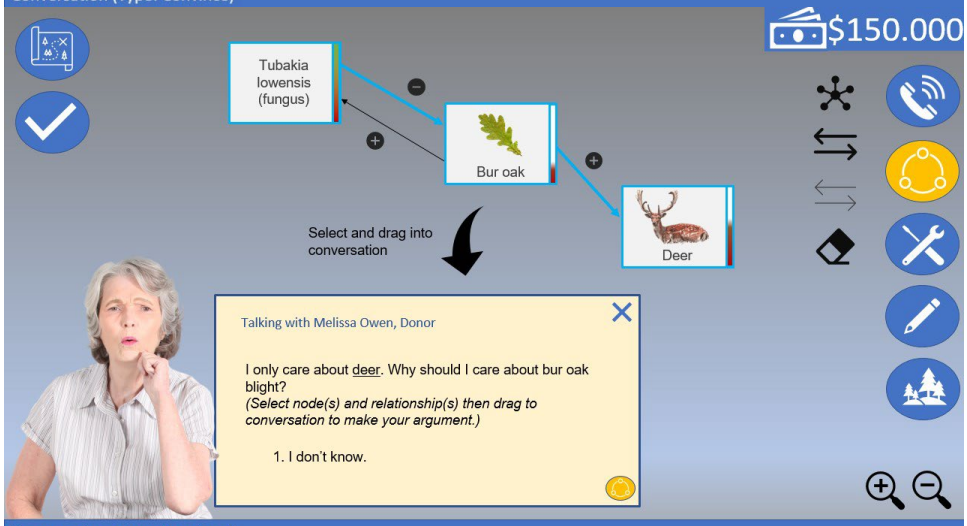


Conversation


$150.000
Tubakia
lowensis
(fungus)
Bur oak
Talking with Anne Raleigh, Botanist
I'm concerned that bur oak blight is spreading in your park. This disease is caused by the fungus Tubakia iowensis and can kill your bur oak trees.
1. Why are bur oaks important?
2. What are the signs of bur oak blight?
3. How can we stop bur oak blight?
Conversation
lowensis
(fungus)
Bur oak
Select and drag into conversation
Deer
Talking with Anne Raleigh, Botanist
I'm concerned that bur oak blight is spreading in your park. This disease is caused by the fungus Tubakia iowensis and can kill your bur oak trees.
1. Why are bur oaks important?
2. What are the signs of bur oak blight?
3. How can we stop bur oak blight?
Conversation
lowensis
(fungus)
Bur oak
Deer
Talking with Anne Raleigh, Botanist
Deer? I don't know much about fauna. You should ask Richard Owens. He's a very experienced zoologist.
(New contact added to contact list.)
1. Return to previous topic.
Conversation

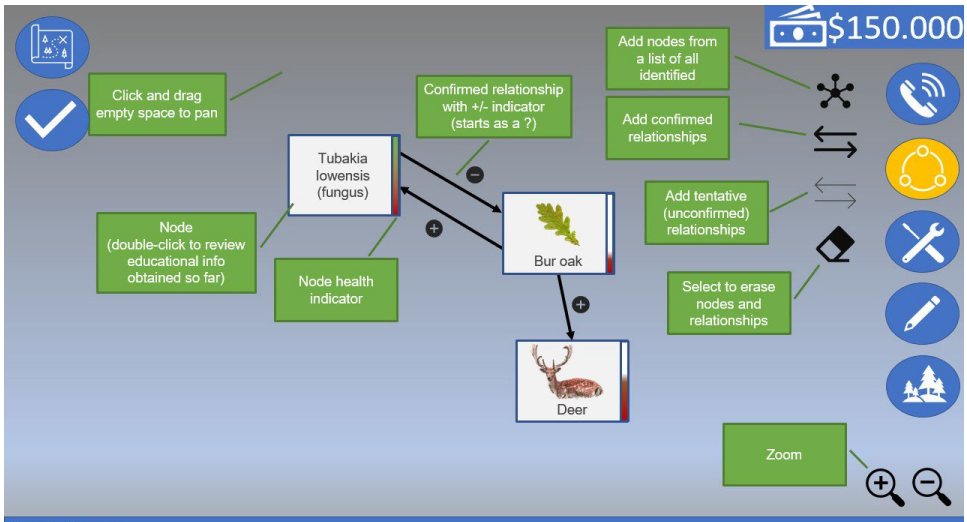


Conversation (Type: Convince)



Conversation (Type: Convince)

Current/New Scene (Part 2: gather information)



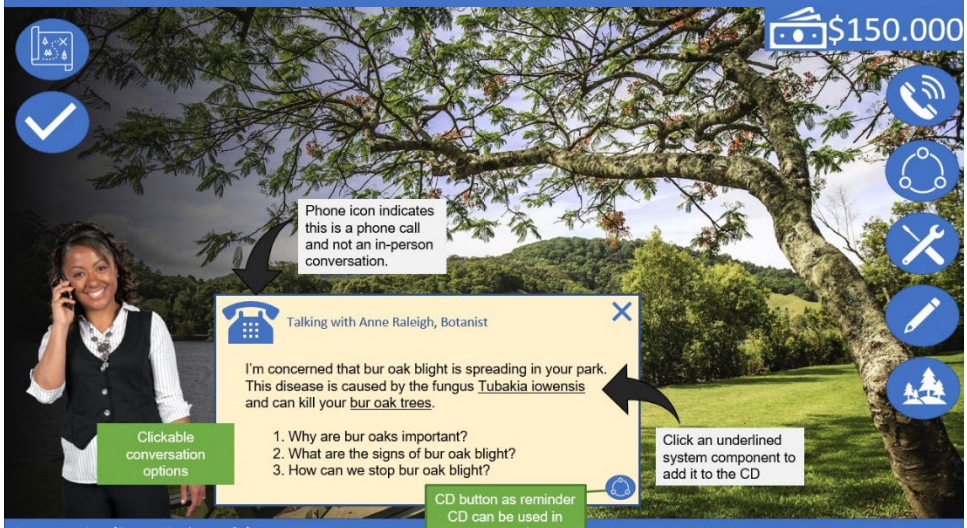
Causal diagram

Building

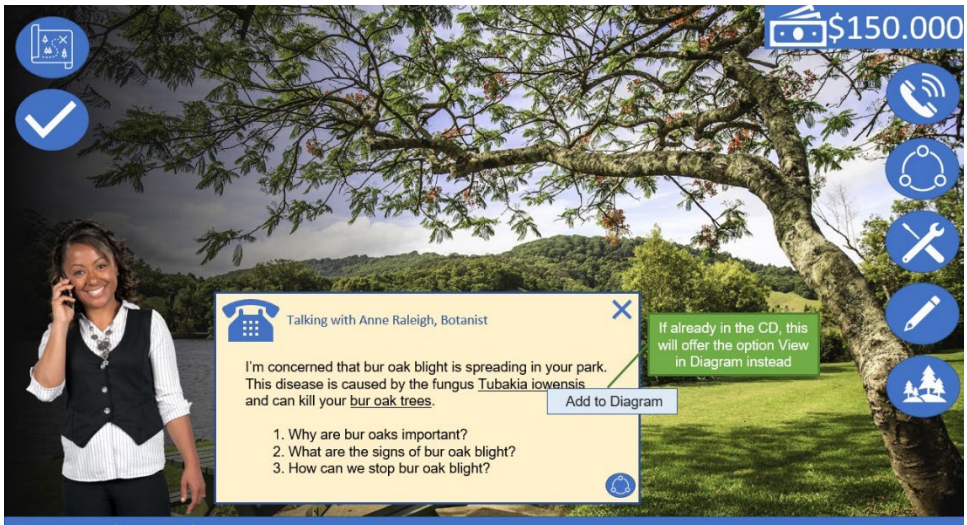
Programs

Nature management

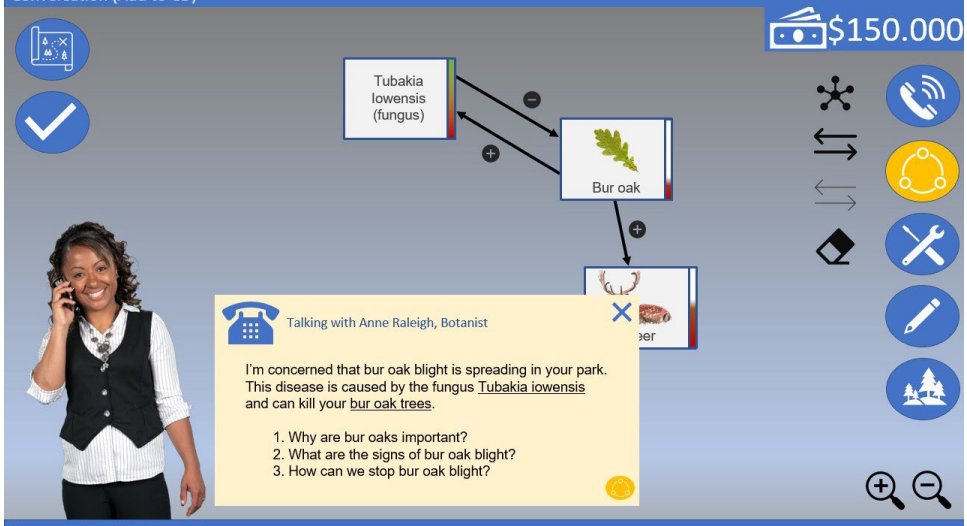
Phone contacts



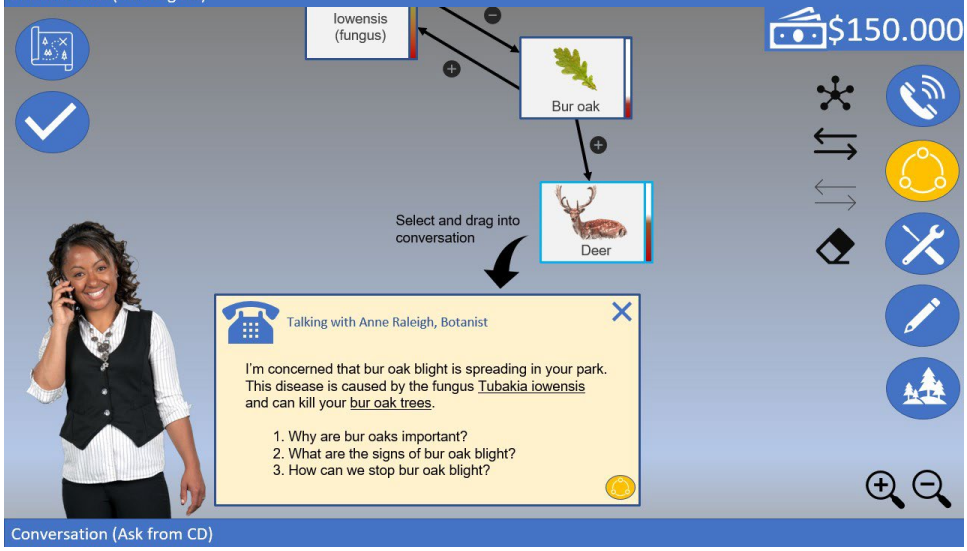
Conversation (Type: Gather Info)



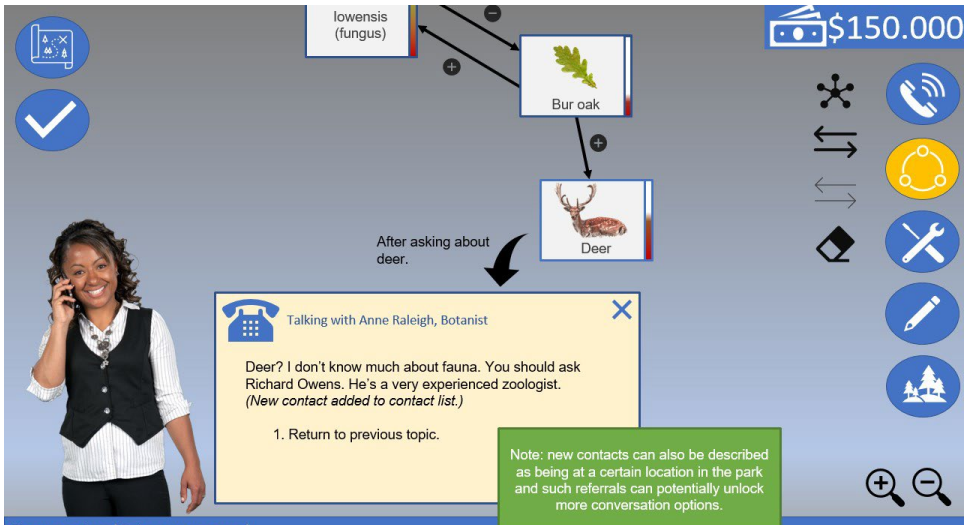
Conversation (Add to CD)



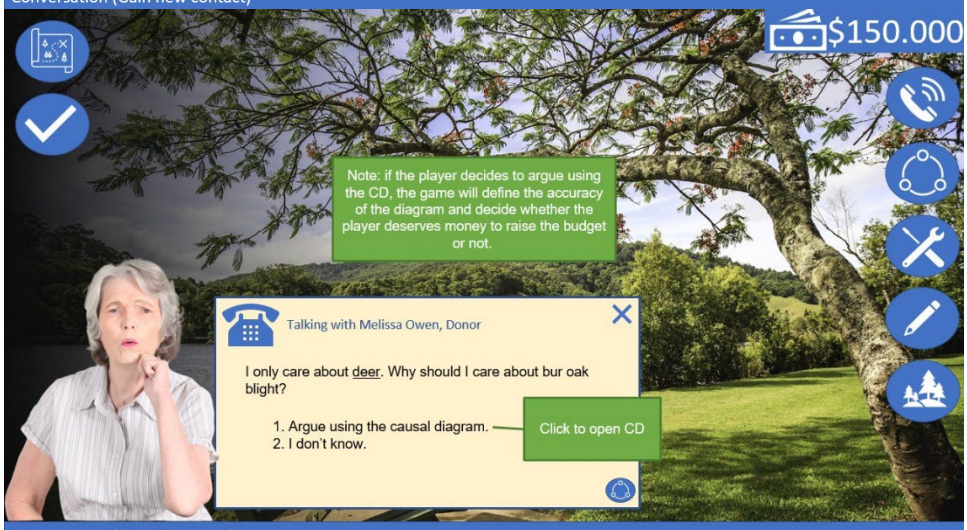
Conversation (Viewing CD)



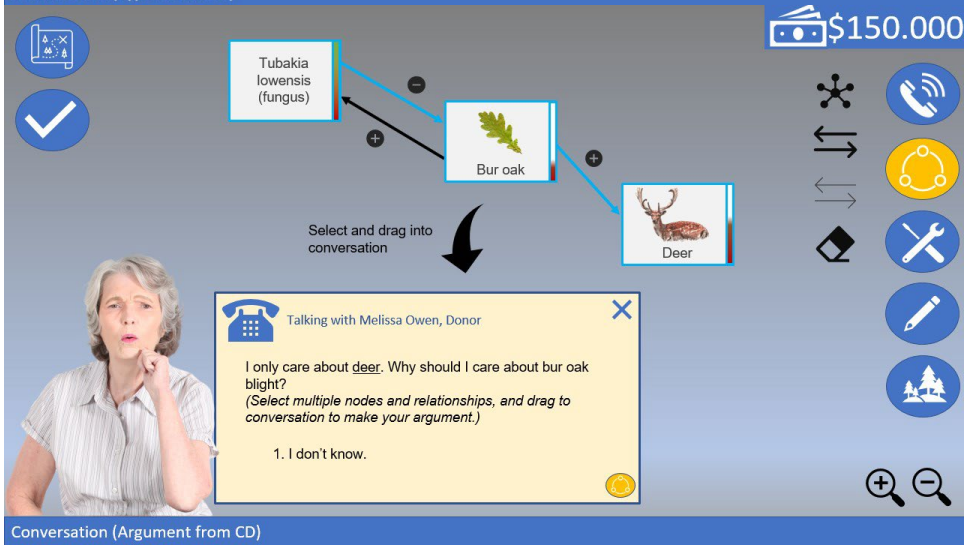
Conversation (Ask from CD)



Conversation (Gain new contact)



Conversation (Type: Convince)



Conversation (Argument from CD)

Appendix C

Difficulty Rubric

1. Total nodes required for all minimal solutions:
    1. 1-3 (0)
    2. 4-5 (1)
    3. 6-7 (2)
    4. 8+ (3)
2. Relationships per node required for all minimal solutions:
    1. (Nodes / relationships)
3. Most complicated relationship in the minimal solutions:
    1. Direct (0)
    2. Direct feedback (1)
    3. Indirect feedback involving 3 nodes (2)
    4. Indirect feedback involving 4+ nodes (3)
4. Total nuisance nodes:
    1. 1-3 (0)
    2. 4-10 (1)
    3. 11-20 (2)
    4. 21+ (3)
5. Nuisance relationships per nuisance node:
    1. (Nodes / relationships / 2)

6. Likely familiarity of the average player with the ecological material and context, as determined by experts:
    1. Not familiar (0)
    2. Familiar (1)
7. Conversations *required* for finding the minimal solutions:
    1. 1-2 (0)
    2. 3-5 (1)
    3. 6+ (2)
8. Additional contacts that experts deem *players* could think are relevant to the solution:
    1. 1-3 (0)
    2. 4-10 (1)
    3. 11-20 (2)