# Dynamic Feedback as Automated Scaffolding to Support Learners and Teachers in Guided Authentic Scientific Inquiry Settings

*Dynamic Feedback as Automated Scaffolding to Support Learners and Teachers in Guided Authentic Scientific Inquiry Settings*

A Dissertation

presented by

Joseph Michael Reilly

to

The Committee on Higher Degrees in Education

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the subject of

Education

Harvard University

Cambridge, Massachusetts

April 2020

Dissertation Advisor: Chris Dede Joseph Michael Reilly

Dynamic Feedback as Automated Scaffolding to Support Learners and Teachers in Guided Authentic Scientific Inquiry Settings

## Abstract

Guided authentic scientific inquiry activities can give students a clear picture of the nature of science and how its fields operate in practice, but such activities are difficult to do well. Too much structure negates the benefit of open-ended activities and student-led investigation, while too little can result in frustration or unproductive floundering. The inclusion of scaffolds for students as well as teachers in such activities is necessary for their successful implementation. Virtual environments can facilitate open-ended inquiry activities by having built-in scaffolds, such as feedback that reacts dynamically to the actions of learners in real time. However, sparse experimental research exists on how formative feedback can best be deployed to students in open-ended immersive environments compared to more constrained tutoring systems or games. Additionally, these tools and curricula must include support for teachers in the design and implementation process if they are to be used in complex classroom settings. Cutting-edge techniques at the forefront of the educational data mining field are being developed for analyzing student actions in such settings, and interest in these techniques is steadily growing. This dissertation documents a pilot implementation of formative feedback tools embedded in the open-ended virtual world of EcoXPT, an experiential curriculum for learning ecosystems science and scientific inquiry. The efficacy of various types of feedback and methods of deployment to both students and teachers are explored, and student outcomes are compared with data from prior baseline EcoXPT implementations. Compared to students in baseline versions of EcoXPT, groups with access to these feedback tools had larger scientific epistemology gains, when

controlling for other factors, and made more complete concept maps. Meaningful sequences of logged actions were identified that correlate positively with outcome measures, and overall trends in feature usage between teachers in the sample were described. Student feedback on the utility of the new feedback features is analyzed, and themes from teacher interviews are described to explore how they used the teacher-facing tools as well as how they saw students reacting to the additional tools.

## Table of Contents

**Acknowledgements**

Over the past five years, many people helped me move through the doctoral program and make adequate progress toward the writing of this document. My success would have been impossible without the patience and support of my wife Sarah and the rambunctious, radiant love of my sons, Oliver and Hugo. I was lucky to join a cohort of passionate scholars who pushed me to see the landscape of education with different lenses and to realize my own blind spots. I would like to thank the Harvard Graduate School of Education doctoral programs office for their assistance navigating the administrative web of this novel Ph.D. program, particularly Dr. Julie Vultaggio, Clara Lau, and Eric Zeckman. I received constant feedback, support, and encouragement on this work from the EcoLEARN team as a whole, including Dr. Shari Metcalf, Dr. Amy Kamarainen, Dr. Shane Tutwiler, Dr. Meredith Thompson, Eileen McGivney, and Emily Gonzales. I owe each of you a debt which I hope to pay forward to a new generation of scholars and researchers in education.

I would like to express my heartfelt appreciation and thanks to my committee. Dr. Matthew Berland has been a much-needed source of expertise in quantitative methodology and log file analysis as well as an invaluable source of information on entering the job market. Dr. Tina Grotzer has been my guide and mentor in transitioning from the science classroom to conducting research that is rigorous and impactful for both practitioners and researchers. Last, but certainly not least, I would like to extend my deepest gratitude to Dr. Chris Dede. As my advisor, his advice shaped my path in innumerable ways and his support in all forms facilitated my burgeoning research career through thick and thin. A five-minute conversation with him would sometimes give me a semester's worth of research ideas.

Portions of this work were supported by the National Science Foundation through Grant No. 1416781 for EcoXPT. The opinions, findings, and conclusions expressed are those of the author and do not represent views of the National Science Foundation. The development and deployment of the LENS feedback tools was supported by a generous grant from the Cheng Yu Tung Research Innovation Fund as well as a Dissertation Completion Fellowship from the Harvard Graduate School of Arts and Sciences. The EcoXPT survey and summary of events per day presented in Appendices A and B were developed by the EcoXPT research team led by Drs. Tina Grotzer and Chris Dede along with Drs. Amy Kamarainen and Shari Metcalf.

## Chapter 1: Introduction

Learning technologies have become increasingly prevalent in science classrooms over the last four decades and have broadly shown promise when used to support types of science learning that could not be accomplished without them (Krajcik & Mun, 2014). Put another way, educational technology can be used in an attempt to *do things better* than traditional methods (e.g., computerized standardized assessments or classroom response systems), or technologies can be developed in an attempt to *do better things* that would otherwise be impossible (e.g., dynamically adapt learning and generate feedback in real-time for all users simultaneously; Fishman & Dede, 2016). Only the latter type of use is transformative, as it aims not to increase the efficiency of the status quo but instead to enable the best possible teaching and learning that will prepare students for the modern workforce. Many transformative technologies in the classroom leverage the large amounts of data generated by virtual or online learning environments, with their use steadily becoming more commonplace as digital native adults enter the teaching workforce (Wade, Rasmussen, & Turnbull, 2013). Drawing insights from these data when available can drive better science instruction, but teachers report low confidence in their ability to leverage these data in consistent ways (Datnow & Hubbard, 2015).

Additionally, these types of educational "big data" can pose challenges related to the volume of data produced (how to store it safely yet cost effectively), the wide variety of structured and unstructured data generated (in potentially incompatible formats with non-aligned data structures across platforms and sources), the veracity of the data (the accuracy and applicability of the data for specific use cases), and the velocity of the data (how quickly it is generated and can be processed to be actionable; Laney, 2001). The educational data mining (EDM) and learning analytics (LA) communities have embraced these challenges and

opportunities across a wide array of student data, providing better supports directly to students as well as providing more actionable information for teachers (Baker & Inventado, 2014). While there is significant overlap between the two communities, EDM focuses more on automated data analysis with a reductionist focus whereas LA foregrounds the role of human judgement with a holistic focus on supporting human intervention (Siemens & Baker, 2012).

Heuristics for dynamic guidance are well understood in tightly structured learning environments such as intelligent tutoring systems (ITS), programs intended to replicate the one-to-one instruction typically provided by a tutor but without human intervention. The constrained problems typical of many ITS (e.g., solving systems of equations or physics problems) are conducive for detecting a student's puzzlement, determining its degree of productivity, and formulating appropriate feedback to be delivered by the computer-based system or the teacher (Roll, Aleven, McLaren, & Koedinger, 2011; Stamper, Eagle, Barnes, & Croy, 2011). While formative feedback and assessment for students and teachers is a well-studied aspect of ITS and structured games for learning (e.g., Sabourin, Rowe, Mott, & Lester, 2012; Sao Pedro, Baker, & Gobert, 2013), open-ended learning environments for complex tasks like collaborative scientific inquiry pose difficult problems for designers in terms of detecting signs of struggle from a wide array of actions. The abundance of this diverse information is problematic in comprehending how best to design and implement interventions to facilitate inquiry; for example, this poses challenges for teachers and students in understanding when struggles are productive or unproductive (Kaplinsky, 2015; Warshauer, 2015). As a result, targeting open-ended and collaborative domains for educational data mining is of increasing interest, as evidenced by this being selected as the theme for the Twelfth International Conference on Educational Data Mining (Lynch et al., 2019).

Another main purpose of this work is to echo calls to provide equitable technology-based learning experiences that are designed for traditionally excluded learners across socioeconomic, racial, cultural, gender, and linguistic spectra (Papendieck, 2018). MOOCs and other popular 21st century learning platforms have been criticized for doing little to combat inequality or actually exacerbating the very problem they aim to alleviate (Hansen & Reich, 2015). While the potential for radical change via educational technology is alluring, efforts to transform education over the past decade have seen their fair share of spectacular failures (Watters, 2019).

Over the past ten years, the EcoLEARN team at the Harvard Graduate School of Education has explored the use of advanced immersive technologies to support learning about the complex causal dynamics of ecosystems (Dede et al., 2019). Through curricula enabled by immersive virtual environments, augmented reality, and virtual reality, students have the opportunity to participate in authentic scientific inquiry and to view science as a dynamic process rather than a body of knowledge or a formulaic procedure. Rooted in the philosophies of design-based research (Anderson & Shattuck, 2012), this work embraces the complexity of real-world classroom settings and has evolved based on feedback from researchers, teachers, and students.

EcoMUVE, the original multi-user virtual environment (MUVE)-based curriculum developed by the EcoLEARN team, simulated a both a pond ecosystem and a forest, letting middle school science students explore the virtual world and tackle problem-based scenarios in ways that mimic how real ecosystem scientists would address these problems (Metcalf et al., 2011). EcoMOBILE, a hybrid curriculum focused on field trips to physical pond ecosystems, was designed to employ augmented reality (AR) to aid transfer of knowledge from the virtual curriculum to the real world and to situate learning in real watersheds (Kamarainen et al., 2013). EcoXPT built on EcoMUVE, adding experimental tools to the virtual pond ecosystem to allow

students to generate additional causal evidence to support their claims of causal relationships between biotic and abiotic factors (Dede et al., 2017). Numerous quantitative and mixed method studies about the efficacy of the EcoLEARN curricula have focused on content gains, shifts in affective dimensions, and teacher perceptions of the technology (Grotzer et al., 2013; Kamarainen et al., 2013; Chen et al., 2016; Dede et al., 2019). These projects and their relevant research findings are elaborated in later chapters.

Building on a decade of research findings from the EcoLEARN projects and using cutting-edge techniques from the fields of educational data mining, learning analytics, natural language processing, and machine learning, this dissertation is focused on leveraging the large amount of student-generated data in near real-time ways to support student learning, maintain engagement, and minimize unproductive struggling while enabling productive struggling (as elaborated below). All of the virtual environments designed by the EcoLEARN team incorporate scaffolds to help guide learners' inquiry, but the scaffolds are static and do not differ based on prior actions taken in the world or provide feedback based on what users have already completed in previous lessons. Additionally, existing scaffolds in the virtual world are all student-facing and do not directly involve the teacher in the classroom (significant teacher guidance focused directly on teacher behaviors is provided in accompanying teacher guides and lessons plans). Further, none of these scaffolds are currently designed to fade over time after students are observed demonstrating mastery of certain concepts, potentially limiting learners' ability to become capable of independent learning and achieve self-regulation (Puntambekar & Hubscher, 2005). While student log file data have been utilized in several EcoLEARN studies, they have not yet been leveraged as a source of evidence to generate dynamic scaffolding. In terms of metacognition, model-based feedback, and guidance for unproductive floundering, a next step in

the transformative work of the EcoLEARN team is to move towards implementing dynamic formative feedback by taking full advantage of the rich data that immersive virtual worlds generate.

In order to determine how best to support learners in an immersive virtual environment, one must explore what types of feedback are appropriate (grain size, specificity, timing, etc.), as well as what insights one can glean from the logged actions of groups using the curriculum. For this dissertation, a modified version of the existing EcoXPT technology and curriculum was developed that provides a two-pronged approach to simultaneously assessing and supporting students in their inquiry activities, as well as teachers in their use of an immersive world-based middle school curriculum. These new features have collectively been dubbed the LENS suite of tools (Learning with Embedded Nuanced Support). This dissertation refers to "LENS" suite of tools or condition as shorthand to refer to the full and previously existing EcoXPT curriculum, videos, posters, and teacher lesson plans with supporting resources such as Powerpoint and classroom hand-outs, developed by the entire EcoLEARN team, to which the LENS modifications outlined above have been made. By comparing performance on pre-post surveys, utilizing emerging analytical techniques on student log file data, and interviewing veteran teachers familiar with the existing EcoXPT curriculum to solicit their input, findings show what types of feedback delivery are most helpful and what types of analyses can be incorporated into open-ended virtual worlds to inform future use. By providing students real-time feedback dependent on their prior actions and formalizing opportunities for reflection and self-explanation, the modified curriculum aims to further the general goal of all EcoLEARN projects to provide a personalized experience that helps all learners engage in authentic scientific inquiry.

Additionally, these dynamic features and formative feedback measures must include the teacher in the design and implementation process if they are to be used in complex classroom settings rather than one-to-one tutor replacements. Models based on log files for use in computer-supported collaborative settings must account for rich conversations within and between groups that are not captured, as well as for the role of the instructor in the classroom. Teacher-facing summaries of student actions in the virtual world can provide indications of which students might need additional help. How to design and implement orchestration tools such as these has been an area of debate in the learning analytics community (Van Leeuwen & Rummel, 2017), and this work adds more evidence to the conversation.

This dissertation is divided into six chapters, with the first being this introduction. Chapter 2 provides an overview of relevant science education, educational data mining, and learning analytics literature that contextualizes the need for this work and the prior research from which it draws. I then outline the three research questions explored in this dissertation. Chapter 3 provides details of the sample and methods used to answer those research questions, elaborating on specific analytic techniques piloted with EcoXPT data and on the general data analytic procedure used. Chapter 4 presents the findings of this study with regard to comparing student performance on baseline EcoXPT with a version containing the LENS tools, exploring how frequently the new scaffolding tools are used and seeing what impacts these tools had on how students conducted investigations in the virtual world. Chapter 5 summarizes these findings in terms of answering the research questions and discusses new contributions this work makes, limitations to the current study, and potential future work. Finally, Chapter 6 concludes the dissertation with final thoughts and synthesizes my findings into takeaway points that are relevant for other researchers in the fields of educational data mining and learning analytics.

## Chapter 2: Background and Context

This chapter briefly summarizes several strands of literature that are drawn from science education, educational data mining, and learning analytics. This foundational literature segues into the research questions for the dissertation as well as hypotheses about what will be investigated.

### Scaffolding Inquiry in Problem-Based Learning Curricula

Employment in science, technology, engineering, and mathematics (STEM) careers grew by 10.5% from 2009 to 2015 compared to 5.2% growth in non-STEM jobs, and the average national wage for STEM jobs is nearly double that of other occupations (Fayer et al., 2017). Despite this, employers report difficulty filling these positions due to a lack of sufficiently qualified job candidates (Langdon et al., 2011). Research suggests one possible reason for this gap is a discrepancy between how science is taught in schools and the reality of what these jobs require (National Research Council, 2011).

An authentic scientific inquiry lesson mirrors a real scenario a scientist might face, and students taught with this method gain "knowledge and understanding of scientific ideas, as well as an understanding of how scientists study the natural world" (National Research Council, 1996, p. 23). Instead of focusing on memorization and "cookbook"-style lab activities, "…learners can investigate the natural world, propose ideas, and explain and justify assertions based upon evidence and, in the process, sense the spirit of science" (Hofstein & Lunetta, 2004, p. 30). While authentic inquiry activities address these issues, they are difficult to implement and assess. Many teachers' attempts at these activities result instead in simplistic inquiry tasks, engaging less complex cognitive processes and utilizing less powerful epistemologies than those real scientists employ (Chinn & Malhotra, 2002).

Even when using lessons based on authentic inquiry, teachers may not intervene sufficiently or rapidly enough, resulting in students becoming lost, frustrated, or conceptually confused (Brown & Campione, 1994). While productive struggle has long been a focus of mathematics education (Hiebert & Grouws, 2006; Warshauer, 2015), similar work has not been done on what productive versus unproductive struggle looks like to science teachers in inquiry settings, and how effectively they respond to these visible signals. Optimal levels of difficulty support learning, comprehension, and retrieval of challenging material that will aid subsequent use of content or skills (Bjork & Bjork, 2014), but educators need to properly support students through challenging moments to make them fruitful (Kapur, 2016). Kapur (2016) elaborates that productive failure does not maximize performance in the short term but can maximize learning in the long term, and that tasks designed for productive failure must afford sufficient problem spaces for exploration yet still direct learner attention toward the critical features of the target concept. Fostering persistence and perseverance has been conceptualized as a collective enterprise among groups and classes rather than as an individual capacity (Sengupta-Irving & Agarwal, 2017), utilizing multiple perspectives on a problem to gain deeper understanding of the material. Researchers in maker education have embraced early failure as a feature of learning in applied settings, with a focus on how learners respond to failure and how educators can aid this process (Maltese et al., 2018).

Scaffolding is needed in problem-based and inquiry learning to provide sufficient support to maintain motivation, reduce cognitive load for learners, and facilitate collaborative learning in complex domains (Hmelo-Silver et al., 2007). In addition to structuring student work, scaffolds can "problematize" tasks by encouraging learners to attend to essential ideas or processes in the learning experience that may be otherwise overlooked (Reiser, 2004). These two aspects of

scaffolding can be complementary but also represent a tension where creating scaffolds may lower initial barriers to engaging with a challenging problem while potentially increasing the difficulty of the task by revealing important but less salient challenges (i.e, students feel they have mastered a task but are actually unaware of what they do not know without scaffolds indicating it). Learning with these types of authentic, problem-oriented tasks can be enabled by technology-supported learning, where access to additional multimedia resources and on-demand delivery of content can foster deep learning rather than a surface-level understanding of a problem (Wang et al., 2017). In particular, computer-based "cognitive mapping" (concept mapping focused solely on causal relationships) activities have been shown to help students problem-solve during complex tasks (Wang et al., 2018)

Virtual environments can overcome many of the traditional pitfalls of inquiry activities by offering a low cost, efficient, valid, and rich context for teaching and assessing complex inquiry-based science learning (Ketelhut et al., 2010; Asbell-Clarke et al., 2012, Code et al., 2013). Helpful scaffolds can be hard-coded into the learning environment and can be configured to shift depending on user actions or patterns over time. These scaffolds have been shown to be most effective in helping students with low prior knowledge in guided discovery scientific settings (Großmann & Wilde, 2019). Virtual learning environments presented in three dimensions are well-suited to providing these types of scaffolds, as they can simulate authentic problems or scenarios and can provide greater opportunities for experiential learning, sustained motivation and engagement among students, and more contextualization for learning than many other platforms (Dalgarno & Lee, 2010). Immersive virtual environments can be explored in an open-ended way by students and teachers, embracing an exploratory learning model that emphasizes reflection and collaboration (Freitas & Neumann, 2009). Virtual environments can

offer different entryways to activities that connect with learners' identities, interests, and competencies while offering a wide variety of supports to keep them comfortable and encouraged (Kolodner et al., 2017).

**Assessing Performance in Open-Ended Environments**

Virtual environments allow for unobtrusive collection of rich user-data streams that can be analyzed via "stealth assessment" to support learning and facilitate the reactive nature of the scaffolds (Shute, 2011). Instead of interrupting activities with assessment items or waiting until the end of an activity to assess learning, stealth assessment uses data generated during the natural course of learning in virtual environments to model understanding of different constructs with the potential to alter the course of the activity based on evidence gathered from these data. Further, these assessments can be reliably and validly applied to learning environments by utilizing the evidence-centered design framework (Shute & Moore, 2017). Despite the increasing prevalence of virtual learning environments that attempt to provide formative assessment, relatively few studies examine the contributions of these assessments to student learning (Wijesooriya et al., 2015).

In addition to supporting learning, these data can be used to assess student engagement and immersion (Calvo & D'Mello, 2010); finding ways to keep learners engaged while confronting difficult content is a central goal of many computer-based learning environments. Maintaining engagement and immersion in game-based learning has positive effects on learning outcomes (Huizenga et al., 2009), and tackling challenging tasks can help students learn more while potentially increasing engagement and immersion (Hamari et al., 2016). Boredom has been shown to be strongly deleterious to learning outcomes in computer-based learning environments, while confusion and frustration were shown to be less persistent and less harmful to learning

(Baker, D'Mello, Rodrigo, & Graesser, 2010). That said, simply adding elements to make activities more fun in an attempt to reduce boredom does not correlate with higher learning gains or cognitive engagement (Ke et al., 2016).

While many virtual tools exist for collecting summative assessment data, these types of data-driven formative assessments and affect detectors are typically seen only in intelligent tutoring systems and in highly structured games for learning or assessment. Most ITS attempt to guide students via hints based on whatever the student previously entered, with Bayesian knowledge tracing (BKT) being the classic approach to modeling a learner's mastery of a subject (Corbett & Anderson, 1994). Many ITS are tightly structured and constrained to facilitate automatic grading and feedback generation. Traditional model-tracing tutors require knowledge engineering and cognitive task analysis of novices and experts. That results in slow, laborious development (Koedinger et al., 2013), and this process becomes even more difficult when applied to larger domains or open-ended tasks.

In contrast, some tutors utilize a data-driven approach where prior student data can be used to generate certain states that learners might be in (Rivers & Koedinger, 2013). Data-driven tutors view a solution space as all possible intermediate states a learner might move through to get to an expert or correct state. When a learner enters an answer or submits an artifact, the tutor makes assumptions about what state the learner is currently in, maps the shortest distance to get them to a desired state, then gives a hint to get them to the next closest state to the correct one (Rivers & Koedinger, 2017). This method has proven popular in open-ended programming tasks where all possible programs students might write could never be known a priori, and such data-driven methods can deliver feedback to teachers to help them adapt traditional and online instruction to help their students (McBroom et al., 2018).

*Hint Factory*, the ITS that pioneered data-driven tutoring, utilizes a Markov decision process (Sutton & Barto, 1999) where all known states are evaluated for goodness and goal-oriented feedback is given to move students to a state with a higher rating (Stamper, Barnes, Lehmann, & Croy, 2008; Barnes & Stamper, 2008). This approach works well when a large amount of actions for students to perform are possible and many of them can be correct. While *Hint Factory* could only generate hints for specific problems utilizing prior data, more recent data-driven methods utilize unsupervised machine learning methods to cluster student solutions based on structural features and to generate feedback based on those patterns (Gross et al., 2012). Another strategy utilizes the weighted average of multiple previous states to create a *Continuous Hint Factory* that can function better in the sparsely populated state spaces common in open-ended environments (Paaßen et al., 2017). These methods allow feedback generation even when there is no one "correct" model for states to move towards.

Additionally, most feedback-generation systems assume a single individual is using the tutor; in collaborative settings such as EcoXPT, a model of group knowledge must instead be constructed. Le, Strickroth, Gross, and Pinkwart (2013) have proposed methods to detect strengths and weaknesses of individual learners to support their co-construction of code in a dyadic learning system. Lessons learned from computer-supported collaborative learning research may aid the development of tools designed with dyads in mind.

Another method that has proven effective in ill-defined domains is using classification algorithms to find desired and undesired actions based on hand-coded training data. This method is especially relevant as it has been employed in scientific inquiry and experimentation with middle school students. Sao Pedro, Baker, Montalvo, Nakama, and Gobert (2010) studied *Science Assistments*, a series of microworlds to teach students experimentation and data

collection skills. They utilized "text replay tagging" of student activities where chunks of log files were analyzed and coded along with video of students to tag certain actions learners took in the world such as "Tested Hypothesis" or "Never Changed Variables." A decision tree algorithm was then trained on this tagged data, resulting in 85% accuracy identifying when students employed a control of variable strategy. In addition, the classifier could identify when students were running experiments to test a specific hypothesis with 86% accuracy.

*Science Assistments* evolved into *Inq-ITS* and now provides a wide variety of assessment and tutoring activities for science inquiry practices, along with a pedagogical agent that provides real-time feedback to learners. Inclusion of this scaffolding allowed for an extension of the traditional BKT framework that improved predictive performance for student mastery of skills; this indicates that these skills can be transferred between topics (Sao Pedro, Baker, & Gobert, 2013). These assessments and others developed since also generate reports for teachers using the platform and can alert teachers in real-time when students might benefit from intervention (Gobert & Sao Pedro, 2016).

Other relevant explorations of scientific experimentation in ITS include the use of natural language processing (NLP) to evaluate reflective dialogue and structured self-explanations prompts (Dzikovska et al., 2014). Instead of having students write explanations as "do now" activities on paper or through a separate interface, capturing and analyzing that data in real-time may allow feedback based on otherwise unobserved understandings or misconceptions. *Inq-ITS* employs a simple NLP automated scoring technique to assess student use of the claim, evidence, and reasoning framework (CER; McNeill & Krajcik, 2011) that strongly correlates with human scores across multiple inquiry activities (Li, Gobert, & Dickler, 2017a). Among the most interesting text-related findings from the platform is that student written explanations at the end

of an inquiry task often don't match up with the logged investigations and actions conducted in the virtual worlds (Li, Gobert, & Dickler, 2017b). Therefore, assessing only investigatory skills or written explanations will not sufficiently capture overall inquiry skill. While written explanations were coded by hand by experts in this study, automation of these methods is certainly possible.

Several games for learning and game-based assessments (GBAs) have also explored the ability to mine student data to provide feedback or hints for players. Many games traditionally use tutorial levels or messages to explain mechanics, but Andersen et al. (2012) found that tutorials work best when presented in context in a just-in-time fashion. In open-ended games where feedback has been found to be more useful, a classifier-based data-driven approach to learn players' behavior in games can model their learning over time (Lee, Liu, & Popovic, 2014). Several "off the shelf" methods such as *Playtracer* and *ADAGE* exist to extract relevant user data from games for use in fitting these classifiers and improving learning (Andersen et al., 2010; Halverson & Owen, 2014). As seen in the discussion of ITS literature, most feedback-related analysis is at the individual level, but increasing attention is being paid to collaborative problem-solving in open-ended games for learning (Bauer & Popović, 2017).

*Crystal Island*, a game-based learning environment designed to teach middle school microbiology (Rowe et al., 2009), has since served as a hotbed for several analytical strategies for assessing student performance in open virtual worlds. Several different levels of narrative-based scaffolding have been tested with the environment, and problem-solving guidance has been shown to improve learning when compared to minimal or no support conditions (Lee, Mott, & Lester, 2012). When this scaffolding takes the form of a virtual learning companion, female students were typically more engaged than males, indicating that scaffolds adapting to student-

level information may offer the most optimal supports (Pezzullo et al., 2017). In addition to modeling player learning, player goal-setting behavior has been modeled via deep learning techniques to predict player goals and actions without the need for laborious feature engineering (Min et al., 2016). While the freedom to explore such open-ended scenarios grants a large amount of agency to players, scaffolding within these environments is key to improving student self-regulation of learning and to maximize learning gains (Dever & Azevedo, 2019).

As their main goal is assessment, GBAs do not always report feedback to users, but some are intended as supplemental practice for learners to identify gaps in their knowledge and improve areas of weakness. Tsai, Tsai, and Lin (2015) showed that immediate elaborated feedback helped high school students learn about energy in an online GBA and may be a valuable form of self-assessment to promote self-regulated learning. *Mars Generation One: Argubot Academy* was designed to teach and assess argumentation skills through a role-playing game and linked key gameplay telemetry actions to relevant constructs being assessed, allowing teachers to visualize student progress through learning goals and how they mapped to Common Core standards (Bertling et al., 2015). Contextualizing data in this way for teachers increases the salience of this data and may lead to more use in the classroom.

Interestingly, an entire GBA exists to measure how students self-regulate learning choices based on feedback. In *Posterlet*, learners design a poster and then receive automatic feedback based on specific design principles (Cutumisu, Blair, Chin, & Schwartz, 2015). The frequency of student selection of negative feedback instead of positive feedback correlated with learning gains, as did the number of iterations of design and feedback. Learners spent more time dwelling upon and revising their posters after receiving negative feedback rather than positive.

In summary, stealth assessment of student interaction data with ITS and game-like systems has shown promise that formative assessment and automatic feedback generation is possible even when the system is more open-ended in nature. Machine learning techniques allow researchers to uncover trends in student behaviors or responses, make sense of them, then report findings to students or teachers. How to report those findings in a useful way, however, is an active area of research and an ongoing conversation in the learning sciences.

**Hint Quality and Help Seeking**

There are many different ways to provide hints to learners in computer-supported inquiry environments, including but not limited to constraints on processes, performance dashboards, prompts, heuristics, and scaffolds (De Jong & Lazonder, 2014). In a review of the efficacy of these diverse forms of feedback in computerized inquiry settings, Zacharia et al. (2015) explain that different types of feedback are better suited for different phases of inquiry activities. While the orientation phase generally requires little guidance, many different types of scaffolds and supports have been designed for the conceptualization and investigation phases but little empirical evidence exists in support of their efficacy in applied settings (Zacharia et al., 2015). Guidance must be delivered in appropriate forms and according to the needs of the student (Belland & Drake, 2013), but extant literature is sparse regarding what these optimal forms and customizations are for science education. A meta-analysis of this literature for feedback systems in all types of computer-supported learning environment indicated that elaborated feedback is generally more effective than feedback specifically on correctness or when systems provide a "correct" answer (Van der Kleij et al., 2015), and that immediate feedback is generally more effective than delayed feedback.

Worked examples are another form of hint popular in ITS and games. While worked example-based feedback is generally as effective as tutored problems and showing erroneous examples, worked examples are by far the most efficient form of feedback (McLaren et al., 2015). While this finding was from stoichiometry problem solving in a chemistry tool and not an inquiry-based environment, the efficiency of feedback and hint tools is an important consideration during time-constrained interventions or curricula. Worked examples are typically effective in aiding deep understanding when presented during initial skill acquisition, although worked examples are best used when combined with prompts and suggestions to encourage self-explanation of the examples (Renkl, 2014).

It is essential that students using conceptually-integrated games or environments (where learning content is fully embedded in manipulating the virtual space) are able to relate their in-game actions to formal learning goals (Clark & Martinez-Garza, 2012; Habgood & Ainsworth, 2011), and one way to make that relationship explicit is through self-explanation of their problem solving processes. Self- explanation encourages students to engage in deeper learning by exploring what they do and do not know, and it has been shown especially effective in multimedia environments (Roy & Chi, 2005). These tasks run the risk of increasing cognitive load of learners, but steps such as offering learners self-explanation options or scaling the abstraction of prompts based on past learner performance can mitigate this issue (Adams & Clark, 2014; Clark et al., 2016). Both very simple and very abstract self-explanation prompts are not useful, indicating an optimal level of reflection and introspection that is useful during game-based learning (O'Neil et al., 2014). Self-explaining (along with note taking and concept mapping) has been shown to increase student engagement learning material in an attempt to increase learning (Chi & Wylie, 2014).

While the ability to generate feedback automatically is exciting, ensuring the helpfulness of that feedback for the learner is paramount. For example, providing hints where none are needed or providing high-level guidance where a learner is struggling with something foundational are both unhelpful. While prohibitively difficult in the past, fitting models in real-time on individual students based on their early actions is becoming more feasible (Koedinger et al., 2013). This could allow initial hints to be more on-target; initial hint quality is very important to reducing negative behaviors in open-ended problems (Price et al., 2017). Early hints are typically seen by more students, help on post-test performance when compared to control groups, and have also been shown to promote near-transfer between problems and improve later performance (Stamper, Eagle, Barnes, & Croy, 2013). In general, bad hints (e.g., ones that incorrectly diagnose issues or increase cognitive load) are worse than no hints at all, and the earlier an appropriate hint is delivered, the larger impact it has on learning.

In addition to content-area experts rating hint quality, feedback and involvement of the teacher in formal educational settings is becoming increasingly more common in hint generation. By showing teachers a list of issues and suggested help automatically detected from a class, they can then change what types of hints are being shown or how often these appear, based on understanding of the prior knowledge of classes or their perceptions of student competence (Ben-Naim et al., 2009). Feedback can be directed toward specific misconceptions seen in multiple cases in the class, so the expertise of the teacher can be leveraged in a tutor environment. Teacher-facing learning dashboards typically work well in collaborative group learning environments, but their efficacy is difficult to measure, and current research has focused on a limited selection of data types from which dashboards can draw (Verbert et al., 2013). Interviews with teachers regarding their needs and desires for a dashboard revealed several central themes

such as helping them intervene at critical points, easing their overall load, seeing if their interventions are actually helping, and learning information about student progress that isn't obvious (Holstein et al., 2017, 2019). These types of co-design projects ensure both the utility of the learning analytics application in the field and empower teachers as key stakeholders in the work rather than simply research subjects.

While some tutors and programs automatically give hints to learners, many provide hints only when the learner asks for one. Designers of these systems view help-seeking as an important process in self-regulated learning and aim to aid students to better regulate their own learning by providing hints on demand. A learner's prior knowledge, beliefs, and thinking all affect how feedback is received and used (Butler & Winne, 1995), making it difficult to say what effect automatically generated hints might have on a given user. Tools designed explicitly to foster self-regulation along with the learning of content are increasingly common (Hadwin & Winne, 2001), as self-regulation is considered essential alongside 21st century competencies (Wolters, 2010).

Maladaptive help seeking behaviors, however, may undercut learning objectives. For example, early work with on-demand, principle-based hints often found students rapidly clicking through hint messages to get to "bottom-out" hints that would explain exactly what to do (Aleven et al., 2016). This and other maladaptive help-seeking strategies are referred to as "gaming the system" (Baker, Walonoski, Heffernan, Roll, Corbett, & Koedinger, 2008). To maximize the usefulness of feedback, the student's frequency of hint access should be low, and time spent thinking about each hint should be high (Wood & Wood, 1999). This may, however, vary based on prior knowledge of the learner. For this and other reasons, Aleven et al. (2016) recommend focusing more work on opportunities for self-explanation and reflection, in addition

to principle-based hints that include what action should be taken, when to do it, and why to do it. Work from Levy and Wilensky (2011) in open-ended environments suggests focusing on content-based interventions that remind students of underlying principles of the subject may be more valuable than procedure-based feedback. Their work indicates that a deep understanding of scientific principles and properties may be more helpful than next step hints.

In general, good formative feedback does not increase cognitive load, reduces uncertainty in goals, shifts student focus to learning instead of performance, and elaborates on how and why to solve different types of issues. Conversely, formative feedback should not make normative comparisons with other learners, interrupt active engagement, or exclusive utilize text; it should neither praise nor discourage the user (Shute, 2008). To understand the potential role that different types of hints and methods of delivery might play in the EcoLEARN curricula, a broad survey of work previously done with the curricula is necessary. Several precursor curricula have focused on assessment and feedback, and multiple EcoLEARN studies have used log file data to assess student learning, but these assessments have not yet been used to generate feedback for students or teachers.

**Previous Work on EcoLEARN Curricula**

Two different immersive virtual environments were developed prior to or concurrent with the start of the EcoLEARN group that greatly influenced the trajectory of the curricula with regard to assessment and feedback delivery. Inspired by earlier multi-user virtual environments (MUVEs) that focused on science instruction, such as Bruckman's *Moose Crossing* (1996) that explored the role of inquiry in a MUVE, *River City* was designed to teach scientific inquiry skills to students in formal educational settings (Nelson & Ketelhut, 2007; Ketelhut et al., 2010). Simulating a city with a sick population, students worked in teams to determine why the city's

population was ill. One version of the activity that utilized individual, reflective feedback found increased viewing of the feedback messages was associated with larger knowledge gains, although the use of the system varied widely between students (Nelson, 2007). The embedded guidance was found to have similar effects for male and female students. In addition to student-facing feedback generated from logged actions, a dashboard allowed teachers to monitor chat logs from their classes, see what activities groups had completed, and set access to different tools (Nelson et al., 2007).

In contrast, the *Virtual Performance Assessment* (VPA) was designed as a performance assessment to observe how learners investigate complex problems in realistic scenarios (Clarke-Midura & Dede, 2010), allowing for evaluation of higher-order thinking skills that are difficult to measure with traditional methods while scaling more effectively and reliably than traditional performance assessments. Baker and Clarke-Midura (2013) developed a model to predict the correctness of a student's conclusion as well as their skill in designing causal explanations using VPA. Based on a combination of a tree-based variable selection procedure, a linear regression model, and decision rules, their work was able to pinpoint essential information in the virtual world students must access in order to have a high probability of success. Specific interventions related to variables negatively correlated to success that may be candidates for just-in-time intervention were identified. Clarke-Midura and Yudelson (2013) developed a method to automatically learn rubrics based on student actions to determine if students draw certain conclusions based on evidence they have been exposed to. Actions were coded as supporting, rejecting, or being neutral to the five possible causal factors in the virtual ecosystem. While preliminary, their results again indicated potentially simple ways to analyze student actions in a complex problem space. Scalise (2017) applied a more complex model to the VPA data that

combines traditional item response theory with Bayesian networks. Utilizing the robust psychometric properties of the former with the flexibility of the latter, this model makes sense of "semi-amorphous data" that is not directly scorable or immediately interpretable to human coders.

After comparing VPA with several other case studies of formative assessment in immersive virtual environments, Code and Zap generated five directions for future research to empower teachers to do the following:

1) clarify and share learning intentions and criteria for success,
2) engineer effective classroom discussions and other learning tasks that elicit evidence of student understanding,
3) provide feedback that moves learners forward,
4) activate students as instructional resources for one another, and
5) empower students as the owners of their own learning (2017, p. 245)

While automated formative feedback was never a goal of VPA, this work nonetheless lays the groundwork for analyzing complex actions in ill-defined problem spaces via a variety of methods ranging from linear regression to cutting-edge hybrid psychometric models. Beyond calculating probabilities, these findings can guide teachers through inquiry-based instruction while also providing guidance to the student directly.

EcoMUVE is a multi-user virtual environment (MUVE)-based middle school ecosystem science curriculum designed to teach science content and understanding of complex causality over the course of ten classes spent exploring a virtual ecosystem (Metcalf, Kamarainen, Tutwiler, Grotzer, & Dede, 2011; Grotzer, Kamarainen, Tutwiler, Metcalf, & Dede, 2013). Two modules based around simulated virtual worlds are available to teachers to use: a pond

ecosystem that allows students to explore a fish kill scenario and a forest that introduces students to the role of predators in ecosystems over time. Both modules share the same overall goals of providing problem-based scenarios and empowering learners to conduct collaborative inquiry of ecosystems using simulated tools that real scientists would use. While exploring the virtual worlds and collecting data, students work in small groups to develop concept maps on paper showing the causal relationships between biotic and abiotic features in the ecosystems and are tasked to support their claims with evidence and reasoning. This framework aligned well with subsequent research by McNeill and Krajcik (2011) that popularized the Claims, Evidence, Reasoning framework. These concept maps represent a shared understanding of the problem and are presented to their peers at the end of the curriculum.

While using the virtual environments, a record of student actions is stored and uploaded to a secure server. Logged information is time-stamped and includes what zones students moved through in the world, what data they collected, any interactions with non-player characters (NPCs), use of a virtual field guide as a reference, and other similar data. Since all the EcoLEARN curricula are collaborative and team-based, these log files capture only part of the overall experience. Many studies of these curricula have been mixed methods in nature, investigating classroom audio and video records along with survey gains and log files. While this synthesis focuses solely on papers centered around log file data, several case studies on the EcoMUVE modules are outlined in Kamarainen, Metcalf, Grotzer, and Dede (2015).

Translating techniques that had performed well in ITS literature to immersive virtual environments, sensor-free affect detectors were built for EcoMUVE (Baker, Ocumpaugh, Gowda, Kamarainen, & Metcalf, 2014). By hand-coding ground truth labels of student affect via observations, models were built to automatically detect five affective states relevant to

education: boredom, confusion, delight, engaged concentration, and frustration. All five detectors performed better than chance, and all detectors highlighted the rate of student actions in the world as being an important feature for classification. Unlike prior work with affect detection in MUVEs, no additional surveys were required that would interrupt student immersion. Sensor-free affect detectors can also help from a design perspective by triangulating what features of the environment are strongly associated with boredom or frustration. Detectors such as these can be used as a basis for interventions to alleviate boredom or to ensure interventions do not occur during periods of engaged concentration when they would be potentially distracting. Research focusing on changes in student interest in science and their epistemic beliefs during the EcoMUVE curricula utilized pre-post surveys to model changes over time, but noted that more fine-grained log file data may reveal additional details and would help establish causal evidence for why these shifts take place (Chen, Metcalf, & Tutwiler, 2014; Chen et al., 2016).

To explore how the physical salience of the virtual data students collected might impact how they construct their explanations in the curriculum, Tutwiler (2014) investigated the longitudinal records of data collection in student log files to calculate the average salience of collected data, how much low-salience data was found, and the ratio of low-salience data to total. When compared to pre-intervention survey findings on student use of low salience data in constructing explanations, students with higher prior knowledge were slightly more likely to utilize low salience data, and students overall were more likely to find low salience data early on in the curriculum. By embedding tutorials and scaffolding into the world that direct student attention to this low salience data, non-obvious causes become more obvious and are attended to early in the activity. A large amount of variation between teachers was also observed in student log files despite high fidelity of implementation (Metcalf, Kamarainen, Grotzer, & Dede, 2013),

indicating that even small teacher variations in delivery of the curriculum may impact student data collection behaviors.

EcoMOBILE was designed as a hybrid curriculum that could blend the affordances of immersive virtual environments with those of augmented reality (AR). Students first completed the pond EcoMUVE module, then went on a real field trip to a physical pond to utilize actual water quality probeware to collect water quality data (Kamarainen et al., 2013). Utilizing mobile broadband devices, students were guided through the data collection procedure, then given an array of side activities to dive deeper into certain aspects of the pond ecosystem. To investigate the potential for transfer from a virtual environment to the physical world, as well to test how well that learning can take place in a physical setting first, a study was designed where one condition completed EcoMUVE then EcoMOBILE (as intended) while the other condition used EcoMOBILE prior to exposure to the virtual world (Grotzer et al., 2015). In addition to rich video data and interviews, log files from EcoMUVE were used to track student movement in the virtual world to compare their paths at the simulated pond and the real one. In the typical student path, students began their explorations of the virtual pond unsystematically prior to discovering the fish kill event, focused closely on the pond after discovering the event, then gradually moved their attention to causes at a distance that impacted the watershed. These paths mirror shifts towards expert understandings of ecosystems that focus on causes over greater spatial and temporal distances (Grotzer, Tutwiler, Dede, Kamarainen, & Metcalf, 2011). Students with EcoMUVE experience explored the real pond in a more targeted, systematic way than their peers, providing strong evidence of transferring aspects of what they learned.

The log files captured by the AR platform used in portions of the EcoMOBILE study recorded time-stamped GPS locations of students, what content they viewed, and any recordings

or notes they made in response to prompts. By calculating how much time students spend exploring and what field trip content they see, designers can pinpoint the most impactful content in their experiences and can consider how to adjust their activities to maximize learning gains in the amount of time allotted to the field trip. In a recent study of log files from an implementation of EcoMOBILE, survey gains were compared to the amount of time students spent freely exploring the world after collecting their initial data, what content they viewed, and the order in which it was viewed (Reilly, Kamarainen, Metcalf, Dede, & Grotzer, 2019). Time spent exploring had a large impact on learning gains, completing certain "learning quests" had disproportionately large effects on what students learned, and the order in which content was viewed was impacted by the physical arrangement of the field trip, which resulted in differential learning gains for different groups. These effects were unintentional and can inform designs of future AR activities. Quests that directly linked aspects of the physical pond to lessons learned in EcoMUVE were associated with higher learning gains and may have aided transfer sufficiently to forge stronger links between material in the two different activities.

EcoXPT was designed to be a second-generation MUVE-based curriculum that expanded the curriculum of EcoMUVE. In addition to observational and water quality data, EcoXPT gives students access to a suite of simulated ecosystem science experiments that empower students to test their hypotheses directly and establish stronger causal evidence (Grotzer et al., 2017; Grotzer et al., 2018). In addition, a new concept mapping tool was built into the MUVE that replaced the paper-based activity done in EcoMUVE. When switching from paper-based to electronic maps, students created larger maps with more connections and included less salient factors (Metcalf et al., 2018). Log files from EcoXPT capture all of the edits and modifications students make to

their maps and provide a scaffolded space for virtual evidence to be provided and for free response reasoning text to be entered.

**Stealth Assessment in EcoXPT**

The electronic concept maps generated in EcoXPT can be automatically graded, and initial work on comparing EcoMUVE concept maps with those from EcoXPT showed that EcoXPT groups made smaller, more accurate maps that focused on less salient, microscale aspects of the world like how bacteria and temperature affect dissolved oxygen (Reilly, Metcalf, Studwell, Dede, & Grotzer, 2019). Exploring sources of experimental evidence for claims (Metcalf, Reilly, Dede, & Grotzer, 2019) revealed that the experimental tools are a powerful means to support many causal claims that are otherwise difficult to justify. Some experiments such as the comparison tanks can provide evidence to justify a wide variety of claims, while others are much narrower in scope and only relate to a handful of possible claims. Emerging work utilizing natural language processing (NLP) techniques with student reasoning has been able to classify student use of causal language as reliably as an expert rater, enabling the potential to fully automate all aspect of this grading. Doing so would facilitate a powerful way to generate feedback based on student misconceptions and erroneous or unsupported claims.

In order to explore how the trajectory of a group's actions may relate to their learning gains, a method of plotting and analyzing groups' investigations was used on EcoXPT log file data (Reilly & Dede, 2019). Rather than focusing on specific subsequences that appear frequently or are potentially meaningful due to expert coding, this analysis considers the entire multivariate time series of logged events to estimate the rate at which groups conducted investigations in the world. By using principal component analysis to reduce the multivariate time series to a univariate, constantly increasing value, this value could be plotted versus time

spent in the virtual environment. By fitting a linear regression through each group's trajectory, a rough rate of investigation can be calculated for each group. When clustered by these rates via k-means, groups fell roughly into low, medium, and high rate clusters. While this study found that higher rates of investigation may be positively associated with higher learning gains, rates varied widely by teacher, and the relationship is likely more complex than "faster is better." These results may suggest scaffolds that can prevent groups from getting stuck (therefore increasing their investigation rate) could result in better learning.

In another study that looked at the overall group log file data from an EcoXPT implementation, deep learning algorithms were compared to traditional classifiers to determine how accurately predictions could be made regarding student performance on surveys as well as the quality of their concept maps based solely on their logged actions (Reilly & Dede, 2019). Post-scores on the science content, understanding causality, and experimental methods constructs were more accurately predicted by the long short-term memory (LSTM) neural network model compared to a support vector machine and Random Forrest, and predictions of concept map quality were also likewise higher. These accuracies were in the same range as similar work done on more constrained games for learning (Min et al., 2015).

In addition to work aimed at student-facing scaffolds, a digital dashboard was designed for teacher use while implementing EcoXPT (Reilly et al., 2018). Logged events were binned into five different categories: exploration of the virtual world, collection of observational data, analysis of collected data and notebook entries, running virtual experiments, and generating hypotheses via the concept map tool. The proportion of types of events occurring in a given time period is shown to the teacher. This time period could vary from 5 minutes to an entire class period. On different days of the curriculum, different actions are typically more prevalent. While

there is no one “right” way to tackle the ill-structured problem presented in EcoXPT, outliers on the graph may warrant teacher intervention to check for unproductive struggle or roadblocks. For example, one pair may be mainly exploring the world, while peer groups are mostly involved in experimentation and analysis. This may signify a search for a missing organism in the virtual world or potentially a pair of students who believe they have nothing left to do. Seeing these patterns gives instructors a snapshot of how classes are progressing through the activity and can direct their guidance to groups that may need it most. While not piloted with teachers in this form, the design of this project heavily inspired the teacher daily reports designed for this dissertation.

**Research Questions**

To investigate how formative feedback tools could be added to an existing open-ended virtual environment for science education, a modified version of the EcoXPT software was developed, incorporating a variety of adaptive tools that provide targeted feedback to students varying based on their progress and their demonstrated understanding of the science topics being taught. A pilot study of this modified virtual environment was conducted with seven science teachers across two middle schools in a suburban New England school district that had previously used the baseline version of EcoXPT. Student performance using this modified curriculum was compared with historical performance data on the baseline version (as described in detail in the Sample and Site section of Chapter 3), and how the newly added tools were used based on students’ log file data was analyzed. A teacher-facing daily report was also developed that summarized student progress in the curriculum and flagged groups for potentially requiring more attention during the next session. For each research question, an *a priori* hypothesis is

offered regarding how the data and models might show evidence of certain changes in student and teacher behavior during the study.

*RQ1: Through same-day automated analysis of students' activities, can opportunities be identified for teachers to provide timely guidance in ways that have the potential to deepen learning and motivation?*

Such opportunities for teachers may be found in three sources otherwise difficult for them to see (types of logged events, types of saved notes, and current state of the concept map) and reported in actionable ways via a daily report emailed to teachers the evening after their classes use the virtual world. By making these data visible and easily digestible for teachers, teachers are hypothesized to target their instruction in the next class period to address common issues and focus their attention on specific groups that appear to be struggling in unproductive ways. Evidence for deeper learning and motivation as a result of this intervention will come from teacher feedback from veteran teachers who have used the curriculum previously without such support.

*RQ2: What types of feedback (model-based, metacognitive, direct guidance) and what delivery methods (automatic or upon request), if any, show evidence of behavioral change (as measured by survey gains, logged actions, and user feedback) when delivered directly to students within the virtual world?*

The relative effects of different types of feedback can be seen on how they alter student logged actions immediately after delivery compared to how students historically have acted in certain scenarios. Meta-cognitive and model-based feedback upon request are hypothesized to have the largest impacts on behaviors, based on evidence from feedback literature in light of the open-ended nature of the EcoXPT task. Despite this, user feedback is hypothesized to state they

found the direct guidance prompts to be most helpful. Due to the design of the study (discussed in Chapter 3), it is difficult to separate the effects of any one intervention on overall survey gains. Students using the added LENS components are hypothesized to have larger gains on the affective and experimental methods portions of the survey, as those aspects of the curriculum are what most of the interventions target.

*RQ3: What features of the logged activities, if any, are informative for generating formative feedback for students and for teachers, and what analytical techniques best reveal these features?*

Certain key sequences of logged events are hypothesized to be most indicative of good learning, and the top performing quartile of groups (according to survey gains and concept map quality) is hypothesized to have significantly more of those sequences than groups in the lowest quartile. Transitions such as exploring to experimenting, exploring to analyzing, and exploring to analyzing to hypothesizing are hypothesized to be more common in the highest quartile, where sequences of actions can be interpreted as following a logic consistent with best practices in conducting scientific investigations. Transitions that cannot be as easily linked to best practices, such as exploring to exploring and hypothesizing to exploring, are hypothesized to be more common in the lowest performing groups. Sequential pattern mining and Markov models can reveal these patterns of transitions. and LSTM-based models can account for the sequential nature of the log file data when making predictions about student success.

# Chapter 3: Design and Methodology

## Sample and Site

Data for this study were collected in the Fall of 2019 from 595 7th-grade students across seven teachers and two schools in a suburban school district in New England. After filtering for signed permission slips, 587 students working in 304 groups remained in the sample (98.7% of students). Average middle school class size in the district is 21.7 pupils, with 15% of students eligible for free and reduced-price lunch, 18% of students receiving special education services, and 7% of students classified as English language learners. The student population of the district is 62% white, 19% Asian, 8% Hispanic or Latino, and 5% African American. Teachers were recruited from a list of teachers at schools that had previously utilized the baseline version of EcoXPT. Four teachers had used the baseline EcoXPT curriculum with their classes within the past two years, while the other three teachers were new to their schools and had not utilized the curriculum previously. Written consent was obtained from each student and teacher included in the dataset. Teachers were compensated $500 each for participating in a teacher professional development session regarding the new tools and modified curriculum, managing permission forms for all classes, implementing the 13-day curriculum, conducting the pre and post surveys with their classes, and being interviewed at the completion of the curriculum.

Historical data used in this dissertation came from two prior implementations of EcoXPT carried out in the Spring semester of 2018. One study focused on the relative efficacy of the EcoXPT world with and without access to experimental tools (Study #1). The other study compared EcoXPT to Environmental Detectives, a paper-based curriculum designed to teach environmental science via a detective approach (Study #2; Beals & Willard, 2001). Of the veteran teachers in the current LENS study, two participated in Study #1 and two participated in

Study #2. The two teachers who previously participated in Study #2 had also served as pilot users of the EcoXPT curriculum the previous year while it was still in development. As both of these studies required preparation of multiple curricula or versions of curricula, these veteran teachers had been asked to go beyond what would normally be required of a novice teacher of EcoXPT.

**Procedures**

Intended to closely match past iterations of the EcoXPT curriculum to facilitate comparisons (for more detail on the original curriculum, see Dede et al., 2017), classes progressed through a 13-day curriculum designed for middle school science classes. A rough breakdown of the daily schedule for the curriculum is provided in Appendix B. Several weeks prior to their planned start date, a one-hour professional development session is provided to teachers that explains the overall purpose of the curriculum, the virtual world, and the newly added features for this version. During that session, teachers are given permission forms for themselves and their students as well as other paperwork required to collect student demographic data. Prior to the start of the curriculum, the teacher is sent usernames for all students for use on the electronic survey as well as in the virtual world.

Lesson plans and accompanying presentations are provided for each day of the curriculum along with short videos and handouts that frame groups' investigations of the virtual ecosystem. Designed for 45-minute periods, most class sessions involve roughly 10 minutes of teaching and discussion along with 30 minutes of students using the virtual ecosystem. Prior to the first lesson of the curriculum, teachers distribute individual links to the pre-intervention survey, which is designed to take one 45-minute class period. Teachers then teach the lessons in order over the course of 3-4 week typically (most teachers using these curricula teach 3 or 4 of

the lessons per school week due to shortened periods on certain days, snow days, illnesses and absences, etc.) After the completion of the 13 lessons, teachers distribute individual links to the post-intervention survey. Soon thereafter, a member of the research team interviews each teacher.

The pedagogical intent of EcoXPT closely aligns with the PBL and ASI literature outlined in Chapter 2, where students are provided minimal direct guidance to investigate an authentic scientific scenario via tools and techniques that mimic how ecosystem scientists would approach the task. Instead, significant inherent structure is designed into the experience to invite students to grapple with scientific concepts such as uncertainty, puzzles about scale of interpretation, and information about the epistemic lenses that ecosystems scientists employ, for examples. Developed in consultation with many ecosystem scientists (Kamarainen & Grotzer, 2019), the curriculum is designed to have a "low floor" and "high ceiling" where groups of any ability level can make meaningful progress and discover different parts of the problem space and aspects of the virtual world. Students and teachers alike are strongly discouraged from considering the curriculum to be a linear activity with one "right" answer at the end.

**The LENS Tool Suite**

To infuse dynamic scaffolds and new types of supports into the virtual world, five student-facing features were added to the software portion of the EcoXPT unit. In addition, a daily teacher report was designed in consultation with teachers from previous classroom studies of EcoXPT. To address features of the second research question, these new features range across several types of feedback (model-based, metacognitive, direct guidance) and delivery methods (automatic or upon request). While all features are available to all students in the new sample,

analyses can reveal which are most used and which types teachers and students report to be the most helpful.

Two NPCs were modified to act as pedagogical agents that offer advice and feedback based on group progress through the curriculum and demonstrated understanding of the complex causal relationships at the pond. Both still offer the same functions they did for students in the baseline EcoXPT curriculum but additional functions have now been added to each. These agents act as e-coaches that can observe a group's progress in context and over time to deliver appropriate guidance (Kamphorst, 2017). Interest in the use of pedagogical agents has been driven by the need for scalable learning that engages students and emphasizes higher-order skills (Johnson & Lester, 2018).

Ranger Susan provides students with information about the pond from the perspective of a park ranger. In addition to her static information on the different virtual days of the activity, in the LENS modification to EcoXPT she delivers general advice to groups on what to do on a given day based on what tools are unlocked and what they have done thus far (Figure 1). This type of direct guidance is available upon request and is intended to echo instruction delivered by teachers in EcoXPT. The intention of this tool is to reduce the load on the instructor in the classroom, thereby freeing them to focus on groups who may be struggling with deeper issues than what to do next. What advice is delivered depends on what tools are currently unlocked (which is linked to specific days of the curriculum), as well as what logged events are present for different groups. In general, her advice guides students through the process of collecting data from all sources available, analyzing it to find patterns, building hypotheses in the concept map, and testing those hypotheses via experimentation.

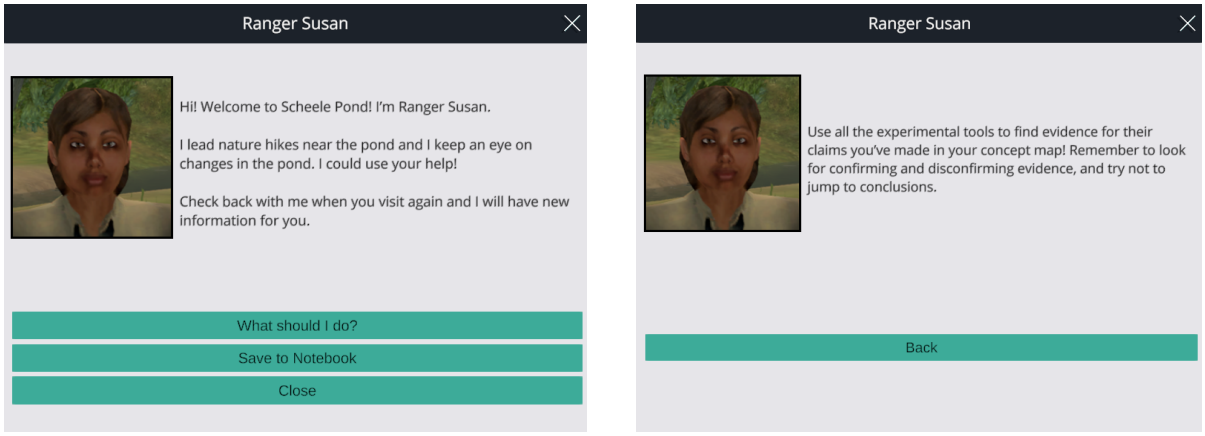



Figure 1. Example dialog with Ranger Susan.

The other NPC turned pedagogical agent is Dr. Jabir Hatami, an ecosystem scientist stationed at the lab building that houses the experimental tools. His previous role was to guide groups to the correct experimental tool to answer different types of questions. This is an example of the static scaffolding present in EcoXPT. For his expanded role in the modification for LENS, Dr. Hatami is now able to look over group concept maps and provide feedback based on the current state of the concept map (Figure 2). He evaluates concept maps in progress and identifies common errors without specific "next step" advice. The algorithm that selects advice looks for if nodes are present without connections, if known incorrect claims are present, if "Does Not Affect" claims are being made, whether evidence and reasoning are present, and if factors are connected to multiple other factors. This type of model-based, upon request feedback again echoes teacher instruction on causality and how to use the concept mapping tool.

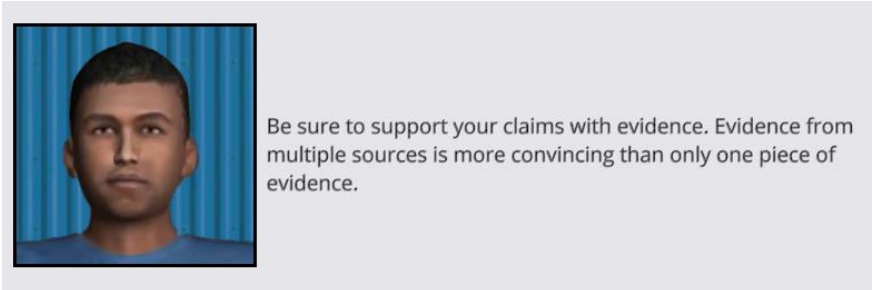



Figure 2. Example dialog with Dr. Jabir Hatami.

Two additional supports were added to the comparison tank tool, which allows students to set up experiments to explore what effect different features of the pond ecosystem have on measurable factors. The first support is a worked example tutorial for the tool that shows students how to use the user interface as well as how to frame a question, to set up an experiment to answer questions, and to save the results (Figure 3). In EcoXPT, all tutorials for the experimental tools take the form of multiple screenshots of the tool with a large volume of accompanying text. Analyses of log files from historical EcoXPT data shows that groups look at tutorial messages for an average of 32.5 seconds, which may be insufficient time to read and deeply process the information presented. By automatically providing this direct guidance to all groups upon encountering the tool, groups are introduced to all aspects of the tool that are necessary to use it effectively. In order, the tool shows students how to fill the tanks with water, how to add factors to the water to address an example question, how to measure an attribute of the tanks, and how to save their data to their notebook.

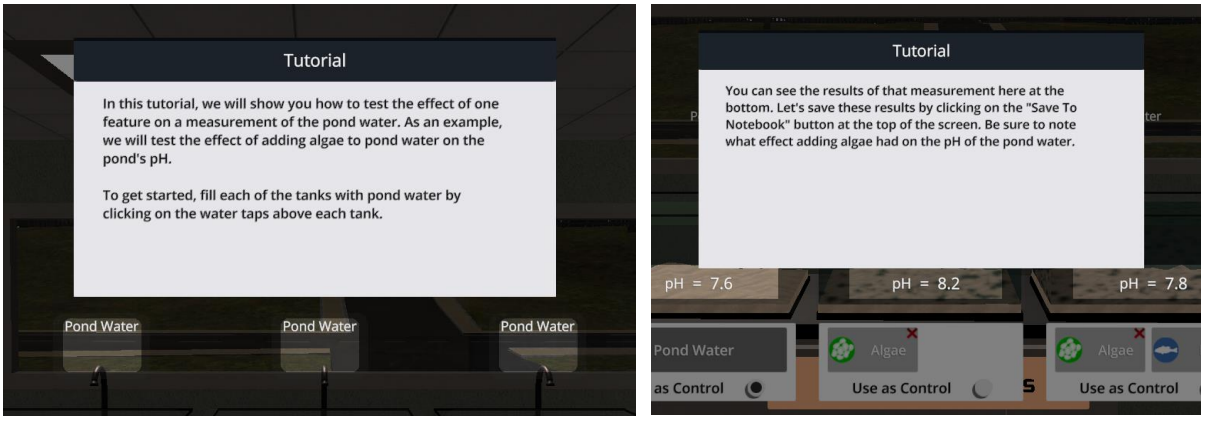



Figure 3. Examples of the comparison tank worked example tutorial.

The other modification to the comparison tank tool is less structured and requires students to use help-seeking behaviors to request metacognitive support. In lieu of repeating the tutorial if having difficulty with the comparison tank tool, students can click a new help button

that frames student questions in terms of causality to highlight the importance of experimental evidence to establish causal relationships (Figure 4). When a student asks for help, a new window appears where students can frame their investigations by selecting what effect adding a factor to the pond water might have on a measureable quantity. The advice echoes classroom instruction on experimentation strategy, such as varying one thing at a time and writing explanations for saved experimental results.

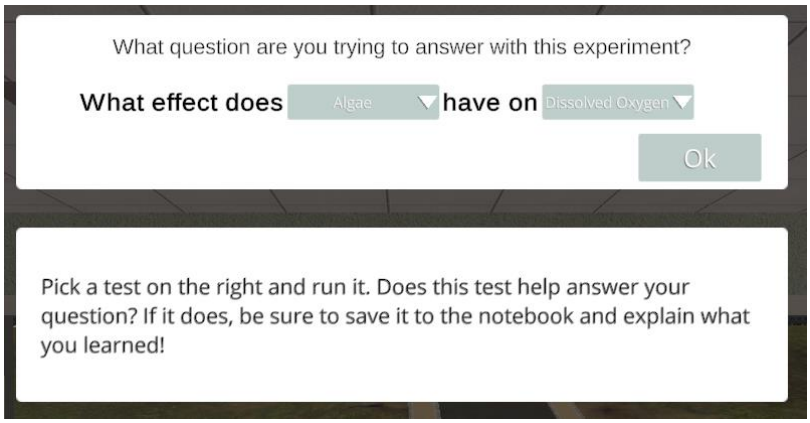


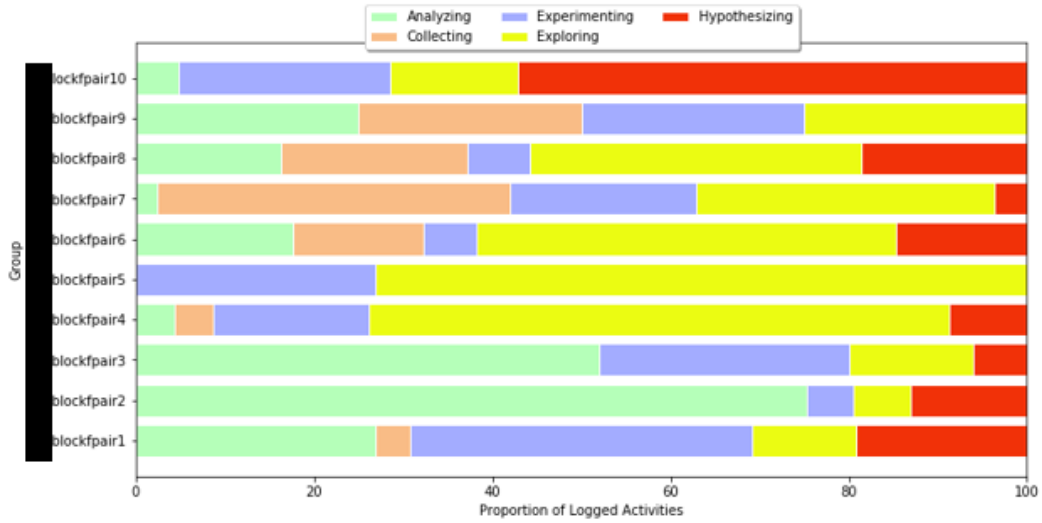

Figure 4. Example of the comparison tank support tool.

The final student-facing support added to the EcoXPT virtual world is a new type of notebook entry specifically to record reflections electronically within the virtual world. Aligned with self-explanation and reflection literature as outlined in Chapter 2, this support provides an automatic, meta-cognitive chance for groups to pause their investigations and think deeply about the nature of the problem they are investigating, how their investigations have gone thus far, and what unanswered questions they have. The standard EcoXPT curriculum provides similar reflection questions as exit ticket activities or optional homework activities. These reflections have not previously been separately analyzed by researchers. In the LENS modifications, additional opportunities for reflection were created for each day of the curriculum, and time in the lesson plans was included at the end of each lesson.

The teacher-facing tool designed for the LENS suite extends the dashboard prototype developed for EcoXPT (Reilly et al., 2018) into a daily report for teachers. Feedback solicited from teachers regarding the real-time dashboard prototype indicated that it would be unlikely to be used due to increasing their cognitive load and introducing another potential technology to troubleshoot. The daily report utilizes the same visualization designed for the dashboard, but presents it as a static view of how groups spent their time over the course of an entire period (Figure 5). While not useful for detecting off-task behavior in real time, this visualization can aid teachers in planning how to allocate their time among groups during the next session. Additionally, a series of actionable observations from the groups' log files are provided beneath the visualization to address certain issues and outliers that can be automatically detected. Between two to four pieces of advice are provided to teachers based on the contents of their students' notebooks and their current concept maps. Advice is contextualized to the day of the curriculum and is generally intended to highlight features that have been missed by most of the class or groups that have not yet completed tasks that are typical by a certain day. Advice centers around common errors in concept mapping, underutilization of certain tools, and progress relative to others in the class. Student use of certain tools in open-ended games for science learning have been shown to influence content-specific gameplay decisions and correspond to a higher likelihood of success (Cheng et al., 2017).

[redacted] Block F – 12/10 Report

**Types of Logged Activities:**




**Notebooks:**

Some groups have a small amount of tolerance tank results relative to others (1, 9, 10)

Encourage groups to use the weather simulator

**Concept Maps:**

Some groups have several incorrect claims in their concept maps (2, 8)

Some groups have included Does Not Affect claims in their maps (1, 2)

Some groups have not used evidence in their maps (3)

NPC testimonial and the field guide are used infrequently as evidence

Figure 5. Example of the teacher daily report (teacher name and initials redacted).

At the conclusion of a school day, all log files from that day were downloaded and run through a Python script to generate visualizations and identify potential features to report to teachers. While this process may be fully automated in the future, a researcher selected which advice to report and manually compiled the reports during this implementation. Reports were then emailed to teachers that evening to assist with their lesson planning. In this implementation, five reports were sent to teachers over the twelve days that students used the virtual environment.

Logged events for groups are very similar prior to the unlocking of the concept map tool, and advice became repetitive when generated every day. At the end of the curriculum, teachers were interviewed about their use of the reports. Due to technical issues at one school site that required a significant amount of teacher attention to troubleshoot, the daily reports were only utilized with the three teachers at the other school site.

**Data Sources and Measures**

Three types of quantitative data were analyzed for this study (individual pre-post surveys, log files from groups of 2-3 students using the software, and group concept maps), and qualitative interviews were conducted with all teachers after completion of the curriculum. Demographic data on students was also collected from teachers. While survey data is at the individual level, the unit of analysis for this study is at the group level. Any models or feedback tools designed for similar virtual environments used in computer-supported collaborative learning environments must be able to make group-level recommendations without access to pre-survey data that would not normally be collected outside of a research setting. Measures of group learning were calculated by taking the average of the group members' normalized learning gains for each survey construct and by applying a rubric to the group's concept map to evaluate their understanding of the complex causal scenario at the virtual pond.

**Assessment Instrument**

Individual students took a pre- and post-assessment designed for EcoXPT as part of a blended assessment strategy to determine the efficacy of the curriculum (Thompson et al., 2016). Assembled from a mixture of pre-validated instruments from science education literature (as cited below by construct), the survey contained six constructs: affective dimensions, ecosystem science content, understanding of causality, correlation versus causation, use of experimental

methods, and epistemology of science. The assessment (see Appendix A) was comprised of a variety of multiple choice and Likert items designed and tested by educational researchers, psychometricians, and an ecosystem scientist. The survey was administered electronically via Qualtrics during science classes immediately preceding and following the implementation of the curriculum. The survey was designed to detect medium sized effects (Cohen's $d > 0.5$; Cohen, 1977) at the 0.05 alpha level with a sample of roughly 200 student pairs and is thus appropriately powered for this sample.

Aspects of the assessment have been validated in earlier studies, including a measure of students' complex causal assumptions (as reported in Tutwiler et al., 2016). The Cronbach's Alpha from the final version of that construct was 0.71. The construct validity of the causality subscales was established via confirmatory factor analysis. The construct assessing students' understanding of correlation versus causation was designed by experts in causal systems and ecosystems to tease apart how observed relationships between observed factors may or may not indicate the presence of a causal relationship between the two. Designed as several scenario-based questions that were aligned and unaligned with the content of the virtual world, the Cronbach's Alpha from this construct was 0.78. The self-efficacy sub-construct of the affective construct was designed and validated for a prior virtual world-based ecosystem science curriculum (Chen et al., 2014). Questions in the science identity sub-construct came from a study of deep engagement in a high school biology curriculum (Pugh et al., 2010). The science interest sub-construct was taken from the interest/enjoyment subscale of the Intrinsic Motivation Inventory (Deci et al., 1994). The Cronbach's Alpha from the final version of the affective construct was 0.91.

The content construct focused on photosynthesis, respiration, biotic and abiotic factors,

decomposition, and food web dynamics. Questions were drawn from the previous EcoMUVE and EcoMOBILE content surveys (Metcalf et al., 2011; Kamarainen et al., 2013), as well as from the Causal Patterns in Science Project (Grotzer, Basca, & Donis, 2002), the National Assessment of Educational Progress (NAEP) and American Association for the Advancement of Science (AAAS). Face and content validity were established by ecosystems content knowledge experts. The Cronbach's Alpha from the content construct was 0.71.

Students' understanding of the use of experimental methods in ecosystem science was determined by presenting a set of scenario-based, Likert-scale questions that explore when to use different experimental methods. Designed by an ecosystem scientist and piloted during the spring of 2016, the final version of the experimentation construct attained a Cronbach's Alpha of 0.74. Questions measuring students' understanding of the nature of science and their epistemological beliefs were modified from the Epistemological Beliefs Assessment for Physical Science (EBAPS) scale designed and validated by Elby, Frederiksen, Schwarz, and White (1997). Questions were modified to better match the reading ability and linguistic level of middle school students. This version of the epistemology construct attained a Cronbach's Alpha of 0.70.

**Log Files**

During a group's use of EcoXPT, 68 unique logged events can be recorded by the software and are saved during every session in a PostgreSQL database on a server located on the Harvard University campus. An example of the raw data format is shown in Figure 6. Each row represents one event logged by one user in the virtual world. Each event is tagged with its anonymous identifier, what anonymized class it is a member of, the type of event, any relevant details (e.g., settings for an experiment or what dialog option was selected), and a timestamp. To reduce the grain size of the data for several analyses, events were binned into five different meta-

categories that encompassed the most common problem-solving behaviors seen in the world, as shown in Table 1. These categories were inspired by work done to analyze the learning pathways of novice programmers (Berland et al., 2013).

| userId | courseId | moduleId | lessonId | activityId | type | data | timestamp |
|---|---|---|---|---|---|---|---|
| 3477 | 104 | 2 | 2 | 13 | Event | END ECOMUVE XPT ACTIVITY | 2017-10-30 10:57:45.846607-04 |
| 3477 | 104 | 2 | 2 | 13 | closeWeatherSimulation | Closing Weather Simulation | 2017-10-30 10:57:43.658474-04 |
| 3477 | 104 | 2 | 2 | 13 | startWeatherSimulation | *NULL* | 2017-10-30 10:57:40.816657-04 |
| 3477 | 104 | 2 | 2 | 13 | resetWeatherSimulation | Resetting Weather Simulation | 2017-10-30 10:57:40.317989-04 |
| 3477 | 104 | 2 | 2 | 13 | closeTutorialEvent | Weather Simulation Tutorial | 2017-10-30 10:57:38.758321-04 |
| 3477 | 104 | 2 | 2 | 13 | OpenWeatherSimulation | Opening Weather Simulation | 2017-10-30 10:57:36.287072-04 |
| 3477 | 104 | 2 | 2 | 13 | openTutorialEvent | Weather Simulation Tutorial | 2017-10-30 10:57:36.285386-04 |
| 3477 | 104 | 2 | 2 | 13 | zoneEntered | Trailer | 2017-10-30 10:57:33.486894-04 |
| 3477 | 104 | 2 | 2 | 13 | openingDialogue | {<br>"objectId":"Dr. Jabir Hatami",<br>"timePeriod":"June 30"<br>} | 2017-10-30 10:57:23.685864-04 |
| 3477 | 104 | 2 | 2 | 13 | Event | START ECOMUVE XPT ACTIVITY | 2017-10-30 10:57:09.971421-04 |

Figure 6. Raw data format of EcoXPT log files.

Table 1. Classification of logged events.

| Category | Examples of logged events included |
|---|---|
| Explore | Moving between zones in the world, changing time period |
| Collect | Gathering population and water quality data, taking pictures of organisms |
| Analyze | Graphing data, editing and saving notebook entries |
| Experiment | Running virtual experiments, placing experimental tools |
| Hypothesize | Constructing a causal map of the different factors at the pond |

**Concept Maps**

Concept maps are saved as part of a large JSON file (the user activity state) that serves as the saved state of the software for each group. The concept map portion saves what connections are present, what evidence is linked to each claim, and what reasoning is provided. As part of the work done with EcoXPT concept mapping, a series of 16 claims were selected as core aspects of

the fish kill scenario (Reilly et al., 2019). In that paper, a series of known incorrect claims were also identified (e.g., fertilizer affects fish) based on commonly seen misconceptions in early group concept maps (see Table 2). To facilitate the automatic evaluation of concept maps, the ability for students to create their own nodes was disabled. Claims 4, 8, and 11 are therefore not possible to make in the modified LENS version. In addition to the presence of evidence, the source of the evidence can be automatically evaluated. Sources of evidence for the concept map include experimental tools, observations, graphed data, the field guide, and NPC testimonial. Most causal claims are better supported by experimental evidence versus the correlational evidence available from other sources (i.e., the data graphing tool), but the quality of student experiments varies (Metcalf et al., 2019). Any reasoning is saved as raw strings of text.

Table 2. Concept map claims.

| Claim | Description |
|---|---|
| 1 | Source (e.g., housing dev, people) affects fertilizer |
| 2A | Rain washes fertilizer into pond (runoff). |
| 2B | Wind causes fertilizer to get into the pond. |
| 3A | Fertilizer affects N, P in the pond |
| 3B | N,P in the pond affect algae (nutrients) |
| 3C | Fertilizer affect algae (3A+3B combined) |
| 4 | Lack of nutrients causes algae to die** |
| 5A | Algae affects dead matter (decomposition) |
| 5B | Dead matter affects bacteria (food) |
| 5C | Algae affects bacteria (5A+5B combined) |
| 6 | Algae affects dissolved oxygen (photosynthesis) |
| 7A | Clouds/sunlight affect algae (photosynthesis) |
| 7C | Clouds/sunlight affect dissolved oxygen (via algae) |
| 8 | Night affects dissolved oxygen (via algae) ** |
| 9 | Wind affects dissolved oxygen (mixing from the air) |
| 10 | Dead matter/algae increases turbidity |
| 11 | Less sunlight (due to turbidity) decreases algae ** |
| 12 | Bacteria affects dissolved oxygen (respiration) |
| 13 | Temperature affects dissolved oxygen |

Table 2 (Continued)

| | |
|---|---|
| 14 | Dissolved oxygen affects fish (respiration) |
| 15 | Big fish and minnows all affect each other |
| 16 | Big fish and herons affect each other |
| DNA | Does not affect claim (can be correct or incorrect) |
| OTH | Correct claim, but not one of the core numbered claims above |
| INC | Incorrect claim |

** = will only appear if a pair makes their own node

A rubric was designed to assess the overall completeness and quality of the causal concept maps (partially adapted from McNeill & Krajcik, 2011). The overall score is out of 6 possible points (three levels for each of the three dimensions) and equally emphasizes the role of claims, evidence, and reasoning (Table 3). This structure aligns with the base explanation rubric for evaluating CER (McNeill & Krajcik, 2008) and is intentionally kept to only three levels to facilitate automatic scoring and classification of artifacts. This scoring is not based on this specific problem scenario nor the number of claims possible in EcoXPT.

Table 3. Concept map rubric.

| Category | Level 0 | Level 1 | Level 2 |
|---|---|---|---|
| Claim | Contains incorrect claims | Claims are correct but represent 6 or less core claims | Claims are correct and represent 7 or more core claims |
| Evidence | Evidence is not consistently used (<50% of connections) or does not match the claim | Appropriate but insufficient evidence used | Appropriate and sufficient evidence used |
| Reasoning | Reasoning is not provided | Reasoning provided but does not discuss causal mechanism | Reasoning discusses causal mechanism |

The presence of any incorrect claims drops the quality for claims to 0, as this indicates a map in the emerging stages of finding support or failing to find support for their claims.Many EcoXPT concept maps contain some errors as is inherent to authentic scientific investigation, but the incorrect claims the rubric checks for (such as pH affecting the fish) have sources of

disconfirming evidence in the virtual world. While a rubric intended for summative assessment would not penalize students for having any incorrect claims, this scoring system does want to detect such errors easily as target feedback may be helpful for those groups. In this way, the rubric was aligned to be aligned with the pedagogical intent of the additional LENS components; it is unaligned with the pedagogical intent of EcoXPT in which errors are considered inherent to the process of science. Claims level 1 indicates an emerging understanding of the causal relationships at work, while level 2 represents a more complete understanding. Evidence quality can be assessed automatically by checking on the consistency of evidence use, exploring how much evidence was used for claims, and assessing whether experimental evidence is used for claims that rely upon it (e.g., those having to do with dissolved oxygen). The presence of causal mechanistic language in student writing has been the subject of prior qualitative coding work on EcoXPT student reflections and can be automated via natural language processing.

To explain the use of the rubric, two examples are provided here. One group currently has a concept map that contains incorrect claims, provides appropriate and sufficient evidence for correct claims, and discusses causal mechanisms in most reasoning statements. The overall score for that map would be the sum of the scores for each aspect of the rubric. In this case, 0 points are awarded for claims, 1 point is awarded for evidence, and 2 points are awarded for reasoning, resulting in a total score of 3. In a different concept map, many core claims are made, appropriate but insufficient pieces of evidence are used, and reasoning is provided but not in a way that addresses causal mechanisms. This map would yield a claim score of 2, an evidence score of 1, and a reasoning score of 1 for a grand total of 4.

### Qualitative Interviews

Due to time constraints and logistical concerns, teachers were interviewed in a semi-structured fashion in focus groups at each school location. Questions focused on efficacy of the new tools, what "stuckness" looks like while using the software, how teachers intervene when they detect unproductive struggle, and what could be added or modified to ease this process (see Appendix C for the interview protocol). Interviews took one 45-minute period as soon after the conclusion of the curriculum as possible. Interviews were audio recorded with participant permission and transcribed. Thematic analysis was used to assign preliminary codes to participant responses from which central themes in all interviews were identified. These themes served as triangulation for the efficacy of the student-facing features and as a primary data source for assessing the utility of the teacher-facing report.

## Data Analysis

Survey responses collected by Qualtrics were downloaded and manually cleaned to ensure student entered usernames were correctly entered to facilitate matching pre and post responses. These data were also automatically cleaned to ensure that no students without signed permission slips accidentally used the electronic survey. Analysis of the survey data was conducted in R version 3.5.2 with the use of the psych, psychometric, nlme, and stargazer libraries as well as "tidyverse" suite of tools (namely ggplot2, dplyr, tidyr, and stringr). Survey data was analyzed via a series of ANOVA models and Tukey's honest significant difference tests to investigate differences by condition and by teacher. Additionally, a two-level random intercept hierarchical linear model (students nested in classes) was built for each survey construct to control for the nested nature of the data as well as student- and class-level covariates. An example of the two-level model is as follows:

Level 1 - Individual Growth

$$PostScore_{ij} = \beta_{0j} + \beta_{1j}PreScore_{ij} + \beta_{2j}ELL_{ij} + \beta_{3j}IEP_{ij} + \beta_{4j}Engagement_{ij} + \beta_{5j}Reading_{ij} + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, \sigma_y^2)$$

Level 2 - Student growth

$$\beta_{0j} = \gamma_{00} + \beta_{01}Veteran_j + \beta_{02}LensTools_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \beta_{11}Veteran_j + \beta_{12}LensTools_j + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + \beta_{21}Veteran_j + \beta_{22}LensTools_j + u_{2j}$$

$$\beta_{3j} = \gamma_{30} + \beta_{31}Veteran_j + \beta_{32}LensTools_j + u_{3j}$$

$$\beta_{4j} = \gamma_{40} + \beta_{41}Veteran_j + \beta_{42}LensTools_j + u_{4j}$$

$$\beta_{5j} = \gamma_{50} + \beta_{51}Veteran_j + \beta_{52}LensTools_j + u_{5j}$$

$$\begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \\ u_{4j} \\ u_{5j} \end{pmatrix} \sim N\left[ \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \tau_{00} & \cdots & \tau_{05} \\ \vdots & \ddots & \vdots \\ \tau_{50} & \cdots & \tau_{55} \end{bmatrix} \right]$$

These models also converged when specified as three-level models (students in groups in classes) and can incorporate group-level features (such as those from log files) but these have not been used in this dissertation.

Log files were manually cleaned for one school that suffered technical issues at the beginning of their implementation that resulted in many shorts segments of use due to connectivity problems. These logged events that did not correspond to productive time using the software were removed. Log file and user activity state data were analyzed in Python 3.7 with the primary use of the pandas, numpy, json, matplotlib, and scikit-learn packages. Some Python

packages and analytical frameworks have been developed for use with educational log file data, but the structure of the EcoXPT log files was not aligned with their schema (Hao et al., 2016). All neural networks were built and trained using Keras with the Microsoft Cognitive Toolkit (CNTK) backend, and all Markov models were visualized with the NetworkX package.

# Chapter 4: Findings

## Differences in Student Outcomes

### Survey Trends

Of the 587 students whose legal guardians granted permission to analyze their data, 484 students completed a pre-survey and 535 completed a post-survey. After keeping only students who completed both the pre- and post-survey, the survey dataset is comprised of 395 students across the two schools that participated in the current study as well as historical data from 179 students from those same schools who participated in EcoXPT two years earlier. During teacher recruitment for this study, it was assumed that there would be no significant differences in prior knowledge between students of teachers who participate. To test this assumption, a series of pairwise t-tests with a Bonferroni correction were conducted on all teachers' pre-scores across all survey constructs. No teachers were found to have students with significantly different pre-scores on any survey construct. These results were confirmed with a Tukey's honest significance difference test. Within teachers, none of their classes were significantly different from each other on pre-survey performance across the six constructs. Simple ANOVA models were used to test if pre-scores across constructs were equal on expectation across conditions. Classes from the EcoXPT historical data had significantly higher pre-scores on content (0.73 versus 0.69, $F = 5.93$, $p = 0.015$) but no other pre-scores differed significantly by condition.

Students in both datasets showed significant gains on all six constructs. Table 4 shows an overview of survey scores and gains for both conditions with all construct scores rescaled to be percentages of possible points. According to mixed model ANOVA for each construct, the only construct that differed significantly between the groups is epistemology ($F(568) = 5.571$, $p < 0.05$), where groups in the LENS classes gained 1.1 more percentage points. According to mixed

model ANOVA model for survey gains by teacher, teacher assignment did have a significant impact on gains in this construct (F = 3.073, p = 0.0035). A Tukey's honest significance difference test revealed that teacher JG's mean gain on this construct was significantly higher than four other teachers' classes (JA, RB, SC, and EJ).

Table 4. Survey gains with statistical significance and effect sizes by condition.

| **EcoXPT:** | | | | | | |
|---|---|---|---|---|---|---|
| Constructs | Pre % | Post % | Gain % | t(df) | p value | Cohen's d |
| Affect | 58.42 | 60.52 | 2.46 | 3.05(165) | $p < 0.005$ | 0.15 |
| Exp Methods | 76.23 | 82.46 | 6.24 | 6.31(175) | $p < 0.0001$ | 0.52 |
| Content | 73.34* | 75.69 | 2.63 | 2.27(170) | $p < 0.05$ | 0.12 |
| Corr vs. Caus | 61.60 | 70.48 | 9.32 | 6.40(173) | $p < 0.0001$ | 0.44 |
| Causality | 60.67 | 63.10 | 2.45 | 2.25(171) | $p < 0.05$ | 0.26 |
| Epistemology | 43.84 | 46.74 | 2.90 | 3.38(175) | $p < 0.0005$ | 0.26 |
| **LENS:** | | | | | | |
| Constructs | Pre % | Post % | Gain % | t(df) | p value | Cohen's d |
| Affect | 59.50 | 62.19 | 2.87 | 4.89(376) | $p < 0.0001$ | 0.21 |
| Exp Methods | 74.67 | 80.32 | 5.63 | 9.26(391) | $p < 0.0001$ | 0.53 |
| Content | 69.23 | 73.16 | 4.11 | 4.44(384) | $p < 0.0001$ | 0.21 |
| Corr vs. Caus | 63.18 | 74.65 | 11.62 | 11.06(388) | $p < 0.0001$ | 0.60 |
| Causality | 60.17 | 64.66 | 4.50 | 7.15(383) | $p < 0.0001$ | 0.55 |
| Epistemology | 45.68 | 49.63*** | 3.95*** | 6.06(394) | $p < 0.0001$ | 0.37 |

$^{*}$ $p<0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p<0.005$

Due to the size of the sample, exploring effect sizes is more meaningful than statistical significance. As defined by Cohen (1977), d = 0.2 should be considered a small effect size, 0.5

indicates a 'medium' effect size, and 0.8 a large effect size. The effect sizes reported here are more specifically referred to by Cohen as $d_s$, meaning the difference of means in terms of the pooled within-group standard deviation observed for both interventions (Lakens, 2013). The values are generally small across most constructs with several breaking into the “medium” size classification. While the simplified cut-off values are of dubious value (Baguley, 2009), the effect sizes of these tests are presented alongside the p values as a means of comparing the impact of these treatments to other findings in the literature that use different instruments.

To explore how gains across these constructs differ based on condition assignment and on student- and teacher-level covariates, a series of two-level random intercept multilevel models (students in classes) predict construct post-scores based on pre-scores for all six constructs. Student-level covariates pertaining to reading level (below/on/above grade level), and teacher estimates of student engagement in science (low/medium/high), English language learner (ELL) status, and the presence of an Individualized Educational Plan (IEP) or a 504 plan were included. High reading level and high engagement are the reference categories in the models, thus coefficients are only shown for low and medium categories in both of those features. A class-level indicator of whether or not the teacher had ever previously implemented EcoXPT was also included. Models including the identity of the teacher as a factor often failed to converge, and thus this factor has been omitted from these analyses. Details for this series of models are shown in Table 5.

Table 5. Summary of multilevel models of survey performance.

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | method post | content post | affect post | Causality Post | corrcaus post | epistemology post |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| methods_pre | 0.423*** | | | | | |
| | (0.051) | | | | | |
| content_pre | | 0.503*** | | | | |
| | | (0.042) | | | | |
| affect_pre | | | 0.671*** | | | |
| | | | (0.043) | | | |
| causality_pre | | | | -0.158** | | |
| | | | | (0.065) | | |
| corrcaus_pre | | | | | 0.389*** | |
| | | | | | (0.044) | |
| epistemology_pre | | | | | | 0.262*** |
| | | | | | | (0.041) |
| LENS tools | -0.004 | -0.010 | 0.010 | 0.014 | 0.024 | 0.033*** |
| | (0.011) | (0.018) | (0.011) | (0.009) | (0.018) | (0.010) |
| ELL | -0.0004 | -0.016 | 0.026 | 0.027 | -0.147*** | 0.017 |
| | (0.031) | (0.046) | (0.032) | (0.028) | (0.052) | (0.027) |
| IEP | 0.015 | -0.057*** | 0.004 | -0.064*** | -0.077*** | -0.040*** |
| | (0.016) | (0.022) | (0.016) | (0.013) | (0.024) | (0.013) |
| Engagement (L) | -0.040** | -0.083*** | -0.040** | -0.065*** | -0.084*** | -0.061*** |
| | (0.017) | (0.025) | (0.018) | (0.015) | (0.027) | (0.015) |
| Engagement (M) | -0.035*** | -0.016 | -0.027** | -0.014 | -0.019 | -0.028*** |
| | (0.011) | (0.016) | (0.011) | (0.009) | (0.017) | (0.010) |
| Reading (L) | -0.040** | -0.062*** | -0.017 | -0.043*** | -0.062** | -0.072*** |
| | (0.017) | (0.025) | (0.017) | (0.014) | (0.026) | (0.014) |
| Reading (M) | -0.015 | -0.046*** | -0.003 | -0.027*** | -0.022 | -0.024** |
| | (0.012) | (0.016) | (0.011) | (0.010) | (0.018) | (0.010) |
| Veteran | -0.009 | -0.003 | -0.015 | -0.019* | -0.027 | -0.024** |
| | (0.012) | (0.017) | (0.012) | (0.010) | (0.018) | (0.010) |
| Constant | 0.531*** | 0.457*** | 0.245*** | 0.786*** | 0.548*** | 0.414*** |
| | (0.041) | (0.040) | (0.030) | (0.040) | (0.039) | (0.024) |

Table 5 (Continued)

| | | | | | | |
|---|---|---|---|---|---|---|
| Observations | 481 | 470 | 459 | 470 | 476 | 484 |
| Log Likelihood | 384.634 | 232.697 | 387.74 | 469.551 | 182.721 | 466.066 |
| Akaike Inf. Crit. | -743.267 | -439.395 | -749.47 | -913.103 | -339.442 | -906.133 |
| Bayesian Inf. Crit. | -688.981 | -385.409 | -695.79 | -859.117 | -285.291 | -851.766 |

*Note:* $^{*}$ p<0.05, $^{**}$ p < 0.01, $^{***}$ p<0.005

Gains on the epistemology construct did differ significantly by condition, with students assigned to the LENS group earning 3.3-percentage points more on the post survey than their EcoXPT peers when controlling for all other student-level covariates ($t(188) = 3.19$, $p = 0.0035$). Epistemology gains were sensitive to student reading level, engagement, and IEP/504 status when controlling for all other student-level covariates. Students below grade reading level earned 7.2-percentage points less on the post survey versus students above reading level ($t(188) = -4.86$, $p < 0.0001$), and students at grade reading level earned 2.4-percentage points less on the post survey versus students above reading level ($t(188) = -2.34$, $p = 0.020$). Students rated as low engagement earned 6.1-percentage points less on the post survey versus highly engaged students ($t(188) = -4.13$, $p < 0.0001$) and those rated as medium engagement earned 2.8 percentage points less ($t(188) = -2.77$, $p = 0.0061$). Students with an IEP or 504 plan earned 4.0-percentage points less than their peers ($t(188) = -2.96$, $p = 0.0034$). Students in classes taught by veteran EcoXPT teachers earned 2.6 percentage points less than their peers when controlling for all other factors ($t(188) = -2.55$, $p = 0.016$). Two-way interactions between all factors were tested but none were found to be significant.

While no other construct's gains differed by the presence of the LENS tools, Table 5 shows that engagement was a significant predictor of survey performance across all constructs when controlling for all other factors. Reading level of students had almost as uniform a presence in most models, with all construct models except the one for affect including it. The presence of

IEP/504 plans was also significant on the content and correlation versus causation construct, while ELL status was significant on that latter construct as well. The model to predict performance on the causality construct was the only other model where teacher veteran status was significant.

To explore which teachers may have had an impact on the significance of the "veteran" designation, a Tukey's honest significance difference test was used to test all pairwise comparisons of teacher gains on the epistemology and causality constructs. Despite not differing significantly on pre-survey scores, one novice teacher's students' gains (JG) were significantly higher than four others (see Table 6). Three of those four teachers are veteran EcoXPT instructors. On the causality survey, one novice teacher's students' gains were higher than those of two veteran teachers.

Table 6. Differences in survey gains by teacher.

| **Epistemology** | | | |
|---|---|---|---|
| Teachers | Veteran? | Difference (in percentage points) | p value |
| JG vs. EJ | No/Yes | 10.45 | 0.015 |
| JG vs. JA | No/Yes | 11.25 | 0.0071 |
| JG vs. RB | No/Yes | 12.41 | 0.0018 |
| JG vs. SC | No/Yes | 14.53 | 0.00065 |
| **Causality** | | | |
| Teachers | Veteran? | Difference (in percentage points) | p value |
| SH vs. JA | No/Yes | 6.78 | 0.018 |
| SH vs. RB | No/Yes | 7.69 | 0.0034 |

### Concept Map Quality

Overall concept map quality was determined by summing quality scores for claims, evidence, and reasoning as outlined in Table 3 in Chapter 3. The coding procedure was designed to be automatic. Cut points for different thresholds of quality were picked without human intervention and with no reference to specific claims or pieces of evidence present only in this curriculum. This type of automated coding makes the coding more applicable to other curricula.

The purpose of this procedure is to score concept maps to generate formative feedback and drive teacher interventions, not to assign a final quality score or grade at the end of the curriculum. In addition to aligning with existing CER rubrics, the trinary split of quality scores allows machine learning classifiers to make accurate predictions in ways they could not with continuous data and provides simple bins for where different groups are facing issues.

The claim coding procedure first set the claim score to 0 if any incorrect claims were present, then tallied the number of core claims described in Table 2. If that count was less than 7 (half of all possible core claims), claim quality was set at 1. Otherwise, claim quality for that pair was set at 2. The mean claim score for students in the LENS condition was 0.53 (SD = 0.66) indicating a large number of student maps still included at least one incorrect claim.

The evidence coding algorithm first calculated the ratio of total evidence used per number of claims for each student group. If this ratio was less than one, that group was assigned an evidence score of 0. Each piece of evidence used was then labeled by source (experimental, reference, data/pattern, observation, or testimony) to estimate the appropriateness of the evidence for particular claims. Lists of claims were generated that benefited from particular types of evidence. For example, claims related to the role of weather on dissolved oxygen levels could be justified well by experimental evidence from the weather simulator and from graphing weather data alongside dissolved oxygen levels (though this is more correlational than causal evidence.) A relevance metric was calculated for each concept map where groups earned one point every time they used an appropriate source of evidence to justify a claim. The median relevance value for all concept maps was 7, with a minimum of 0 and a maximum of 35. Groups with a relevance metric below 7 were assigned a 1 for evidence score, while those higher were given a 2. The

mean evidence score for LENS students was 1.23 (SD = 0.85) indicating that a large number of concept maps did use evidence well to justify claims.

The reasoning scoring algorithm first calculated a reasoning ratio similar to that used by evidence scoring, where the ratio of reasoning statements to claims made was calculated for all groups. The mean reasoning ratio for all concept maps was 0.73, with ratios ranging from 0 to 1. If this ratio was less than 0.73 (i.e., less than 73% of claims had reasoning included) then those groups were assigned a reasoning score of 0. To determine the frequency of causal language in group reasoning statements, natural language processing was used to evaluate if reasoning statements used causal language. A set of 467 hand-coded reasoning statements from EcoXPT concept maps was used as the basis of this coding. Reasoning text was vectorized to transform text to a matrix of token counts, then the count matrix was run through a term-frequency times inverse document-frequency (TF-IDF) transformer to scale down the effect of commonly used words and focus on more meaningful ones. A linear support vector machine (SVM) model with stochastic gradient descent (SGD) training was trained on 70% of the available labeled data, with parameters optimized via grid search with 5-fold cross validation. The SVM model achieved 82.8% accuracy on the test set of data. While some of the claims in the concept map are only supported by correlational evidence, most have causal experimental evidence available and a manual investigation revealed that the vast majority of meaningful reasoning statements from students do include causal language. This model was then used to classify each reasoning statement as containing causal language or not. If less than half of the reasoning statements contained causal language, the group's reasoning score was a 1, while concept maps with higher levels of causal language were assigned a 2. The mean reasoning score for LENS students was 0.92 (SD = 0.80) with a nearly even split between all three scores.

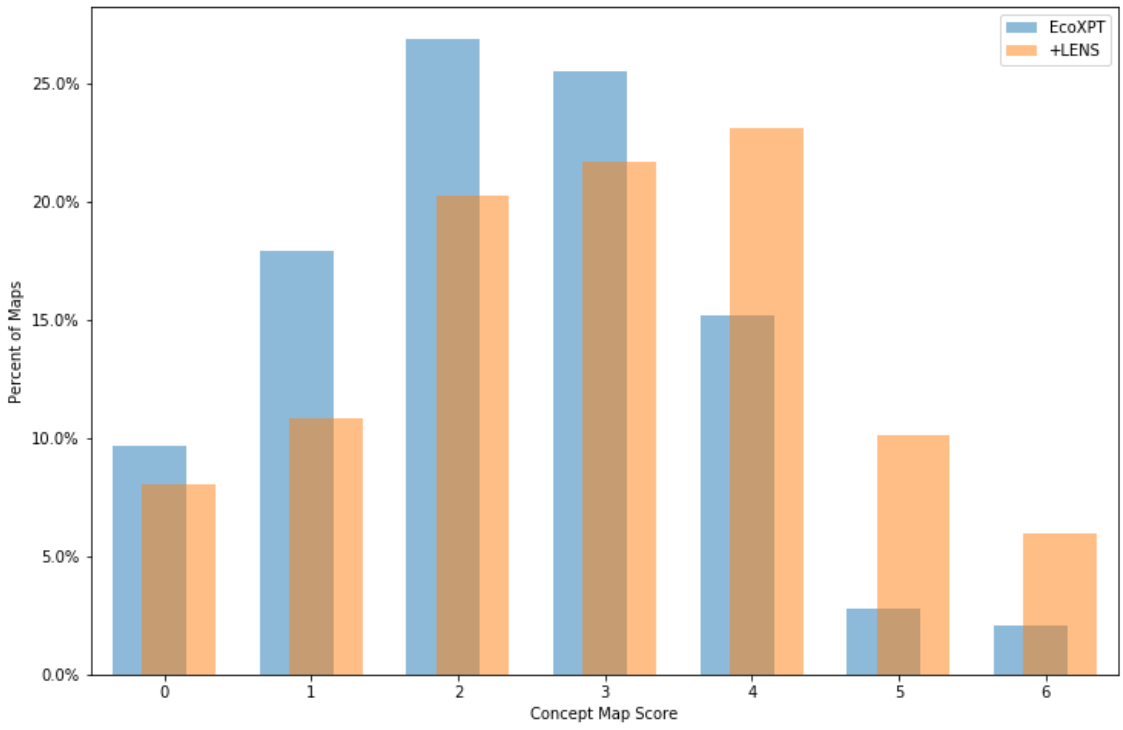



Figure 7. Distribution of concept map quality scores by condition.

To generate the final concept map quality score, the claim, evidence, and reasoning scores were added together. As an assessment of the validity of these scores, concept map scores were correlated with normalized learning gains across all six constructs as shown in (1). As the

$$\text{normalized learning gain} = \frac{\text{post}-\text{pre}}{100-\text{pre}} \quad (1)$$

survey scores are at the individual level, average normalized learning gain was calculated for each group. Both evidence scores and total scores were positively correlated with average group normalized learning gain on the causality construct ($r = 0.17$, $p = 0.0082$; $r = 0.17$, $p = 0.0091$). The distributions of scores for both LENS classes as well as those school's prior EcoXPT classes are shown in Figure 7. LENS classes (mean = 2.95, SD = 1.59) had significantly higher concept map scores than EcoXPT classes (mean = 2.35, SD = 1.37) according to a Welch's *t*-test ($t = 4.05$, $p < 0.0001$). This variant of a *t*-test accounts for the unequal variance and sample size

between the groups. Within each subcategory, LENS classes had higher average claim scores (0.77 vs 0.54, $t = 3.08$, $p = 0.0022$) and evidence scores (1.23 vs 0.89, $t = 4.17$, $p < 0.0001$) than baseline EcoXPT classes, although average reasoning scores did not differ significantly (0.95 vs 0.92) between conditions.

In order to test whether the rubric features that were aligned or unaligned with the pedagogical intent of each condition, and to assess any potential impact of the rubric scoring breakpoints on the significance of these differences by condition, raw counts and ratios can also be treated continuously and tested between groups. When analyzing these data, LENS classes had higher average core claim counts (3.45 vs 2.07, $t = 4.07$, $p < 0.0001$) and evidence relevance scores (8.32 vs 4.80, $t = 5.92$, $p < 0.0001$) than baseline EcoXPT classes. These findings make sense given the intended support in the LENS condition for figuring out the underlying connection in the causal dynamics of the eutrophication scenario as designed into the technology. Average ratios of evidence and reasoning use as well as the ratio of causal language used in reasoning statements did not differ significantly between conditions.

**Sequential Pattern Mining and Markov Models**

To explore how groups moved through the virtual world, Markov models were used to calculate transition matrices for groups to see if the increased prevalence of certain transitions was indicative of better understanding or a superior investigatory strategy. For these analyses, a subset of data from days 6 through 8 of the curriculum was used, as all features were unlocked by that point and development of the concept map did not yet dominate the majority of the logged activities. Unsupervised learning was first used to identify different types of behaviors exhibited by groups in their logged actions. Next, transition matrices from groups in different

quartiles of learning gains and concept map quality were compared to detect any differences in behavior between those groups.

A 6x6 transition matrix was calculated for each group then flattened into one dimension. These vectors were then appended to each other, resulting in a dataframe with one row per group and 36 features corresponding to the 36 different possible transitions. The k-means algorithm was used to cluster the data, with an optimal number of two clusters selected by silhouette analysis. State transition diagrams are plotted for each cluster in Figure 8 and reported in Table 7. The largest differences are that groups in cluster 0 will go from hypothesizing to seeking feedback (while groups in cluster 1 never do) and are much more likely to go from hypothesizing to experimenting and are not as prone to go from exploring to hypothesizing. Despite these differences, cluster membership is not significantly associated with normalized learning gain on any construct nor is it associated with concept map quality. Means of groups in each cluster did not differ significantly by normalized learning gains on any construct or by concept map quality.

Table 7. Transition matrices by cluster.

| **Cluster 0:** | | | | | | |
|---|---|---|---|---|---|---|
| | Analyzing | Collecting | Experimenting | Exploring | Hypothesizing | Feedback |
| Analyzing | 0.75 | 0.06 | 0.08 | 0.08 | 0.00 | 0.03 |
| Collecting | 0.14 | 0.44 | 0.04 | 0.33 | 0.04 | 0.01 |
| Experimenting | 0.20 | 0.03 | 0.54 | 0.18 | 0.03 | 0.02 |
| Exploring | 0.05 | 0.23 | 0.11 | 0.60 | 0.00 | 0.01 |
| Hypothesizing | 0.10 | 0.15 | 0.20 | 0.34 | 0.16 | 0.06 |
| Feedback | 0.09 | 0.02 | 0.03 | 0.06 | 0.00 | 0.79 |
| **Cluster 1:** | | | | | | |
| | Analyzing | Collecting | Experimenting | Exploring | Hypothesizing | Feedback |
| Analyzing | 0.74 | 0.07 | 0.09 | 0.08 | 0.01 | 0.02 |
| Collecting | 0.17 | 0.46 | 0.04 | 0.30 | 0.01 | 0.02 |
| Experimenting | 0.23 | 0.03 | 0.52 | 0.17 | 0.04 | 0.01 |
| Exploring | 0.06 | 0.20 | 0.13 | 0.59 | 0.01 | 0.01 |
| Hypothesizing | 0.04 | 0.11 | 0.70 | 0.08 | 0.06 | 0.00 |
| Feedback | 0.10 | 0.02 | 0.04 | 0.07 | 0.00 | 0.78 |

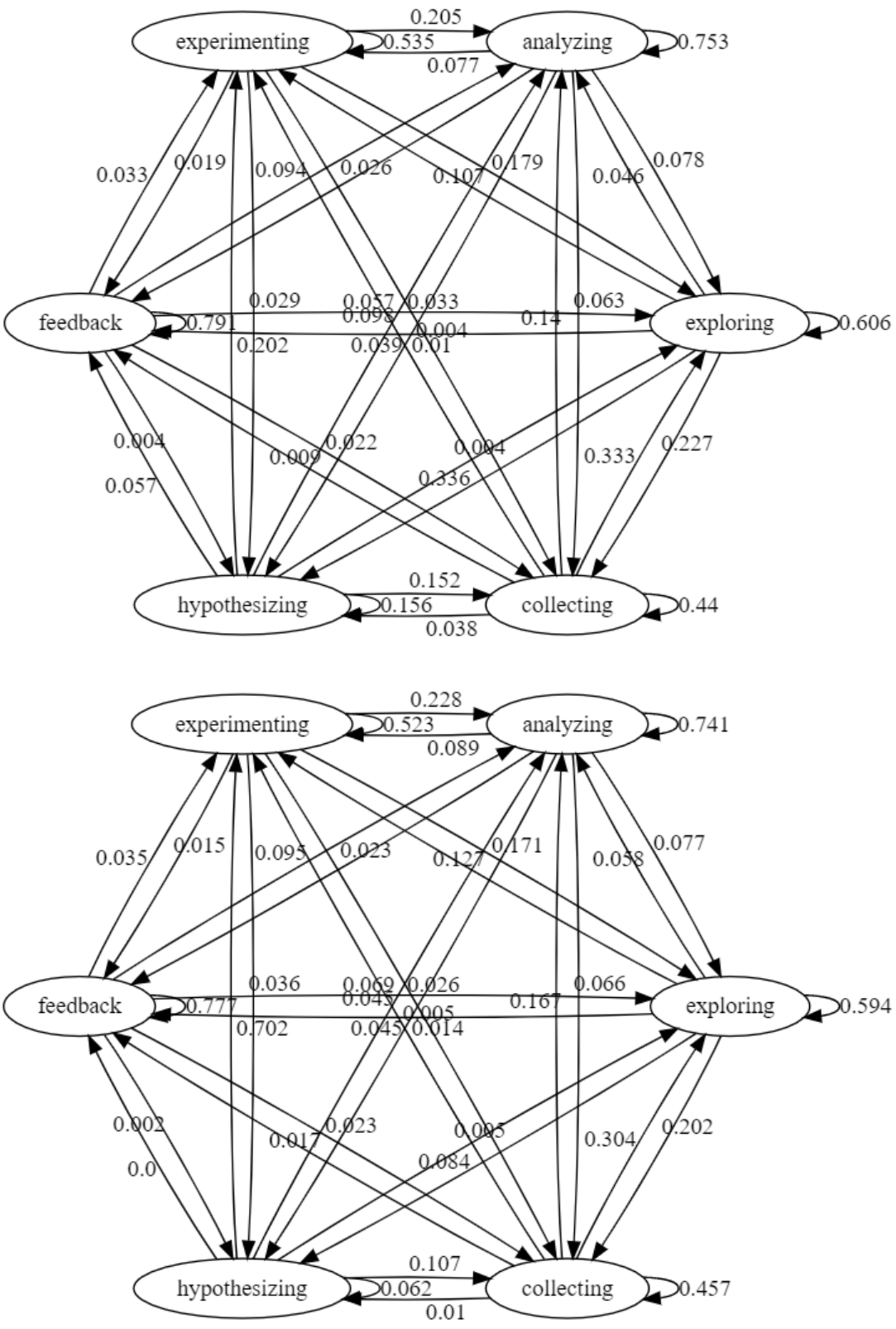



Figure 8. Markov models of clusters 0 (top) and 1 (bottom) of LENS classes.

To further explore potential connections between normalized learning gains, concept map quality, and transition matrices, subsets of the data were constructed for groups in the highest and lowest quartiles for normalized learning gains on the epistemology construct. As survey data are at the individual level, yet these transitions are at the group level, average normalized learning gains were calculated for each group. State transition matrices were plotted for each quartile, but no major differences were observed. In a similar fashion, no large differences were observed between transition matrices from the highest quartile concept map quality groups and the lowest.

Actions are categorized into different meta-categories at the log line level, although not all actions are included. Actions corresponding to adjusting settings of tools or logging UI actions such as closing a dialog box are not included, while events corresponding to meaningful investigations or reflection are categorized. For example, simply selecting an experimental tool is not categorized as an "experimenting" action, but running the experiment is. Likewise, adjusting the zoom in the submarine tool is not classified as a meaningful action, but measuring the population value of a microscopic creature is categorized as "collecting." These distinctions are meant to reduce uninformative events and avoid all processes becoming a series of the same n events in sequence (e.g., to use an experiment one might need to always click the tool, adjust the settings, and run the experiment). In the current system, meaningful actions take place an average of 24 seconds apart (SD = 39.7).

To gain detail on what specific transitions might be more characteristic of a group that is performing well or poorly, sequences of actions for groups in the lowest and highest concept map quality quartiles were extracted and examined. The most frequent 2- and 3-grams

(sequences of two or three logged actions) were extracted from group log file activity over the same days six through eight time period as in the previous analysis. The most common sequences are reported in Table 8 with n-grams consisting of only one meta-category omitted (e.g., (‘exploring’, ‘exploring’, ‘exploring’)). The most common 2-grams are virtually identical between the two groups, so little differentiation is possible.

Table 8. Most common sequences of logged activities by concept map quality.

| Low Quartile 2-grams | High Quartile 2-grams |
|---|---|
| ('exploring', 'collecting') | ('exploring', 'collecting') |
| ('collecting', 'exploring') | ('analyzing', 'exploring') |
| ('analyzing', 'exploring') | ('collecting', 'exploring') |
| ('exploring', 'analyzing') | ('collecting', 'analyzing') |
| ('collecting', 'analyzing') | ('exploring', 'analyzing') |
| ('analyzing', 'collecting') | ('analyzing', 'collecting') |
| ('exploring', 'experimenting') | ('exploring', 'experimenting') |
| ('experimenting', 'exploring') | ('experimenting', 'analyzing') |
| ('experimenting', 'analyzing') | ('experimenting', 'exploring') |
| ('analyzing', 'experimenting') | ('analyzing', 'experimenting') |
| ('hypothesizing', 'exploring') | ('collecting', 'experimenting') |
| ('collecting', 'experimenting') | ('experimenting', 'collecting') |
| ('exploring', 'hypothesizing') | ('hypothesizing', 'exploring') |
| ('hypothesizing', 'analyzing') | ('analyzing', 'hypothesizing') |
| ('analyzing', 'hypothesizing') | ('exploring', 'hypothesizing') |

| Low Quartile 3-grams | High Quartile 3-grams |
|---|---|
| ('exploring', 'exploring', 'collecting') | ('analyzing', 'analyzing', 'exploring') |
| ('analyzing', 'analyzing', 'exploring') | ('collecting', 'exploring', 'collecting') |
| ('collecting', 'exploring', 'collecting') | ('exploring', 'exploring', 'collecting') |
| ('exploring', 'collecting', 'exploring') | ('exploring', 'collecting', 'collecting') |
| ('collecting', 'exploring', 'exploring') | ('exploring', 'collecting', 'exploring') |
| ('exploring', 'collecting', 'collecting') | ('collecting', 'analyzing', 'analyzing') |
| ('exploring', 'analyzing', 'analyzing') | ('analyzing', 'collecting', 'analyzing') |
| ('collecting', 'collecting', 'exploring') | ('exploring', 'analyzing', 'analyzing') |
| ('analyzing', 'exploring', 'exploring') | ('analyzing', 'exploring', 'exploring') |
| ('collecting', 'analyzing', 'analyzing') | ('collecting', 'collecting', 'exploring') |
| ('analyzing', 'exploring', 'analyzing') | ('analyzing', 'exploring', 'analyzing') |
| ('analyzing', 'collecting', 'analyzing') | ('analyzing', 'analyzing', 'collecting') |
| ('exploring', 'exploring', 'analyzing') | ('collecting', 'exploring', 'exploring') |
| ('exploring', 'exploring', 'experimenting') | ('experimenting', 'analyzing', 'analyzing') |
| ('analyzing', 'analyzing', 'collecting') | ('analyzing', 'analyzing', 'experimenting') |

When examining the 3-grams, many of the analyzing, exploring, and collecting activities seem similar. Differences begin to reveal themselves, however, when focusing on 3-grams that include experimentation. The high quartile's high incidence of ('experimenting', 'analyzing', 'analyzing') may indicate attempts to rectify experimental evidence with correlational evidence observed in plotted weather and water quality data. Likewise, the ('analyzing', 'analyzing', 'experimenting') 3-gram that occurs almost as often shows that, while groups are still in the experimental trailer (since an 'explore' event would mark their movement back to the outdoor space), they observed the correlations between their different data and then attempted to test hypotheses or answer questions these data raise. While one of the main goals for EcoXPT is to engage in epistemologically authentic experimentation (which includes that conducted in the ecosystem), currently 5 out of 6 experimental tools are located inside the trailer.

In the lowest quartile, the most common 3-gram involving experimentation is ('exploring', 'exploring', 'experimenting'), meaning a significant amount of time was spent solely moving through the world before running an experiment. While it is possible to experimentally test relationships that are observed while exploring the world, exploration activities are typically logged less frequently by days 6, 7, and 8 of the curriculum. Behaviors such as these may indicate a lack of focus to investigations, but more possibilities are considered in the Discussion chapter. The lower quartile group has a higher prevalence of exploring events in the top 15 3-grams reported in Table 8.

To investigate if any 3-grams were significantly correlated with the outcome measures from the survey and concept map evaluation, a count vectorizer was used on each group's logged events to generate a "bag-of-3-grams" with a row for each group and features for each 3-gram observed in the entire dataset with counts of the occurrence of that 3-gram for each group. These

counts were then correlated with average normalized learning gain and concept map quality. As this analysis involved exploring correlations with 207 unique combinations of 3-grams observed in the data, it was very likely that a high number of correlations would be below the alpha threshold of 0.05. Instead of selectively analyzing these relationships further and running the risk of "p-hacking" (Nuzzo, 2014), these relationships can be viewed in light of the hypotheses made in Chapter 2. Significant correlations are reported in Table 9.

Table 9. Significant correlations with outcome measures.

| Outcome Measure | 3-grams | Correlation (p-value) |
|---|---|---|
| Concept Map Quality | analyzing analyzing experimenting | 0.220 (0.0017) |
| | collecting feedback feedback | 0.187 (0.0035) |
| | experimenting analyzing analyzing | 0.196 (0.0049) |
| | collecting exploring collecting | 0.172 (0.0074) |
| | exploring collecting exploring | 0.155 (0.017) |
| | experimenting experimenting feedback | 0.139 (0.0184) |
| | exploring collecting collecting | 0.146 (0.0225) |
| | collecting collecting exploring | 0.135 (0.0371) |
| | analyzing experimenting analyzing | 0.136 (0.0399) |
| | analyzing hypothesizing experimenting | 0.130 (0.0439) |
| | collecting feedback hypothesizing | -0.187 (0.0016) |
| | exploring experimenting feedback | -0.146 (0.0244) |
| Content normalized learning gain | experimenting collecting exploring | 0.136 (0.034) |
| | exploring experimenting feedback | -0.185 (0.0051) |
| | experimenting hypothesizing exploring | -0.152 (0.0203) |
| | hypothesizing exploring exploring | -0.141 (0.0344) |
| Causality normalized learning gain | exploring analyzing experimenting | 0.174 (0.0087) |
| | analyzing collecting hypothesizing | 0.162 (0.0128) |
| | hypothesizing experimenting analyzing | 0.151 (0.0203) |
| | exploring collecting collecting | 0.141 (0.0294) |
| | collecting collecting analyzing | 0.138 (0.0334) |
| | collecting collecting exploring | 0.133 (0.0394) |
| | collecting exploring collecting | 0.131 (0.0418) |

Table 9 (Continued)

| | | |
|---|---|---|
| Causality normalized learning gain | feedback hypothesizing collecting<br>experimenting hypothesizing hypothesizing<br>exploring exploring exploring<br>exploring experimenting collecting<br>experimenting experimenting hypothesizing | -0.308 (<0.0001)<br>-0.160 (0.0121)<br>-0.152 (0.0168)<br>-0.144 (0.028)<br>-0.134 (0.0339) |
| Affect normalized learning gain | exploring experimenting exploring<br>exploring hypothesizing analyzing<br>hypothesizing hypothesizing collecting<br>exploring experimenting hypothesizing<br>hypothesizing exploring exploring<br>experimenting exploring experimenting<br>exploring exploring exploring | -0.178 (0.0068)<br>-0.151 (0.0207)<br>0.144 (0.0278)<br>-0.141 (0.0299)<br>-0.140 (0.031)<br>-0.139 (0.0352)<br>-0.134 (0.0432) |
| Epistemology normalized learning gain | analyzing experimenting hypothesizing<br>hypothesizing analyzing experimenting<br>hypothesizing experimenting experimenting<br>experimenting hypothesizing hypothesizing<br>experimenting experimenting hypothesizing<br>hypothesizing collecting analyzing<br>experimenting hypothesizing experimenting<br>hypothesizing hypothesizing experimenting<br>exploring collecting experimenting<br>exploring hypothesizing analyzing<br>analyzing exploring exploring | 0.222 (0.0016)<br>0.151 (0.0153)<br>0.157 (0.0177)<br>0.180 (0.0208)<br>0.154 (0.0306)<br>0.130 (0.0353)<br>0.129 (0.0384)<br>0.147 (0.0441)<br>0.112 (0.0458)<br>-0.152 (0.021)<br>-0.142 (0.0389) |

For the counts of 2-grams, ('exploring', 'experimenting') and ('exploring', 'analyzing') were not more common in high quartile groups, but ('hypothesizing', 'exploring') was more prevalent in the lowest quartile groups. The specific 3-gram ('exploring', 'analyzing', 'hypothesizing') is not observed among the significant correlations with any outcome measures. While certain n-grams were highlighted in the hypotheses, the driving factor is whether the n-grams show an indication that groups are conducting scientific investigations in an authentic, effective way. The other major way to tease meaning out of these findings is to compare these sequences to the ground truth of classroom observations of more or less effective groups, though this is not pursued here.

Many positively correlated 3-grams from Table 9 can be mapped onto a logical sequence of events that would happen in an effective investigation. For example, ('analyzing', 'experimenting', 'hypothesizing') was correlated positively with gains on the epistemology construct and it likely means a group looked at correlational relationships in observational data, ran an experiment to generate evidence, then made or edited a claim in their concept map. Conversely, a negatively correlated 3-gram such as ('exploring', 'hypothesizing', 'analyzing') suggests that a group moved about the world, edited their concept map, then began looking at data that had been collected beforehand. While this may be a productive series in some cases, the overall negative association indicates that this is likely not an ideal chain of events. The large amount of negatively correlated 3-grams with normalized learning gain on the affective construct largely involve exploring, which should not be a major type of event in the latter half of the curriculum.

Feedback events are present in six significantly correlated 3-grams across concept map quality, content gains, and causality gains. ('collecting', 'feedback', 'feedback') and ('experimenting', 'experimenting', 'feedback') were both positively correlated with concept map quality, while ('collecting', 'feedback', 'hypothesizing') and ('exploring', 'experimenting', 'feedback') were negatively correlated. Parsing the role of feedback in these analyses is difficult as it is unclear what feedback tool was used and for what purpose. Additionally, the intention of these feedback tools was that they would be used more often by struggling learners, thus it is not surprising that some feedback uses are more common among groups with weaker concept maps. ('exploring', 'experimenting', 'feedback') is negatively associated with content gains and ('feedback', 'hypothesizing', 'collecting') is negatively associated with causality gains. Further analyses of n-grams containing which feedback tool was used may shed more light on these

patterns. Additionally, confirming evidence from classroom observations would help ground these findings and ensure they are meaningful.

**Predicting Success**

In an attempt to replicate the ability to predict the quality of student concept maps from log file data in EcoXPT (Reilly & Dede, 2019), classification models were used to predict the tertile of overall concept map quality for all LENS groups. LSTM-based models can account for temporal dependencies in time series data, and the log file data was transformed into a 3-dimensional tensor with dimensions of [samples, time step, features]. Support vector machine (SVM) and random forest (RF) models are not able to account for the temporal order of logged activities, so these models were fit on the cumulative counts of the same data. Two-thirds of the groups were randomly selected to act as a training set, while the remaining third of groups acted as the test set. 5-fold cross-validation was used to tune all model parameters and hyperparameters.

To reduce overfitting on a small number of groups, a simple LSTM model consisting of three bidirectional layers with 5 neurons each was built, with a dropout of 0.5 to on the first layer. The third LSTM layer output to a fully connected layer with a softmax activation that made the class assignments. The model utilized the Adam optimizer (learning rate = 0.01, epsilon = 0.01) with a categorical cross-entropy loss function. The model could train for up to a maximum of 100 epochs with an early stopping condition to halt training when validation loss did not decrease for two consecutive epochs. Class weights were provided during model fitting to combat the class imbalance in the training data.

In prior work with EcoXPT data, an LSTM-based model achieved the highest predictive accuracy of concept map quality with 48.5% accuracy, while the RF model achieved a higher F1

score due to having higher precision (Table 10). With the addition of the feedback features generating new logged events, all models are able to make more accurate predictions. The accuracy of all models for the LENS condition classes are similar, although the SVM model slightly outperforms the LSTM model. The SVM model also has the highest recall (avoiding false negatives), while the RF model has a superior precision (avoiding false positives). Taken together as an F1 score (the harmonic mean of precision and recall) as defined by (2), all classifiers are again similar, but the RF model is slightly more optimal.

$$\text{F1 score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (2)$$

Table 10. Performance of concept map quality classifier by condition.

| **EcoXPT:** | | | | |
|---|---|---|---|---|
| | Accuracy | Precision | Recall | F1 Score |
| SVM | 37.9% | 0.392 | 0.379 | 0.371 |
| RF | 43.9% | **0.462** | 0.455 | **0.442** |
| LSTM | **48.5%** | 0.339 | **0.485** | 0.399 |
| **LENS:** | | | | |
| | Accuracy | Precision | Recall | F1 Score |
| SVM | **66.3%** | 0.440 | **0.663** | 0.529 |
| RF | 62.1% | **0.493** | 0.621 | **0.535** |
| LSTM | 65.6% | 0.430 | 0.656 | 0.520 |

## Use of New Features

### Amount of Use by Type

Use of pedagogical agent-based feedback tools varied between different NPCs and over time throughout the curriculum. Groups asked Ranger Susan for feedback an average of 2.56 times during the curriculum (SD = 2.60), with some groups never using the feature and one

group utilizing it a maximum of 17 times. Jabir Hatami was consulted by groups an average of 2.02 times during the curriculum (SD = 2.20), with a minimum use of 0 and a maximum of 12 times. The two pedagogical agents were used differently over the course of the curriculum, with Ranger Susan being used heavily on Day 2, then not being consulted as frequently while Jabir Hatami was used more consistently after the experimental tools were unlocked (see Figure 9). This is not surprising as Dr. Hatami is not present in the virtual world until experimental tools are unlocked, resulting in no interactions with him prior to that day of the curriculum. A more detailed breakdown of use by teacher is provided in Table 11.

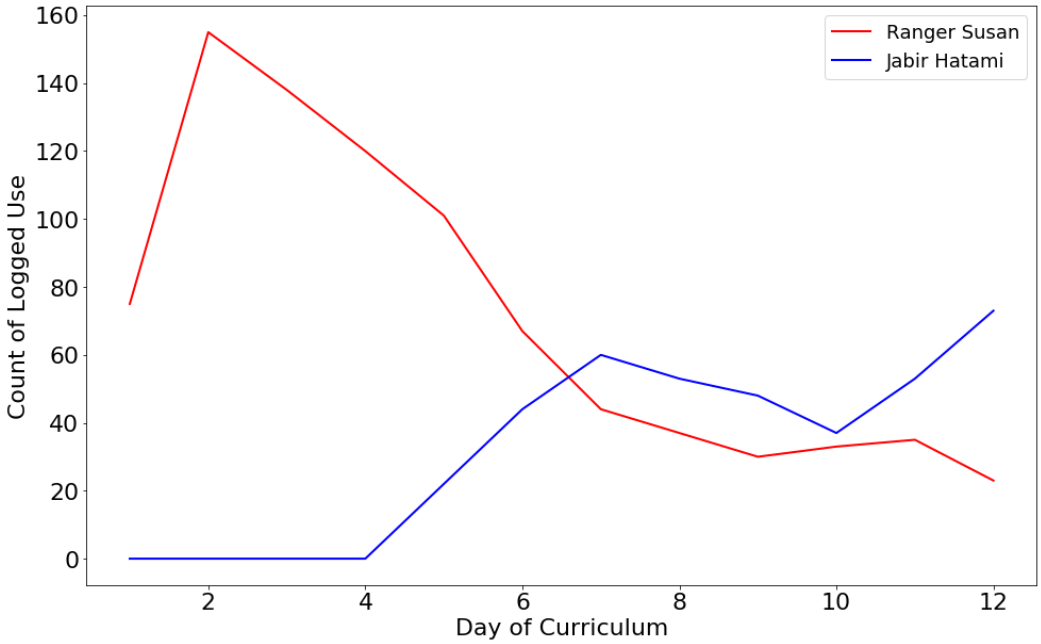



Figure 9. Use of the pedagogical agents by day of curriculum.

Average use of new features by teacher is provided in Table 11. Classes in School 2 appear to have used the pedagogical agents much more than their peers in School 1. All groups made use of the comparison tank tools and thus completed the interactive tutorial at least once, resulting in a mean number of uses of 1.58 per group (SD = 0.85). Some groups chose to repeat

the tutorial up to 6 times. Groups in School #2 were the only ones to repeat the tutorial, with groups in school #1 utilizing it the one mandatory time. The introductory PowerPoint slides for Day 6 describe the ability to repeat tool tutorials if groups become stuck, but this is not explicitly described in the lesson plans for teachers. The optional comparison tank help that groups could activate was used sparingly, with an average of 0.51 questions framed by the tool per group (SD = 2.38). While many elected not to use the tool, the 27 groups that did make use of it posed an average of 3.78 questions with a maximum of 28 uses observed in the log files for one group. In keeping with observed trends, groups in School #2 were more likely to utilize this framing feature during their use of the comparison tank tool.

Table 11. Use of new features by teacher.

| Category | EJ | JA | SH | ML | RB | SC | JG |
|---|---|---|---|---|---|---|---|
| School | 2 | 2 | 2 | 1 | 1 | 1 | 1 |
| Avg. Use of Ranger Susan | 3.93 | 2.72 | 3.23 | 0.50 | 2.88 | 0.62 | 1.47 |
| Avg. Use of Jabir Hatami | 1.20 | 1.41 | 3.79 | 1.13 | 0.31 | 0.37 | 0.35 |
| Comparison Tutorial Use | 2.00 | 1.51 | 1.42 | 1.00 | 1.00 | 1.00 | 1.00 |
| Comparison Questions | 0.67 | 0.45 | 0.98 | 0.13 | 0.00 | 0.00 | 0.47 |
| Average # of Reflection Notes | 0.51 | 2.49 | 5.37 | 3.26 | 0.48 | 0.46 | 1.82 |
| Average Word Count | 19.8 | 20.5 | 31.9 | 25.5 | 14.1 | 53 | 25.3 |

Group save states show a varying amount of reflection notes between groups and between teachers throughout the curriculum. A total of 587 reflection notes were saved by

groups during the curriculum, with a minimum of 0 and a maximum of 21 notes. As there were only 11 prompts for reflection provided, it is likely that some groups inadvertently saved observational data as reflection notes or saved multiple notes per reflection in lieu of editing one note per prompt. 162 groups made at least one reflection note (53% of participating groups), resulting in an overall mean of 2.13 notes per group and a mean of 3.62 notes per group for groups who completed at least one reflection. When examined by teacher, four teachers out of 7 appear to have used the reflection tool more consistently than the remaining three (Table 11). Manual inspection of the reflection notes saved by groups with teacher averages of less than one revealed the presence of observation notes that were incorrectly saved.

Reflection notes contained an average of 26.6 words per note, with values broken down by teacher provided in Table 11. This was confirmed with a visual inspection of the text data that revealed most reflections to be structured as 1-2 sentences answering the prompt, with some elaborating in detail and others providing terser replies. The visual inspection also revealed that most notes contained meaningful content that addressed the reflection prompt, with only 5 reflection notes comprised on nonsense or irrelevant text meant to take up space. While not explicitly part of the curriculum, students may have assumed teachers would want to see these reflections and thus stayed on-task to a high degree. The most frequently used words from the reflection notes are shown in Figure 10. Many reflections centered on the fish kill event, thus the large presence of “fish,” “bass,” and “die.” The prevalence of “fertilizer” and “bacteria” indicate two foci of reflection for groups and exploring the context for how they are used in reflections may be illuminating. The word “wonder” appears as several prompts asked groups to explain what they are still wondering about the pond at certain points in the curriculum.

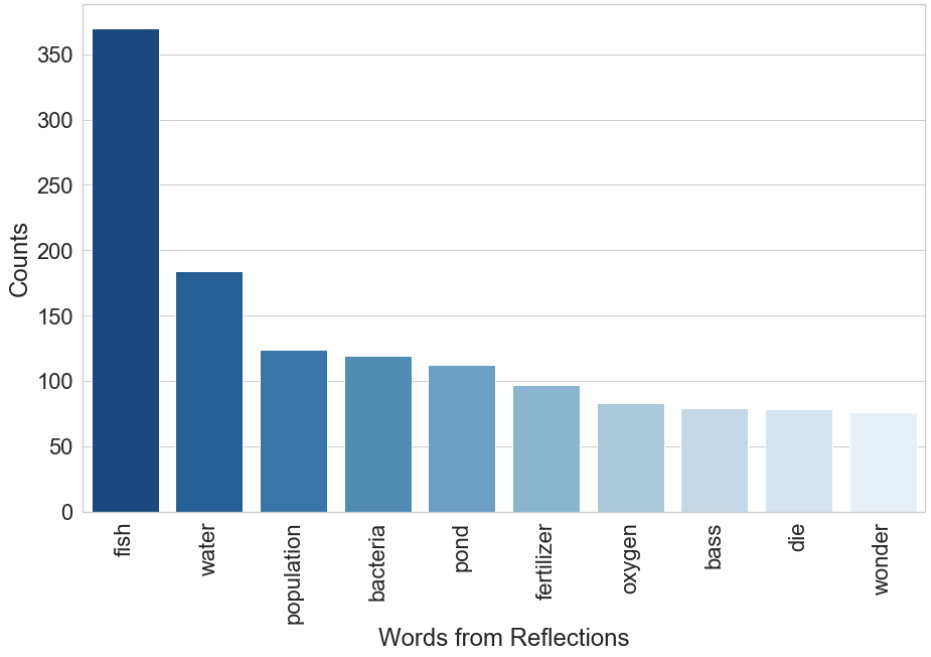



Figure 10. Most frequent words from student reflection notes.

**Student Feedback on New Features and Stuckness**

Three additional questions were added to the end of the EcoXPT post-intervention survey soliciting student feedback on the five new student-facing features added to the virtual environment. One question asked students to rank the five new features in order from most helpful to least helpful, the second asked students to elaborate on how the most useful feature helped them, and the third question asked students to describe a time they felt lost or frustrated in the curriculum and to suggest what might have helped them. Thirty-six percent of students rated the step-by-step comparison tank tutorial to be the most helpful new feature, and 22% found Dr. Hatami's advice on the concept map to be the most helpful. In contrast, 43% of students found Ranger Susan's advice on what to do throughout the curriculum to be the least helpful feature, and 30% of students ranked the reflection activities as the least helpful. The fifth feature, the help button on the comparison tank tool to scaffold question asking, was consistently ranked between these extremes (Figure 11).

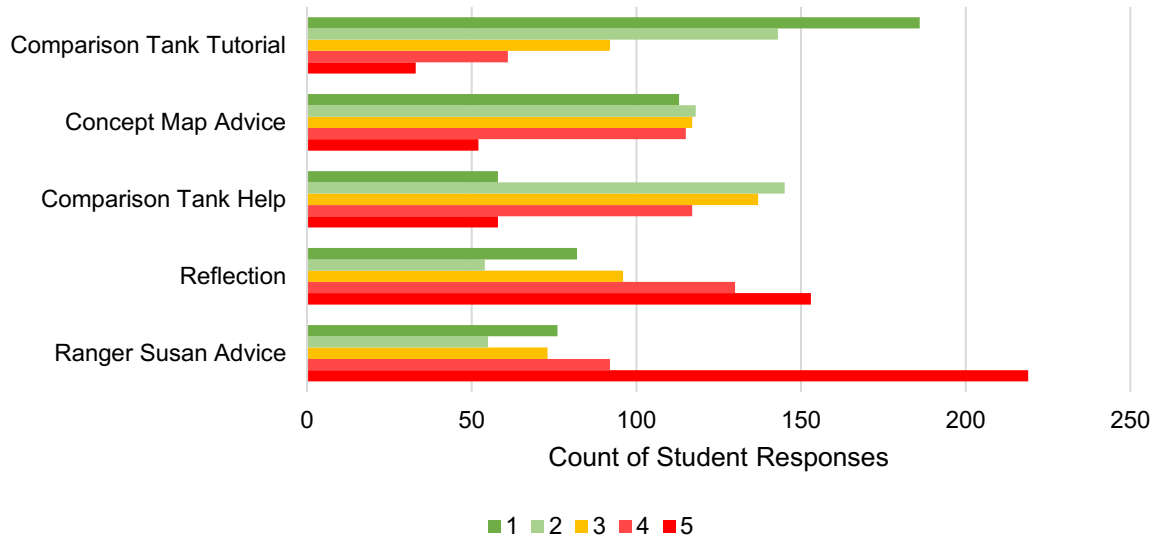



Figure 11. Student rankings of new features (1 = most helpful, 5 = least helpful).

Student replies elaborating why they chose a particular tool as the most helpful were thematically coded to identify common themes for the different tools. Students who selected the comparison tank tutorial largely describe the tool as seeming overwhelming at first glance but that multiple uses of the tutorial showed them how to use it correctly to generate causal evidence. As described by one student:

> The step by step tutorial was very helpful because it helped me understand the comparison tanks better. Without it I would have had no idea what to do. The comparison tanks later helped me provide good evidence.

Other groups focused on the utility of the tutorial as they felt the comparison tank was the most important tool for generating causal evidence:

> The comparison tank tutorial really helped me to learn how to use it. When I finally understood how to use the comparison tank, it helped me a lot to figure out the mystery.

> Learning how to use the comparison tank was very helpful because it showed me many patterns throughout the ecosystem. Also, the comparison tank showed causal evidence that was important for the concept map.

> The tutorial on the comparison tank was most useful because the comparison tank provided my claim with a lot of evidence. If I did not know how to use the comparison tank, my argument wouldn't of [sic] been as strong.

An additional theme mentioned by some students was a time saving aspect of the tutorial:

> The tutorial helped the most because it helped us learn how to use it. Even if we figured out how to use it by ourselves, it would have taken us a lot of time. The tutorial saved us some time, so we can do more things in the world.

This observation is consistent with worked tutorial literature summarized in Chapter 2. While not adaptive to particular groups, students felt this feature nonetheless served as an effective scaffold for their investigations.

The next most helpful feature, the concept map advice, was largely described by students as mitigating confusion by providing direct hints that were actionable and specific:

> I thought that Dr. Hatami's advice on the concept map was the most helpful because he was able to give you feedback on things to improve to get the best concept map you are capable of. It was helpful to have another opinion on your work.

> Me and my partner made our concept map stronger using the advice he gave us, we added more reasoning and more evidence because of what he told us.

> Dr. Hatami's advice was the most helpful as he could see the gaps in our concept map. However, he did not tell us what was wrong with our concept map. He only told us to look for gaps, not what the gaps were.

Other students selected Dr. Hatami's advice because it made less obvious aspects of the tool more apparent (such as the ability to make "does not affect" claims):

> it helped because i learned that scientists also put things that do not affect their goal for their concept maps

Some students described the tool as facilitating a deeper understanding of the complex causal relationships at the virtual pond:

> Dr. Hatami's advice was the most helpful. His advice was the most helpful because it really encouraged us to look deeper and pay more attention to smaller details. Without him, we wouldn't have been able to find the evidence that we needed for our claim, and we wouldn't have been able to make such a detailed concept map. These reasons are why this feature was most helpful.

Students were also aware that the concept map was to be the final deliverable of the curriculum and that they would be graded on their presentation, so they felt that this feature was the most important due to the higher stakes of the concept map. It is worth noting that this is not the only method that participating teachers used to assign grades in EcoXPT.

Students who chose to elaborate on the help button inside the comparison tank tool chose this feature for many of the same reasons students chose the tutorial, namely that this feature made it easier to generate causal evidence:

> The help button on the comparison tank that showed how to ask a question was what I found most helpful because it helps you to generate a claim/hypothesis based on evidence.
>
> It helped me, because I figured out what combinations of materials produced the most dissolved oxygen, and turbitity etc.

Several students mentioned that the help tool and the question framing it provided aided them in generating disconfirming evidence:

> The help button on the tanks helped the most because the tanks knock out many hypotheses. It helped when I was lost on how to test certain stations, because knowing how to use them is helpful.

Others seemed to appreciate the on-demand nature of this feature versus the mandatory tutorial every group had to complete:

> For the help option, I wouldn't have to search to find what I needed, I just asked and it appeared!!!1!
>
> When i didn’t understand I could use the help button and it would get me back on track.

While relatively few students selected this feature as most helpful, it is possible that this subset was influenced to do so because of this main difference.

The reflection activities were selected as most helpful by students who found them a good way to summarize their thoughts for the day and plan ahead for what else to investigate. Some students also specifically mentioned getting unstuck via reflecting:

> It helped me by configuring all the loose strands of ideas and knowledge floating around my head. The reflection activities helped me bring them all together and develop a complicated reason why the fish were dying in the pond.

> The reflection activities helped me the most because when I am on EcoXPT sometimes I get very stuck in what to do next and in class we had questions in the beginning of the class that we right out, so we can have a sense of what to do on EcoXPT.

> Although I thought there were a bit too many reflection activities at the end, the really helped me organize my thinking and understand my explanation and experiences in EcoXPT more.

Teachers who made use of the reflection prompts often had students share their findings with classmates at the end of the day or would use their reflections as a jumping off point for whole class discussions, and many students found this aspect helpful:

> in my opinion the reflection activities at the end of class helped the most because we got to talk about the answers we got and it helped us move [move] on.

> Reflection activities were the most helpful for me because it helped me understand if I was missing information by showing my classmate's thinking and discoveries. If there was something I couldn't figure out, I could hear about what my classmates did, and get an idea of what to do. These were also good times to ask my peers for help.

Since all reflections are stored in the electronic notebook in this modification, students can freely look back on them and see how their thinking evolved over time or see where they left off the previous day. Several students commented on this aspect:

> The reflection activities helped me gather my thoughts and write how they might be causing the fish to die. When you have written enough reflections you can look back and see how your opinion has changed.

> it was important and helpful because I could always go back to it.

While relatively few students overall (82) selected the reflection tool as the most helpful feature, those who did valued the ability to step away from their investigations and synthesize their thoughts with the intention of revisiting them and eventually sharing them with the class.

Those students who found Ranger Susan's "What should I do?" advice to be the most helpful found her advice to be essential to navigating the complex inquiry task set before them, with some students mentioning aspects of engagement such as facilitating "on task" behavior:

> Ranger Susan kept my group on task and gave us something to do when we were not sure what we should have done.
>
> it helped me because I would have been lost on about what to do if it weren't for ranger Susan.
>
> She got me started on the trail and I would be running around the pond with nothing to do without it.

Many other responses in this category specifically mentioned Ranger Susan's advice helped their group avoid being stuck by suggesting heretofore unused features or different tools to investigate:

> It was helpful because it gave you a good idea, for when you are stuck and need someone to steer you in the right direction.
>
> Ranger Susan's advice really helped me because, if me or my partner got stuck at one point, we could easily go to her for help on what to do next.

While previously both of these roles fell to the teacher, this feature allows students to get answers to simple questions and avoid minor issues of stuckness, while freeing the teacher to work with groups struggling on deeper issues or other concerns.

Emergent themes from student descriptions of times they felt lost or frustrated and what might have helped them illuminated several key areas where students are often feeling stuck or discouraged. 320 respondents provided meaningful answers to this prompt that addressed one of both aspects of the question. Of these, six different aspects of the features and curriculum were

discussed by 30 or more students: concept mapping, feedback from ranger Susan, NPC hint generation, disconfirming hypotheses, collecting data, and the buoy data.

Students who discussed concept mapping often described this aspect as the main area where they felt lost. These feelings often occurred near the first time the tool was used or close to the end of the activity. Those who mention feeling lost near the beginning typically described not knowing what to do with the interface or where to start assembling their map. Students who describe feeling lost in the concept map later on relate difficulty filling gaps in their concept maps and using existing evidence in effective ways:

> I felt lost when I had some gaps in my concept map and didn't know what data and tools would help. What could have helped me is looking at the data shown by the different tools to help fill the gaps in my concept map.

> I was working on my concept map with my partner and we had the beginning part of it with the fertilizer and the end part of it with the fish and the dissolved oxygen, but we were missing the big part in the middle. We tried to look around and try to use the graphs and look at the data table more but we couldn't find anything for a while. And in the end I met with my teacher and she guided me to find a part of the concept map we didn't have before.

These replies often discuss finding a part of the world they missed or a tool they could use to solve the problem after consultation with a teacher. Others describe getting through these periods of stuckness by relying on feedback from Jabir Hatami:

> I felt frustrated when I had pieces of my concept map but couldn't put them together. I used Dr. Hatami's help to figure out what I was missing.

> I felt frustrated when I couldn't figure what to put in my concept map next. When Dr.Hatami gave advice on are concept map, that really helped.

While not all students found the NPC concept map feedback helpful in its current form, many who struggled with mapping felt that additional hints and clues would help them get started more effectively in this portion of the activity. Others also mentioned wishing the experimental tools were unlocked sooner to generate causal evidence for their maps earlier.

Students who mentioned Ranger Susan's feedback with regard to stuckness and help were largely negative about that feature's ability to aid them. In general, students felt that her suggestions were not helpful or that more specific advice is necessary:

> One time I felt lost in EcoXPT is when my partner and I did not know why the fish had died. …What could have helped us is if Ranger Susan helped us a bit more and gave us more useful advice.
>
> I did not get why ranger Susan could help because she never showed any data.
>
> I wish Ranger Susan gave more detailed instructions because they were similar to what we were told in class and I wish maybe she pointed us in the right direction more.

Some of the responses seem to reveal a lack of understanding about her purpose in the world (high-level day-to-day hints that mirror the framing from the lesson plans and not a next step advice tool), while others bemoaned the inability to ask her specific questions. These replies describe craving the dialog that they typically have to go to their teacher for in order to work through their problems. The idea of a "right" direction or answer pulls against the goals of EcoXPT to deliver authentic scientific instruction and highlight the true nature of science, and this was not an intended purpose for Ranger Susan.

While Ranger Susan's feedback may not have been as helpful as some students wished, many others expressed interest in receiving more hints and clues in the virtual world as a way to become unstuck. Several respondents mention desiring more just-in-time hints when stuck:

> When nothing was working and I was going in circles. It could give hints if the group was stuck for too long.
>
> I felt lost when evidence ended in a dead end. Maybe it could have helped if there was a little more guidance after you discover something, so that you know what this means and what to do next.

Other students reinvented several supports that already exist in the world, such as an NPC that helps decide what experimental tool to use (Dr. Hatami) or clues about what happened on days

when you weren't at the pond (buoy data). This indicates that some groups were still not aware of all aspects of the virtual world after 13 days of using it, although this is not necessarily an issue. The curriculum's design emphasizes a "low floor, high ceiling" approach where different pairs discover different aspects of the world.

Many students described maximal frustration and stuckness when encountering a piece of disconfirming evidence that required accommodation of existing schemas. In general, students who described this feeling thought that earlier access to experimental evidence would reduce the likelihood of making so many naïve hypotheses that need to be disproven, as well as that nudges toward certain tools or aspects of the world might have been beneficial:

> I felt lost when my hypothesis that the fertilizer had killed the fish turned out to be incorrect, after we ran an experiment and discovered that the fertilizer had no effect on the fish. I think that maybe if we got to open the shed a bit earlier we wouldn't become to set on our first explanation.
>
> We felt lost when we ran an experiment and found out that what we had thought caused the fish to die out actually didn't affect the fish. It would have helped if one of the people had helped us find out what to look at next.

While this cycle of disconfirming initial hypotheses and generating new ones to test is an essential part of authentic scientific inquiry and is an essential goal of EcoXPT, these responses are a reminder of how foreign a concept this process may be for certain students.

Respondents also expressed frustration in the early days of the curriculum with the onerous task of data collection throughout the virtual summer. Some complaints focused on the user interface and the difficulty of finding or being able to click on certain species (the bacteria and red-tailed hawk are both mentioned several times.) Several students suggested ways of speeding up the data collection process:

> One thing that could have helped me would be if the field guide had alerted me when a tadpole was nearby (or any other living organism). It could also have

> helped if we did not have to take pictures of microorganisms for the field guide. It would be easier if the microorganisms were already there.

Features to make data such as these more salient already exist in EcoMUVE Forest and EcoMOD. While intended to be a simple introduction to the user interface and the virtual world in general, early tasks that generate frustration and annoyance will not be beneficial for engaging students or making them feel as though their time is being used productively.

Related to the general issues with data collection, many students specified feelings of frustration related to gathering and using the sensor buoy data in particular. Using this data is the only way to definitively say that the dissolved oxygen levels in the pond dipped low enough on a certain day to kill the large fish. Students describe erroneously discounting the importance of dissolved oxygen for a long time after seeing it never drop low enough on the days they visit the virtual pond. Additionally, many students described being frustrated with the volume and complexity of the buoy data when only one aspect of it is essential to the problem at hand:

> we got lost when using the [buoy's] immense data. There was SOOO much that we couldn't find the important one we were looking for. Maybe you should've highlighted the first one of each day.

Others were simply frustrated that they did not find the right NPC until late in the curriculum or that they did not realize they needed to click her tablet to gather the data (despite dialog explaining this).

**Teacher Feedback on New Features and Stuckness**

In general, the seven teachers were positive about the curriculum and the four veteran teachers noted that the implementation seemed similar to the existing EcoXPT curriculum from their perspective since they didn't directly interact with most of the new features. In comparison to their experiences with the baseline version of EcoXPT, two veteran teachers noted that more

groups seemed to get farther along on their own and that teachers had to ask fewer leading questions, instead focusing on encouraging reflection and planning effective experiments:

> ML: I feel like more students got further along on their own. I feel like I was able to ask less leading questions. I guess my questions were just more open-ended and kind of repeated: take a step back, what can kill the fish, and then go from there. I did a lot of clarifying what tool would be the most useful. So this is what we want to find out, and then I would direct them; I wouldn't tell them how to set anything up, but it was more, this is a tolerance tank question or this is a comparison tank question.

The other veteran teacher at the school noted that she had fewer students with very incomplete concept maps at the end of the activity and that fewer groups ended up stuck in places she could not help them out of without direct instruction. Despite the new tools, teachers did still occasionally get questions pertaining to what groups should do next, and the quality of group experimentation still varied greatly between groups. All teachers noted that they rarely saw students directly using the new features, since groups would use teacher attention as a time to ask the teacher questions instead, but they would hear conversations about the tools.

Teachers heard a litany of complaints about Ranger Susan's advice that mirror the comments students made in their surveys, namely that students wanted more specific guidance to get the right answer and that they were afraid of being wrong:

> RB: I think they thought she was going to tell them what the answer was, or where to go next. Like specifically, because… they're 12 and 13; they're still obsessed with the right answer. And they're so afraid to tell you what they're thinking… I'm like, "You gotta tell me what you're thinking right now. I need to see it." And they're like [sits in silence for a moment], "Start with that sentence." "I'm really not sure." Like they have to give you a disclaimer.

Another teacher referred to the difficulty of using feedback tools with this age group:

> SH: Honestly, I think it's developmental. Like I said, they tend to like the path of least resistance. They're like, "We already looked there." I'm like, "Nope. Look again." They're impatient, just as a whole, 12- and 13-year-olds. They're impatient and they like to do the easiest, right? "I clicked on that..." and it's like, "But did you read it?"

SH also hypothesized that few students followed up with Ranger Susan, as her role was not as emphasized in the introductory lessons (despite being introduced early on in the slides), and that students might make more of a connection if she was connected with Jade in the thinking move videos. Two veteran teachers were not aware that Ranger Susan's role had changed in the curriculum and did not realize she was giving hints until students mentioned it. As observed in the student surveys, teachers heard more positive comments about Dr. Hatami's concept map feedback. One teacher, however, expressed frustration that the message the NPC gives when a large map is presented with no obvious errors might be interpreted as a sign that groups are done working and do not need to add more claims or strengthen existing ones.

Teachers reported that the new tutorial reduced the number of low-level questions students asked them about the interface and how to run experiments. While some groups still rushed through it at first, they were observed repeating the tutorial more deliberately when confused instead of going directly to the teacher or giving up. While some teachers doubted that the question scaffolding "help" tool taught students much content knowledge about proper experimentation technique, they agreed that it did align with classroom instruction on the topic. One teacher described a typical poorly designed experiment:

> ML: For example, they would put bacteria and dead matter in a tank, then put a tank with bacteria, and then not put anything for a control. So they're like, "Oh, look, bacteria increases the amount of bacteria."

Teachers at both schools expressed a desire for this tutorial to be expanded to cover the concepts of control and when to use multiple factors, with the understanding that this would also require more introductory lessons on experimentation and could extend an already lengthy curriculum. One novice teacher expressed interest in lifting the cap on two factors added to each tank, with the understanding that this would make controlled experiments harder to set up.

Teachers reported using the reflection prompts in very different ways. Two veteran teachers reported not using them at all with their classes, instead prioritizing maximizing the amount of time students could spend in the virtual world. Another veteran teacher described not using them initially to let students focus on data collection, but then incorporating them later when the problem space became more complex. Conversely, one of the novice teachers said she used the prompts at the beginning, then stopped doing it as the material became more complex as she felt her students needed more time focused on building their concept maps. She described her main struggle as getting the students started on the "right path" and was very hesitant to then interrupt them when engaged and working productively. The fourth veteran teacher used them occasionally but as homework assignments to allow for more grades to be entered for students during the curriculum. At the start of her next class, she would have students share their homework responses as a warm-up discussion before starting the lesson for that day. This mirrors a suggested use of the prompts from the original version of EcoXPT. Discussing the importance of the reflection activities with teachers at one school lead them to discuss the importance of the reflections in fostering the practices of science in their students versus driving home additional content knowledge.

The addition of the new reflection notes also lead some teachers to discuss general issues they saw with student notebooks. Students in some classes were confused about when to use reflection notes versus observations, and how reflections couldn't be used as evidence, but observations could. Teachers reported many students struggling saving a large amount of notebook entries with insufficient notes to go back and find important information later. One teacher suggested expanding an entire lesson to note taking in a scientific notebook. Teachers identified the notebook as a potential area for more support in the virtual world as well, with the

ability to check what students have put for notes and provide feedback that might be helpful for them when they go to use the notes as evidence.

When asked what behaviors indicated to teachers that groups were stuck or struggling (aside from hand raising), a main indicator that teachers mentioned was to look at what tools are being used midway through the curriculum and see if they are appropriate (e.g., not collecting water quality data still). Some suggested limiting the ability to use certain tools after certain points in the curriculum to avoid policing these types of behaviors. There is an option in the curriculum to fill in all missing data for a specific group, but teachers reported struggling groups using these data collection tools even after being given all possible values. Another main indicator of stuckness for teachers was when groups succumbed to confirmation bias and were reluctant to let go of an incorrect hypothesis despite evidence to the contrary. Once the concept map is partially built, these claims are easy for teachers to notice when observing a group. Certain key phrases that teachers overheard indicated that student frustrations had reached a point where teacher intervention was necessary. Examples given were “I’m done,” “This is stupid,” and “I’m finished.” One teacher summarized her feelings about student reports of feeling bored during the curriculum:

> JA: I think most of the kids stayed engaged. I remember kids telling me towards the end of it that they were bored with it. And I'm like, “You know, the kids who are actually doing it are not bored. Your choice to screw around is actually affecting how much you're enjoying science class.”

Additionally, teachers emphasized the importance of observing body language within groups to gauge how things were progressing:

> SH: I would definitely look at body language between partners. Sometimes one kid was stuck and would just check out [pantomimes slouching and looking into space] and the other kids were persisting. I would go over there and that would be one of the times [where groups were struggling].

One veteran teacher felt that issues related to disengagement from the non-computer-using student was a big enough issue to suggest doing the activity individually next implementation. While she acknowledged the benefits of working in groups, she felt that issues she'd had with classroom management from bored partners were overall more detrimental. Other observed behaviors associated with stuckness include groups asking questions of their nearby peers or spending too much time on one experimental tool.

Teachers describe distinguishing between groups being stuck and productively struggling mainly by sitting with pairs for several minutes and listening to their version of the causal story that lead to the fish kill. Looking at their concept map was usually an easy way of telling when groups may require nudging in a productive direction versus being left to investigate. Several teachers highlighted the importance of including "Does Not Affect" relationships in concept maps as a way for students to not feel as though the work they've done prior to disconfirming a hypothesis was wasted or that they had accomplished little:

> SH: I liked [the DNA arrows] and I was like, that really shows me that you tried stuff and you learned that it wasn't right, and that's really helpful. I wished I had never told them to take stuff out of their concept maps… I actually wished on the contrary, that I had told them to leave stuff in because their disconfirming evidence was important. That told them things that they had tried, that they knew weren't right, so that they didn't have to check it again.

Veteran teachers reported seeing more students using DNA claims than in prior implementations, and they found this useful for giving groups feedback as well as giving struggling groups something to talk about during their presentations. When looking at a map with naïve hypotheses with regard to the large fish dying, the most common piece of advice teachers delivered was to send groups to the tolerance tanks to test claims and see that the evidence does not support them.

At the school that used the teacher reports, the two veteran teachers and one novice teacher were all very positive about this new feature of the curriculum. As one veteran teacher

put it, “It helped me catch the kids who were falling through the cracks that I didn't think were falling through the cracks.” These teachers found it empowering to see data they’d never had access to presented in a manner they could understand and use to directly help struggling groups. Teachers described students being shocked that teachers could know about the data stored in their notebooks and what they had written. Jokingly referring to it as “Big Brother in EcoXPT,” one veteran teacher noticed groups being much more conscientious about labeling their notes and systematically storing data after teacher reports indicated they had had issues in these areas. Teachers avoided what they referred to as “public shaming” and would deliver feedback directly to groups in private or make general announcements at the start of class about widespread issues. The novice teacher did note several times where feedback on the report did not seem to match what she saw on the groups’ screens, but these were seemingly isolated incidents. Teachers were aware that work on this research project would be wrapping up in 2020 and were concerned that similar reports could not be generated in the offline version of the curriculum as easily.

These teachers found the visualization of the proportion of logged activities to be easily interpretable after an initial explanation with the first report and were satisfied with the concept map and notebook hints provided in the bottom section of the report. Teachers said the visualization helped find off-task groups easily. One veteran teacher showed the visualization section of her reports to her classes directly and let them draw conclusions about their own time management and explained what types of activities go under each category. When asked about other features that might be added to the report, teachers wondered if it would be possible to track which student in a pair was at the keyboard at any given time. Linking back to the potential issues with disengagement mentioned earlier, teachers wanted to ensure as equal participation in the curriculum as possible.

# Chapter 5: Discussion

In general, the LENS suite of feedback tools helped students learn about epistemology and the nature of science, increased the number of claims groups made, increased the relevance of the evidence they used, and increased the accuracy of classifiers designed to predict student success by 37%. Student opinions on the tools suggest that some view them as helpful while others view them as insufficient as they still crave direct instruction and are not used to ASI activities. Teachers used these tools in varying ways in their classrooms and provided very positive feedback on the visualization and daily report. This chapter summarizes the research findings detailed in Chapter 4 in terms of how they inform the research questions for this study. Overall trends from the different analyses are teased out and interpreted as part of this effort. Limitations of the work are outlined along with possible remedies. Potential future work inspired by these findings is also discussed.

## Overall Trends

### Teacher-Facing Support

The first research question in this study focuses on the teacher-facing support added to the EcoXPT curriculum. As previously reported, the three teachers who made use of the summary teacher reports found them useful in guiding their whole class instruction to address issues faced by multiple groups, as well as attending to struggling groups. The three hypothesized sources of information for the reports (types of logged events, types of saved notes, and current state of the concept map) proved sufficient, and teachers reported feeling empowered by having access to this otherwise difficult- or impossible-to-observe data.

During the design of the feature, the impact on student behavior of knowing that everything they entered in the virtual world could be observed had not been considered as a

major feature. As teachers described, however, even knowing this level of observation was possible led to observed behavioral changes in students. While the psychological aspects of user interface and user experience design of dashboards is a well-studied problem (Xhakaj et al., 2016), less work has been done on the indirect impact of these tools on the students whose work is being observed. Much human-computer interaction research on teacher dashboards considers their effect as a causal chain that tracks any potential effects on student learning as being mediated by teacher ingestion and processing of dashboard data (Xhakaj et al., 2017). The reported behaviors from veteran teachers in this study, however, indicate that the causality is not quite that linear (such as the reported observer / "big brother" effect).

As has been observed previously with EcoXPT, teacher fidelity of implementation varies significantly across several dimensions (McGivney et al., 2019), and teachers' use of this summary report was no exception. While not meant to be prescriptive or a scripted curriculum, there is always a risk that students or teachers may use features or tools meant to help them in unintended ways that may be harmful to learning or undercut the goals of an ASI curriculum. While it is impossible to quantify the isolated effect of the teacher tool, both veterans and the one new instructor found it helpful for classroom management and providing formative feedback. The teachers in this study self-selected to pilot a new suite of technology-enabled tools for a cutting-edge inquiry-based science curriculum and can safely assume to have been acting in good faith during their implementations in an intrinsically motivated fashion (despite the monetary remuneration they received for their participation.) When deploying these tools at scale or providing them online for teacher use at-will, proper documentation and refinement of the presentation must accompany the feature to make sure its use aligns with the broader goals of the curriculum.

### Types of Feedback

The second research question asks about what types of feedback show evidence of behavioral change when offered to students along with the regular EcoXPT materials. It was hypothesized that students using the added LENS components would have larger gains on the affective and experimental methods portions of the survey, but they did not outperform the EcoXPT groups from earlier implementations. The LENS groups did have higher gains in epistemology. A manual investigation of student surveys revealed no particular type of question that LENS groups performed better or worse on within the epistemology survey, but overall performance was slightly higher on average. Several of the LENS tools explicitly relate actions in the virtual world to authentic scientific practices (e.g., experimentation strategies in the companion tank feature). The experimental methods survey focused more on specific experiments that ecosystem scientists can use in different scenarios and thus had a different focus than the final form of the feedback features. The affective dimension was hypothesized to differ, in that it was expected that reducing student risk of unproductive floundering during the activity might increase student sense of self-efficacy but this was not the case. It is also possible that there is less unproductive floundering in EcoXPT from the perspective of students than perceived by teachers and researchers. Overall, these results are promising as they demonstrate that the added LENS tools do not, at the very least, harm student learning on any measured dimension.

Gains on several constructs were moderated by teacher reported student engagement in science, with students rated “low” on engagement consistently gaining less across constructs when controlling for other factors. Only 66 out of 575 students from the survey dataset were classified as “low” engagement (11%), but these are students who might benefit the most from a

virtual environment-based curriculum that typically achieves high levels of student engagement (Metcalf et al., 2014). Future feedback tools could be designed with engagement as a higher priority alongside methods for detecting group disengagement and intervening accordingly.

Differences by teacher on the epistemology and causality constructs likely do not indicate any particular issues with the curriculum or the condition assignment (as the presence of the new tools did not interact significantly with veteran status), but add more to the conversation regarding teacher variation in implementation. Different teachers focus on different aspects of the curriculum, and at times one's focus can result in larger gains on certain constructs than those of their peer teachers. This may not be an issue with a larger sample, but the effect of having previous experience with the curricula may be a consideration in future studies. As these technologies become more prevalent, experience with them and with teaching authentic scientific inquiry may become more significant.

Issues with lower reading level being associated with lower learning gains have been previously observed in as yet unpublished analyses of large-scale implementations of EcoXPT, but survey gains do not typically differ by IEP/504 or ELL status. EcoXPT has been used successfully with a diverse population of ELL and special education students in the past, and it is likely that this is due to the particular sample assembled for this study. The issues related to student reading level may also be more of an issue with the assessment than with the curriculum, as the pre- and post-survey involve much more reading than the immersive environment requires. As this was an issue when controlling for LENS feature use, it does not appear as though the new features made the curriculum less accessible. This result may be due to the small populations of students in both of those categories, with only 18 students in the final dataset classified as English language learners and 81 designated as having an IEP or 504 plan. Design decisions in

virtual worlds can be made to leverage the affordances of the virtual environment to replace some written text with audio or video from NPCs to reduce reading load.

Concept map quality also differed by condition, with LENS groups making higher quality maps on average according to the rubric. All thresholds for the different quality metrics and algorithms for determining them were implemented prior to testing for these differences while still being based in CER literature (as described in Chapter 3). While this analysis does not show which tools made the difference, it suggests that the tools offered in the LENS package do make a difference in how well students are able to model the particular complex causal relationships of the pond ecosystem. On balance, and as discussed above, while the rubric was unaligned with the broader pedagogical goals of EcoXPT, it was aligned with the intents and purposes of the LENS suite of tools: to move towards formative assessment and scaffolding to support building a concept map that mirrors the one designed into the technology.

Trends in how different new features were used by different classes shed light on which aspects were deemed valuable and how this valuation differed by teacher and school. From the summary statistics listed in Table 9, it is clear that teachers in School 2 implemented a more faithful version of the curriculum with added features. While EJ's use of reflection notes appears to contradict this trend, she revealed in her interview that she used the prompts as homework assignments, as she had previously in her last use of EcoXPT, rather than have students save them in the virtual world. At School 1, ML and SC classes barely used Ranger Susan, while only ML's students there made extensive use of the concept map feedback. All students at this school only used the comparison tank tutorial once, and RB and SC classes never utilized the comparison tank help function. These differences do not break down by novice versus veteran instructors. Despite being given identical training on the new formative feedback features and

having the same introductory slides that describe these features, it is clear that differences in teaching style and presentation lead to very different behaviors in the virtual world. This is consistent with Tutwiler's (2014) findings with regard to between-teacher variation in student data collection in EcoMUVE. These variations add difficulty to the task of assessing which tools made an impact. JG's mean epistemology gain was significantly higher than JA, RB, SC, and EJ, but her overall use of new features was low.

The varying use of the reflection tool between teachers is reflective of a long-standing tension that teachers have expressed with regard to the EcoLEARN curricula. The curriculum designers encourage discretion about too much sharing of findings so as not to eclipse the opportunity for all students to engage in inquiry. Some teachers, none-the-less, encourage whole class discussion and sharing of major discoveries as they happen. As a natural part of these discussions, groups adopt each other's connections in the concept map and learn about causal connections they may not have discovered yet in the world, but they learn less about how to persist through challenging unstructured problems and how to approach their own inquiry. While, the lesson plans currently leave this to the discretion of the individual teacher, the newly added "reflection" feature does push lessons more towards encouraging communication rather than sequestering. The main goal of EcoXPT is not the discovery of a final "right" or "correct" concept map, but rather an understanding of the nature of science and of the difference between correlation and causation.

Student feedback at the end of the survey illuminated some perceived issues with the newly added tools. The hypotheses for the second research question conjectured that students would find direct guidance prompts to be most helpful. While the direct guidance of the tutorial was found very helpful, Ranger Susan's guidance was ranked last in helpfulness of new features.

An unusual amount of vitriol was put into some comments about Ranger Susan that was not observed in reaction to Jabir Hatami, an NPC with a similar role that provided slightly different advice. Hatami's advice is aligned with the greater curricular goals of the program by helping students bridge the gap between their current understanding and goals related to evidence gathering and use. His advice echoes what was presented to students via PowerPoint, instruction at the beginning of lessons, and through the Thinking Move posters. Students were better able to attend to the information when it arrived just-in-time in a dynamic way. Ranger Susan, on the other hand, pulled against the goals of the curriculum by providing low-level direct advice and was not found as helpful.

Ranger Susan acted as a lightning rod for students taking out frustration as one of the most consistently visible NPCs in the world. Gender differences of NPCs and player avatars have been studied in entertainment games (Bergstrom et al., 2012) and may play a role here. Students and teachers both asked for more dialog options with NPCs and for video or at least animations to make pedagogical agents come to life. This is consistent with Lester et al.'s (1997) work on affective aspects of pedagogical agents, where lifelike representations of characters have a strong positive impact on student perceptions of the activity. Student relationships with and use of pedagogical agents can also vary widely by student gender (Pezzullo et al., 2017), and this aspect of the negative feedback must be considered. While graphics in EcoXPT are simple to function well on a wide range of school hardware and authentic by design, other simulations might be designed such to make different artistic choices to sacrifice realism for a more expressive, abstract style.

Within student comments about Dr. Hatami's feedback, they mention his model-based feedback as very helpful. While his importance may have been skewed by the knowledge that the

concept maps were a critical piece of the curriculum and would serve as the basis for their presentations, this still represents a different ranking of feature utility than was hypothesized. It is impossible to tease apart solely the influence of his advice versus other tools, but this feature is the one that would most directly impact concept map quality. Given the difference between conditions on that outcome measure, these data suggest model-based feedback may have a larger impact on student behaviors than other forms. In their written responses, students expressed that they wanted direct guidance about what to do and how to figure out the links that should be in their concept maps and that Dr. Hatami's model-based feedback was helpful towards this end.

This level of interest in getting to the right answer pulled against the broader EcoXPT goals of learning to engage in inquiry and learning the epistemology of science. This reflects both students' lack of experience with similar activities, as well as a tension inherent in balancing a need for support and guidance with the importance of authenticity in these types of inquiry-based activities. These tools have tried to strike an optimal balance of support and struggle, but it is clear that not all students agree that this level of support was what they wanted. The tendency to want to simplify the world to make it easier to comprehend is tempting for both instructional designers and classroom teachers implementing such curricula, but these simplifications will not happen in the real world. Striking the right balance between structuring and the problematizing represented in authentic problem spaces (Reiser, 2004) is important to helping students learn in these complex spaces, such that the inquiry process mimics the shared negotiation of meaning and scientific argumentation that are hallmarks of the real nature of science (Schwartz & Crawford, 2006).

Sharing responses to reflection prompts can unfortunately condense outcomes since more students see what the teacher views as the right answer in terms of claims and evidence. In turn,

they likely learn less about the process of inquiry, the value of uncertainty, and the important process of deeply exploring alternative hypotheses. An abundance of sharing during the curriculum risks taking away learning opportunities from those who are working at lower levels. The reflection tool was intended to be used solely at a group-level, but this intention was not always honored as revealed by teacher comments on their use of the prompts.

Teacher feedback from their interviews regarding the efficacy of different types of tools echoed student sentiment and noted ways for the pedagogical agents to be more obvious in the world. NPC delivered hints are intended to suggest courses of action without defining next steps or bottom-out hints, but teacher feedback suggests this may cause more anxiety and frustration from students who are used to direct instruction and being told what to do next in a procedure-like format in science classes. Feedback can be reworded and reframed in such a way to mitigate these feelings of frustration while still being helpful and not overly limiting in the open-ended activity. Feedback can also help re-center students on the main goals of the curriculum and away from the notion of right answers and arriving at one correct solution at the end of the curriculum.

**Key Sequences and Features**

The third research question asks about key sequences of events or other features of the logged information that can help generate formative feedback. The specific examples given in the hypotheses are not seen in the significant n-grams listed in Table 6. The overall trend of positive correlations with outcomes when sequences can be mapped to best practices in conducting scientific investigations does largely seem to hold. While these correlations are not large, they still suggest features from the logged data that may be informative in future model building. While the Markov models did not reveal major differences between the lowest and highest quartiles of students across several outcome measures, the clustering analysis reveals two

distinct yet equally successful typical state transitions that lead to roughly equivalent outcomes. This aligns well with the goals of an open-ended curriculum where there should not be one optimal path to follow to maximize learning. Clustering student approaches to problem-solving has been effective in helping computer science teachers provide feedback for novice programmers (McBroom et al., 2018), and similar results may be possible if clustering is used more extensively in future teacher reports.

Several sequences of logged events have been discovered that may indicate lack of focus or suboptimal uses of group time in the virtual world, but log file data is generally insufficient to explain exactly what each sequence signified for each group or to deterministically label each instance of these sequences as a negative. These sequences will need to be compared to the close observations of the variety of student behaviors observed in EcoXPT during previous case studies. Triangulation and the accrual of confirming or disconfirming evidence is one of the few ways to ultimately support this study's interpretation of what occurs during group work. Text replay tagging such as that used by Sao Pedro et al. (2010) would likely provide this much-needed triangulation and context for group n-grams.

The newly added hint and feedback tools resulted in a much greater ability to predict concept map quality than previously possible. These accuracies (as summarized in Table 10) approach those achieved in previous work when also using student pre-survey information as features (Reilly & Dede, 2019). The other interesting takeaway from this work is the efficacy of simple classifiers that do not utilize neural networks or time series data to make their predictions. A parsimonious model such as an SVM classifier is much faster to train and easier to embed in a virtual world, and these findings are significant for the development and implementation of feedback tools based on these predictions. Ultimately, this type of classifier could be used on the

first several days of a group's logged actions to predict the groups most likely to struggle and to direct scaffolds and teacher attention their way, in a similar fashion to Mao et al.'s (2019) work done with novice programming tasks.

**Limitations**

One of the main limitations of this study is the difficulty of teasing apart which features have measurable effects on student affect and outcomes. A design decision was made at the outset of the study to treat all recruited students with the same suite of tools, as teacher recruitment had been challenging and it was felt that a larger sample size would aid comparisons to historical EcoXPT data even in light of the resulting uneven sample sizes (as discussed below). Now that data has been collected and analyzed from the entire suite of tools, refinements to the tools can be made and A/B testing of different variants of the feedback features can be done in a future study. For example, some classes may be offered solely the metacognitive and model-based tools while others see only the direct instruction feedback. Smaller divisions would also be possible, but a different survey might be necessary as the EcoXPT survey is powered for a sample size of roughly 200 student groups.

One issue with detecting these smaller changes, however, is related to variation in fidelity of implementation between teachers. While relatively little within-teacher variation is seen between classes, different teachers utilize the curriculum in very different ways. These differences will only increase in magnitude as teachers begin using these tools without the assistance of researchers during the implementation. Even within this sample, several veteran teachers were not aware of certain changes made to the tool for this study and made little use of them or let students discover them for themselves in inconsistent ways. While teachers are always going to foreground their purposes, the tools developed must be resilient to this type of

variation. This type of between-teacher variability has been observed previously in immersive virtual environment-based curricula and will likely make generalizability of feedback tools and detectors challenging (Tutwiler, 2014). The issue of veteran teachers adapting to new versions of feedback tools should be considered for any future design-based research studies utilizing teachers with experience using similar technology-enabled curricula. The same training on new features was presented to new and returning teachers, but more explicit guidance may be necessary to combat the tendency to implement similar activities in the same way as previous years.

Incorporating group reflection activities as a more emphasized part of the virtual world had the unintended consequence of having teachers use these reflection and self-explanation prompts as a basis for end of class or beginning of class discussions. Teacher guidance against this was kept the same as with baseline EcoXPT, but the new feature may have been alluring for teachers to incorporate into their instruction. While some amount of leakage of information between groups is inevitable, more sharing of answers risks yielding more complete concept maps but less inquiry-based investigation. This is one possible cause of the difference in quality of concept maps between treatments, but features are confounded in this study design and it is impossible to examine the effect of one specific change.

While the LENS tools developed for the curriculum have shown promise that bears further investigation, many of them are currently dependent on online connectivity to store and process log file data on a remote server. With the conclusion of the EcoXPT project, an offline version has been made available to teachers, as funding is not available to keep the university servers running in perpetuity to process and store log file and save state information. The types of tools developed for this dissertation will need to be re-engineered to work on local machines

used at schools and to process data without internet connectivity. Some of the more advanced analyses like natural language processing may not be able to be embedded in the Unity game world. These issues are solvable and are typical of technology needing to adapt and scale after development, but associated costs must be considered during the design phase.

The use of historical EcoXPT data also presents limitations for the validity and generalizability of these findings. In both prior studies described in Chapter 3, entire classes were randomly assigned by teacher so that half of every teacher's classes were EcoXPT and half were the modified version or comparison curriculum. Especially in the case of an entirely different curriculum, this represented a significant increase in work and thus less time teachers could potentially devote to EcoXPT preparation. In this study, all teachers taught all sections with the identical features and curriculum. Additionally, this opportunistic sampling method resulted in a much larger sample size for students in the LENS condition than was available in the historical data. While efforts have been made in the modeling and statistical tests used to account for the imbalanced sample, this is not ideal. Teacher recruitment was challenging for testing the LENS condition and it was originally thought that only one school would be participating, thus all classes were assigned the modified version. Ideally (or in a follow-up study), random assignment within teachers would be done as in the historical data to get a more accurate picture of what benefits and drawbacks are associated with the addition of the LENS tools.

Another consistent issue in this work is the potential for the LENS tools to provide advice that runs counter to some goals of authentic scientific inquiry. For example, Ranger Susan's advice suggests students explore the experimental tools to generate confirming and disconfirming evidence if she detects that a group has not yet used the tools after they are unlocked. While rather innocuous advice, some might argue that an ASI activity should not give

the advice to try all of the experimental tools at once. Instead, the focus should be on helping groups focus on answers to the questions that they are asking as part of the hypothesis that they are developing at any given time, mapping to the authentic context in which a scientist would be testing their hypothesis. These hints run the risk of turning the activity into a checklist or "workbook"-style activity where different parts are completed in sequence by all groups. This goes back to the earlier tension between structuring and problematizing via scaffolding (Reiser, 2004) and the LENS tools tend to currently focus (perhaps too much) on structuring. Some amount of struggle in intentional in problem-based learning, yet these feedback tools may have given students too much guidance where they felt reliant on it.

Much like the use of an opportunistic sample, the use of EcoXPT as a framework for the LENS suite of tools is also opportunistic. The design of an immersive virtual environment like EcoXPT is far beyond the scale of a dissertation, yet it served as a convenient test bed for embedding new feedback and assessment features. The curriculum and software were not originally designed for these purposes, and the use of historical data limited some of the modifications that could be made to the curriculum and pre-post assessment strategy. Opportunities to deal with complex, ill-structured problems are essential for preparing scientists to enter the workforce, but these opportunities can also make students in high-pressure settings uncomfortable. The design of EcoXPT is based on thorough interviews with ecosystem scientists and a deep understanding of the nature of science (Kamarainen & Grotzer, 2019) but that does not preclude the possibility of using different types of feedback to reduce the risk of unproductive struggle A curriculum and software suite designed from the ground up to incorporate formative feedback and passive assessment might do a better job at striking this balance than this current suite or tools.

Finally, this study was done with a rather homogenous sample of students from a district the EcoLEARN team has worked extensively with nearly ten years. Other studies conducted on EcoLEARN curricula have worked with diverse populations across New England, but these feedback and assessment tools have only been used with a small sliver of possible students and settings. In order to be effective for teachers internationally and valuable to the research community at large, the efficacy of interventions such as these need to be tested in different geographical regions with as wide a range of students across racial, gender, and socioeconomic lines. There currently exists a barrier between what interventions work at a research and development scale versus what can be deployed at scale (Behrens et al., 2019), and continuing to push these boundaries may provide evidence that funding the expensive development of immersive environments and games for learning is worthwhile. While that falls beyond the scale of a dissertation, securing funding for program evaluation of virtual environment-based feedback and assessment tools would be a good next step. This aligns with one of the Baker Learning Analytics Prizes that focuses on generalizability of models across different populations (Baker, 2019). Including more diverse populations during design and testing of these tools will help avoid tailoring to "roaming autodidacts" (Cottom, 2015) and assuming a baseline of western, male whiteness against which other learners are compared. Fairness of these tools can be evaluated with techniques such as slicing analysis (Gardner et al., 2019), which allows researchers to evaluate model efficacy across demographic groups. As these types of interventions scale, steps must be taken to minimize the risk of "re-identification" of anonymized student data (Fischer et al., 2020) and questions of data ownership, collection, and access will become important to reconsider over time and as legislation adjusts to novel uses of big data in education (Lynch, 2017).

**Future Work**

A simple next step for this type of work would be to expand what worked in the first implementation and measure any increased efficacy as a result. Expanding worked example tutorials in ASI activities to link to more experimentation concepts would likely be beneficial based on expressed teacher desires. Additionally, electronic concept mapping tools could use an in-world tutorial that gets groups fluent in the tool rapidly. More broadly, the role of worked examples in the curriculum could be re-examined. Some of the introductory lessons provide worked examples of how to test a hypothesis, but the just-in-time nature of the explanation via tutorial could be leveraged more in ASI open worlds. While the attempt at clustering student behaviors was not very informative here, other work has shown success using these methods in authentic scientific activities (Peffer et al., 2019), and more work in this area is warranted. Use of automated text scoring to generate customized feedback on scientific arguments has been shows to help students revise their arguments (Lee et al., 2019), and similar features could be implemented for student reflections in virtual world-based curriculum.

Additional work studying teacher behavior in their classes could shed more light on the viability and utility of teacher-facing dashboards or reports. Unlike students, there is no automated collection of data on teacher moves made in their classrooms while instructing ASI lessons. Valuable work on teacher fidelity of implementation (McGivney et al., 2019) has resulted in a qualitative coding rubric for teacher behavior that can be applied in conjunction with new tools. Many teachers do circulate while students use EcoXPT and readily detect off-task behaviors, but others may spend too much time in one place or not circulate sufficiently. Assessing if a dashboard or report may increasingly aid teachers who are not used to circulating

adeptly would be a valuable way of understanding how such features could impact instruction in different classrooms.

While showing teachers more detailed student data via the summary report was viewed as very helpful, one teacher's decision to show the visualization portion of the daily report to her students directly opens additional avenues of research as well as interesting ethical questions. From a research perspective, this teacher utilization of aggregate data may present internal review board issues for students who have permission to participate, but not to have their data used by researchers. If otherwise unavailable log file information is now being made available to teachers, it may behoove designers to show students these data as well. Student-involved data use required both teacher capacity to engage with data as well as the ability to foster that ability in their students, but the empowerment and increased motivation seen in students with these capacities is worth investigating further (Kennedy & Datnow, 2011; Jimerson et al., 2016). While the "big brother"-like effect of keeping students on-task with the knowledge that their actions could be observed by teachers and external researchers was seen as valuable and comical by some teachers in this study, this view of data-driven formative feedback is not an intended perspective on these features. A simple way to combat this is to make the data as transparent and visible for both students and teachers, and to build student capacity for evaluating their own learning data as well as water quality and experimental data from the curriculum.

While modeling group behaviors and shared representations of knowledge has potential to lead to effective group-level interventions, there is currently no way to know which group member is controlling the keyboard at any given time. In many computer-supported collaborative learning activities, there is often a "driver" that is more actively in control of the interface and activity, while a "passenger" assumes a more passive role and offers verbal

suggestions (Shaer et al., 2011). Knowing which member is at the keyboard at any given time could let researchers test if time is being split evenly, how time divisions vary by teacher, and if different amounts of time spent "driving" are related to differential learning gains or greater prior knowledge. It would also be possible to return to using multiple computers per group and assigning each group member a role within the multi-user virtual environment as done in EcoMUVE (Metcalf et al., 2013), but limitations of school hardware may prevent this, and the communication between group members is altered when each student has a device.

In a collaborative setting such as this, the logged data do not currently capture the rich dialog occurring within groups nor any of the non-verbal aspects of their collaboration (eye gaze, body language, etc.) Advances in multi-modal learning analytics have shown how low-cost, high frequency sensors can be used in classrooms to capture such data in real time and incorporate it alongside other features (Worsley & Blikstein, 2015), and these techniques are being increasingly used in game-based learning (Henderson et al., 2019). Drawing on the rich data and findings gathered from mixed methods and qualitative case studies of EcoLEARN curricula, novel quantitative information may be captured quickly and reliably that could augment log file data and improve model quality. A case study comparing sensor data gathered from focal pairs to qualitative observations made on a video recording of their activities would illuminate how these novel features align with human coding. These quantitative data could then be incorporated in feedback models to achieve a more holistic view of the group's dynamic and current understanding of the problem space. Sensor data can also be leveraged along with teacher dashboards to enhance orchestration systems (Dillenbourg & Jermann, 2010) that include sensor data (Muñoz-Cristóbal et al., 2017).

Increasing interest in the game-based learning community has focused on narrative-centered collaborative problem-based learning (Mott et al., 2019) and this work will generate open-ended data sets and continue to drive interest toward analytical techniques able to make sense of the less linear activities. In addition to the potential to expand learning analytics to support users, some of the narrative experiences themselves can adapt the flow of the narrative based on participant actions in the virtual world. Work such as this also aligns with more general efforts to design competency models for collaborative problem solving (Sun et al., 2020) that can evaluate how well co-located student groups co-construct shared knowledge and divide responsibilities dynamically. These analyses can look at how group dynamics change over time and can be used to analyze some of the potential driver/passenger issues observed in immersive virtual worlds.

## Chapter 6: Conclusions

In summary, this study shows via multiple methods that a suite of simple tools derived from group log files generated by an open-ended virtual environment for learning can positively impact the quality of group work completed and their understanding along certain dimensions. During comparisons with baseline EcoXPT, none of the newly introduced tools has shown to be deleterious to performance in any respect due to cognitive load or any other unforeseen consequence. In the process of analyzing this data, new opportunities for intervention have been identified, and the efficacy of extant tools can potentially be improved based on user feedback. Navigating the challenge of maintaining authenticity in scientific investigations while reducing the occurrence of unproductive struggle is an ongoing challenge. Collapsing a complex task such as that presented in EcoXPT to a series of "correct" moves to appease an algorithm is not a desired outcome, but the addition of dynamic scaffolding to the virtual world can potentially aid new teachers adopting such tools for the first time, as well as students for whom authentic scientific inquiry is an entirely foreign concept.

While the longitudinal, fine-grained log file data provides a deep understanding of certain aspects of the curriculum, it misses aspects of the group's collaboration and has no way to record teacher moves taking place during instruction. Ongoing work in NLP and MMLA can aid some of these shortcomings, but analyses based on logged events will always miss certain aspects of discourse or interaction that a teacher can readily observe. Conversely, the logged data can capture information otherwise invisible to the teacher and report it in a timely, digestible manner. The types of learning analytics presented in this dissertation are meant to empower teachers in their classrooms and encourage students to best utilize their nascent investigatory skills. These features can scaffold student learning in dynamic ways that fade as mastery of certain techniques

are demonstrated. Continued work on these types of feedback tools can aid the EcoLEARN group's suite of curricula by leveraging one of the main affordances of virtual environments that had not previously been utilized during implementations of the curricula.

In conclusion, this dissertation presents three major takeaway lessons for researchers and designers interested in formative feedback and automated assessment in ASI/PBL or other open-ended virtual learning environments:

1) It is possible to design tools such as those in the LENS suite, automate their use within an extant learning technology, and deploy them with a considerable level of fidelity. While there are benefits to designing a platform and curriculum with tools like these as a core feature, they still function well when added as a supplement to an already vetted virtual environment.
2) Do not assume teachers will use the tools as instructed. Each classroom and instructor is different and explicit guidance and instruction is needed to ensure new features are not used in ways that might be deleterious to desired outcomes or run contrary to core design philosophies.
3) Ensure that feedback aligns with overall curricular goals and classroom instruction as well as best practices from the literature. Any time these tools strayed from this guidance, the tool functioned poorly for students.

The successes of this suite of tools is a testament to the success of the design-based research philosophy embraced by the EcoLEARN team, as many features and tools were designed explicitly around prior feedback from teachers and research findings from previous implementations. Co-designing tools for their use and soliciting honest feedback at every step of the process is an essential component towards achieving meaningful results in learning analytics.

Continuing to triangulate model output with observations and interviews is the only way to ensure the utility and veracity of these results. Standing on the shoulders of a decade of prior work in this group as well as several decades of related work in ITS and game-based learning, this study will hopefully be of use to other developers of virtual environments for learning who hope to embrace inquiry or open-ended collaborative learning.

## Appendix A: EcoXPT Survey

Which of the following is an example of an experiment?

- Drawing a graph that shows the levels of different substances (like nitrates) found in a pond.
- Recording observations in a notebook of the animals that visit a specific part of a wetland.
- Running a computer simulation to predict the weather based on air temperature, wind speed and cloud cover.
- Adding different amounts of food to different fish tanks to see how it affects fish health.

Which of the following is an example of an experiment?

- Adding baking soda and vinegar to a model of a volcano to simulate an eruption.
- Planting some seeds in the sun and some in the shade and measuring how much they grow.
- Recording the number of birds that visit your bird feeder every day for a month.
- Installing sensors to measure the amount of air pollution in different neighborhoods.

Which of the following is an example of an experiment?

- Providing different kinds of toys in different hamster cages to see how toys affect hamster activity levels.
- Measuring and comparing the size of apples from wild apple trees to apples bought at the store to decide which is best for your health.
- Using a formula and mixing together two chemicals in a beaker to create a solution in the laboratory.
- Asking 10 neighbors about their opinions on a proposed community recycling program to report back to the city council.

Scientists may use multiple methods to investigate an ecosystem. Below is a list of possible methods an ecosystem scientist might use to investigate changes taking place in a wetland. Their goal is to generate evidence to support an explanation.

*Rate how strongly you agree with each option (strongly disagree, somewhat disagree, somewhat agree, strongly agree):*

"To generate evidence to support an explanation, an ecosystem scientist would..."
... put a harmless chemical in the water of the wetland to see how it moves.
... conduct an experiment outside in the wetland.
... make observations and take notes about changes in the wetland.
... put up signs that ask visitors to pick up litter they find near the wetland.
... set up an experiment in a lab to create conditions similar to the wetland.
... build a computer simulation to explore the relationships among factors in the wetland.

An ecosystem scientist is interested in whether drought affects plants in a grassland ecosystem. (note – drought is a period of below-average rainfall). Below is a list of possible methods she might use to investigate changes taking place in the grassland. Her goal is to generate evidence to support an explanation.

Development of this survey was supported by the National Science Foundation through Grant No. 1416781 and was developed by the EcoXPT research team led by Drs. Tina Grotzer and Chris Dede along with Drs. Amy Kamarainen and Shari Metcalf. Reprinted with permission.

*Rate how strongly you agree with each option (strongly disagree, somewhat disagree, somewhat agree, strongly agree):*

"To generate evidence to support an explanation, an ecosystem scientist would..."
... set up a lab experiment to control the amount of water plants get and measure how they respond.
... add a harmless chemical to the water to measure how much water is taken up by plants.
... run a computer simulation to predict rates of plant death given different amounts of rain.
... create an outdoor experiment to mimic a drought and see what happens to the plants.
... set up a system to water the grassland to make sure there aren't any more droughts.
... make observations about the number and kinds of plants that grow in the grassland.

An ecosystem scientist is investigating changes in an ecosystem. She will use multiple approaches to generate evidence to support an explanation.

*Rate how strongly you agree with each option (strongly disagree, somewhat disagree, somewhat agree, strongly agree):*

"To generate evidence to support an explanation, an ecosystem scientist would..."
... release a natural chemical into the ecosystem to see where it goes.
... set up an experiment outside in the ecosystem to see how weather affects it.
... build a model or simulation to better understand the system.
... do what she can to stop the changes from happening.
... make observations and take notes to understand what is typical.
... set up an experiment in a lab to control the conditions.

Data show that in countries where people eat more chocolate, there are more college graduates. The quotes below tell what other students decided must be true based only on the underlined information.

What do you think about each statement? Choose agree or disagree for each statement:
"The data show that eating chocolate causes you to graduate from college."
"The data show that college students like to eat chocolate."
"The data show that people who go to college have more money to buy chocolate."
"The data show that there is a correlation between higher levels of chocolate eating and higher levels of college graduates."
"The data show that eating chocolate helps you to study and therefore to graduate."

Choose which of these two students you agree with more:

Student A: "There may be no causal relationship at all between the amount of chocolate eaten and the number of college graduates. It may be a coincidence."

Student B: "There has to be a causal relationship between the amount of chocolate eaten and the number of college graduates. Otherwise they would not both go up at the same time."

Choose which of these two students you agree with more:

Student C: "When two variables go up and down at the same time, it proves that one causes the other to change."

Student D: "When two variables go up and down at the same time, a third variable might cause both of them to go up and down at the same time."

Development of this survey was supported by the National Science Foundation through Grant No. 1416781 and was developed by the EcoXPT research team led by Drs. Tina Grotzer and Chris Dede along with Drs. Amy Kamarainen and Shari Metcalf. Reprinted with permission.

Choose which of these two students you agree with more:

Student E: “Chocolate eating and college graduation both go up and down at the same time, so it proves that one causes the other to change.”

Student F: “Something else might cause both higher amounts of chocolate to be eaten and higher numbers of college graduates.”

Data show that when rainfall levels go up, the numbers of squirrels increase. The quotes below tell what other students decided must be true based only on the underlined information.
What do you think about each statement? Choose agree or disagree for each statement.
"The data show that squirrels need a lot of water to live."
"The data show that squirrels are happier when it rains."
"The data show that squirrels live high in trees so they don’t wash away in floods."
"The data show that there is a correlation between higher amounts of squirrels and higher amounts of rain."
"The data show that more rain results in more plants and that causes more food so more squirrels survive."

Choose which of these two students you agree with more:

Student A: "There may be no causal relationship at all between the number of squirrels and the amount of rain. It may be a coincidence."

Student B: "There has to be a causal relationship between the number of squirrels and the amount of rain. Otherwise they would not both go up at the same time."

Choose which of these two students you agree with more:

Student C: "When two variables go up and down at the same time, it proves that one causes the other to change."

Student D: "When two variables go up and down at the same time, a third variable might cause both of them to go up and down at the same time."

Choose which of these two students you agree with more:

Student E: "The number of squirrels and the levels of rain both go up and down at the same time, so it proves that one causes the other to change."

Student F: "Something else might cause both higher numbers of squirrels and higher levels of rain."

Data show that when algae levels in a pond go up, the numbers of daphnia increase. The quotes below tell what other students decided must be true based only on the underlined information.
What do you think about each statement? Choose agree or disagree for each statement.
"The data show that daphnia need algae to live."
"The data show that daphnia like when there is a lot of algae."
"The data show that daphnia hide in the algae so that they don’t get eaten by predators."
"The data show that there is a correlation between higher amounts of algae and higher amounts of daphnia."
"The data show that more algae result in more food for things that eat daphnia so more daphnia survive."

Choose which of these two students you agree with more:

Student A: “There may be no causal relationship at all between the amount of algae in a pond and the number of daphnia. It may be a coincidence.”

Development of this survey was supported by the National Science Foundation through Grant No. 1416781 and was developed by the EcoXPT research team led by Drs. Tina Grotzer and Chris Dede along with Drs. Amy Kamarainen and Shari Metcalf. Reprinted with permission.

Student B: “There has to be a causal relationship between the amount of algae in a pond and the number of daphnia. Otherwise they would not both go up at the same time.”

Choose which of these two students you agree with more:

Student C: “When two variables go up and down at the same time, it proves that one causes the other to change.”

Student D: “When two variables go up and down at the same time, a third variable might cause both of them to go up and down at the same time.”

Choose which of these two students you agree with more:

Student E: “The levels of algae and the levels of daphnia both go up and down at the same time, so it proves that one causes the other to change.”

Student F: “Something else might cause both higher amounts of algae and higher numbers of daphnia.”

Imagine that one day there are many dead frogs at the edge of a local pond. Below are a few statements made by different residents of the neighborhood near the pond, suggesting what to do to find out why the frogs died.

*Show how much you agree or disagree with the different statements (Disagree, Somewhat disagree, Neutral, Somewhat agree, Agree):*

We need to look for clues in areas next to the pond, like neighborhoods.
It's a bad idea to only focus on the things at the edge of the pond.
It is important to search for clues far away from the edge of the pond.
Looking for clues far away from the edge of the pond is a waste of time.
The most important clues can be found at the edge of the pond.
We need to think about what was happening a few days before we found the dead frogs.
It's a bad idea to only look for clues from the day we found the dead frogs.
We can find out what happened to the frogs if we try to find clues from days before they died.
Thinking about what has been going on in the days before the frogs died is a waste of time.
The most important clues can be found on the day that the frogs died.
We need to focus on the things we can see.
It's a bad idea to only focus on the things we can see.
It's important to look for hidden things when trying to discover why the frogs died.
Looking for things we can't see is a waste of time.
The most important clues are the ones we can see easily.
We need to find the people who made the frogs die.
Something specific must have caused the frogs to die.
Someone must be responsible for killing the frogs.
We need to find the one thing that made the frogs die.
A person or animal probably killed the frogs.

Development of this survey was supported by the National Science Foundation through Grant No. 1416781 and was developed by the EcoXPT research team led by Drs. Tina Grotzer and Chris Dede along with Drs. Amy Kamarainen and Shari Metcalf. Reprinted with permission.

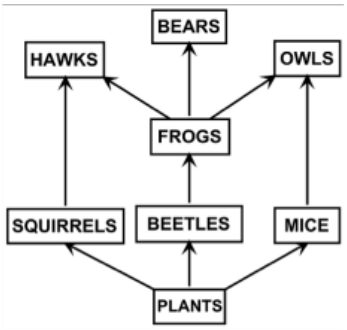
Which is a living factor within an ecosystem?
- The type of climate in a given region
- The animals that consume other animals
- The amount of helium gas in the air
- The rate of flow of water in a river

During photosynthesis, plants use sunlight and _____?
- Release oxygen into the environment.
- Take carbon out of the soil.
- Break down nitrogen into energy.
- Release carbon dioxide in the air.

Which is a nonliving factor within an ecosystem?
- The average rainfall in the ecosystem
- The animals that consume other animals in the ecosystem
- The number of trees in the ecosystem
- The amount of bacteria in the ecosystem

Which is released as a result of plants doing photosynthesis?
- Carbon dioxide
- Hydrogen
- Oxygen
- Water

The diagram shows the feeding relationships between populations of organisms in an area. The arrows point from the organisms being eaten to the organisms that eat them.




Using only the relationships between the organisms shown in the diagram, which of the following populations of organisms could be affected if the number of frogs changes?
- Only the frog population could be affected.
- Only the populations of plants, beetles, and bears could be affected.
- Only the populations of beetles, hawks, bears, and owls could be affected.
- The populations of all of the organisms shown in the diagram could be affected.

How do decomposers obtain their food?
- By consuming living plant materials.
- By hunting and killing other organisms.
- By absorbing food from dead organisms.
- By producing food from oxygen and sunlight.

Which process uses oxygen to break apart sugars for energy?
- Circulation
- Digestion
- Photosynthesis
- Respiration

Development of this survey was supported by the National Science Foundation through Grant No. 1416781 and was developed by the EcoXPT research team led by Drs. Tina Grotzer and Chris Dede along with Drs. Amy Kamarainen and Shari Metcalf. Reprinted with permission.

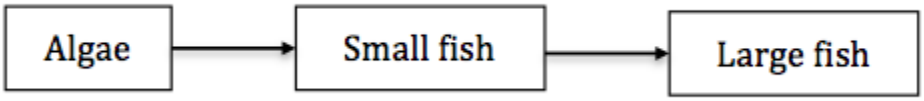
Which of these best describes what happens to dead matter in an ecosystem?

"Nothing happens to it because it is dead."
"It gets buried in the soil or mud and stays there."
"It is consumed by tiny bacteria that are too small to see."
"Rotting makes the dead matter disappear."

The diagram below shows the feeding relationships between insects, birds, and cats.


Which of the following statements best describes what will happen to the number of birds if most of the cats move away from a neighborhood?

The number of birds will increase because there will be fewer cats to eat them.
The number of birds will decrease because there will be fewer cats for the birds to eat.
The number of birds will stay the same because changes in the number of cats will not affect them.
The number of birds will decrease because the birds will follow the cats.

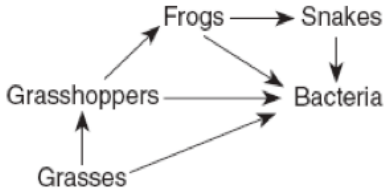



If all of the small fish in the pond died suddenly from a disease that killed only the small fish, what would happen in the next month to the large fish in the pond?

There would be more large fish in the pond.
There would be fewer large fish in the pond.
The amount of large fish would be the same.
The amount of large fish is not affected by the small fish.

Use the food web to answer the following question.
According to the diagram, which of the following living things would be affected if the grasses decreased?

only grasshoppers would be affected
only grasshoppers and bacteria would be affected
only grasshoppers, bacteria, and frogs would be affected
grasshoppers, bacteria, frogs, and snakes would be affected




Which of the following best describes what happens as decomposers (like bacteria) get energy to live?

they make animals sick during the process
they release water during the process
they use energy from the sun during the process
they take up oxygen during the process

Development of this survey was supported by the National Science Foundation through Grant No. 1416781 and was developed by the EcoXPT research team led by Drs. Tina Grotzer and Chris Dede along with Drs. Amy Kamarainen and Shari Metcalf. Reprinted with permission.

*Click the button that shows how confident you are that you can do the following science tasks.*
*How confident are you... (not at all confident = 1, completely confident = 6)*
... that you can investigate what causes change in an environment?
... that you can do the kinds of things that scientists do?
... that you can look at data that you collect and notice patterns?
... that you can investigate the relationships between organisms and their environment?
... that you can investigate the reasons why things happen in nature?
... that you can use tables and graphs to figure things out?
... that you can investigate the kinds of problems that an ecosystem scientist would investigate?

*Click the button that best describes how true or false each statement is for you.*
*(Definitely False, Mostly False, A Little Bit False, A Little Bit True, Mostly True, Definitely True)*

Being involved in science is a key part of who I am.
I can see science-related activities as being a part of my future.
I consider myself a science person.
I can imagine myself being involved in a science-related career.
I am interested in learning about nature.
I have no interest in learning about the environment.
I would like to learn more about environmental science.
I am interested in learning about ecosystems.

*Click the button that best describes how much you agree with each statement.*
*(Strongly Disagree, Disagree, Somewhat Disagree, Somewhat Agree, Agree, Strongly Agree)*

The most reliable scientific knowledge is the kind that has been confirmed by many types of evidence.
To decide whether something I read about science is true, I have to check if other sources of evidence say the same thing.
To find claims that are false in science, it is important to check several different sources of information.
One test is usually enough to decide what is right in science.

Two students are arguing about the best way to do science.
Brandon: In order to do good science, scientists have to do controlled experiments to test their hypotheses.
Jamal: While experiments are one important way to do science, the strongest science ideas are supported by multiple kinds of evidence.
Brandon: Ok, but if a scientist does one good experiment to prove the idea, that can be enough.

With whom do you agree? Read all the choices before choosing one.

- I agree almost entirely with Brandon.
- Although I agree more with Brandon, I think Jamal makes some good points.
- I agree (or disagree) equally with Jamal and Brandon.
- Although I agree more with Jamal, I think Brandon makes some good points.
- I agree almost entirely with Jamal.

Development of this survey was supported by the National Science Foundation through Grant No. 1416781 and was developed by the EcoXPT research team led by Drs. Tina Grotzer and Chris Dede along with Drs. Amy Kamarainen and Shari Metcalf. Reprinted with permission.

Two students are arguing about the best way for a scientist to investigate changes in an ecosystem.
Emma: As long as a scientist does an experiment to test their hypothesis, they will figure out what is causing changes in the ecosystem.
Jasmine: In order to figure out what causes change in an ecosystem, scientists need to consider whether the results of their experiment match with other evidence.
Emma: Well, that might be true, but if the two don't match, then the scientist should base their conclusion on the results of the experiment.

With whom do you agree? Read all the choices before choosing one.

I agree almost entirely with Emma.
Although I agree more with Emma, I think Jasmine makes some good points.
I agree (or disagree) equally with Jasmine and Emma.
Although I agree more with Jasmine, I think Emma makes some good points.
I agree almost entirely with Jasmine.

Two students are discussing ecosystem science: Leticia: My own knowledge about ecosystem science may change as I learn more, but overall ecosystem scientists know how ecosystems work. Nisha: I agree that my knowledge about ecosystem science may change as I learn more, but I also think scientists are always learning new things about how ecosystems work.
With whom do you agree? Read all the choices before choosing one.

I agree almost entirely with Leticia.
I agree more with Leticia, but I think Nisha makes a good point.
I agree (or disagree) equally with Nisha and Leticia.
I agree more with Nisha, but I think Leticia makes a good point.
I agree almost entirely with Nisha.

Jose: In my opinion, science ideas can change. Things we think are "facts" today may turn out to wrong later.
Miguel: I have a different opinion. Once experiments have been done and an explanation has been made, the matter is pretty much settled. There's little room for argument.

With whom do you agree? Read all the choices before choosing one.

I agree almost entirely with Jose.
Although I agree more with Jose, I think Miguel makes some good points.
I agree (or disagree) equally with Miguel and Jose.
Although I agree more with Miguel, I think Jose makes some good points.
I agree almost entirely with Miguel.

Lupe: Some scientists think the dinosaurs died out because of volcanic eruptions, and others think they died out because an asteroid hit the Earth. Why can't the scientists agree?
Samantha: Maybe the evidence supports both ideas. There's often more than one way to interpret the facts. So we have to figure out what the facts mean.
Lupe: I'm not so sure. In stuff like personal relationships or poetry, things can be ambiguous. But in science, the facts speak for themselves.

Development of this survey was supported by the National Science Foundation through Grant No. 1416781 and was developed by the EcoXPT research team led by Drs. Tina Grotzer and Chris Dede along with Drs. Amy Kamarainen and Shari Metcalf. Reprinted with permission.

With whom do you agree? Read all the choices before choosing one.

I agree almost entirely with Lupe.
I agree more with Lupe, but I think Samantha makes some good points.
I agree (or disagree) equally with Samantha and Lupe.
I agree more with Samantha, but I think Lupe makes some good points.
I agree almost entirely with Samantha.

Development of this survey was supported by the National Science Foundation through Grant No. 1416781 and was developed by the EcoXPT research team led by Drs. Tina Grotzer and Chris Dede along with Drs. Amy Kamarainen and Shari Metcalf. Reprinted with permission.

## Appendix B: EcoXPT Daily Activities

Day One: Students begin exploring EcoXPT and focus on getting to know the layout of the world, what organisms live there (both micro- and macroscopic), and how the field guide tool works. They are introduced to a thinking move called Deep Seeing. The camera, field guide, submarine, and notebook tools are available.

Day Two: Students continue exploring EcoXPT and focus on traveling over time and seeing what can be learned on different days. They may also start collecting water quality measurements and gathering data for those measurements across time. The weather tool, population tool, and Data View are also unlocked on the second day and some students will find them and use them. They will be more formally introduced on Day Three.

Day Three: Students discover the fish die-off, focus on their initial hypotheses about what may have happened, and begin collecting evidence in support of their hypotheses. They are introduced the Evidence Seeking move. As they collect pieces of information, or evidence for what might be happening in the world, they are able to collect evidence in relation to each claim. The opening PPT draws their attention to the Population Tool, Data View, and Weather Tool.

Day Four: Students continue seeking evidence in support of their ideas about what happened to the fish. They are introduced to the Pattern Seeking move as they explore patterns in the data that suggest what might be going on.

Day Five: Students continue seeking evidence in support of their ideas about what happened to the fish. They are introduced to a concept mapping tool that will help them to make possible connections and seek evidence for each claim represented in their concept map.

Day Six: Students continue seeking evidence in support of their ideas about what happened to the fish and exploring patterns in the data that suggest what might be going on. Once they have discovered patterns between algae, bacteria, and the fish die off, they are introduced to the differences between correlation and causation and the Analyzing Causality move. All experimental tools are unlocked so that they can begin to conduct experiments to confront some problems in reasoning only from patterns and will begin to see how it is important to explore the mechanisms behind the patterns.

Day Seven: Students focus on asking questions about what might be going on in the ecosystem and investigating what might be happening. They continue working with the Evidence Seeking and Analyzing Causality moves to hypothesize about what might have happened in the world. The Atom Tracker is introduced.

Day Eight: Students reflect on what they do and do not know with a focus on getting the information they need to understand what is going on. As part of a class discussion, they

consider the difference between seeing patterns and determining causality. They continue to refine their questions and make sure they have evidence to back up their claims.

Day Nine: Students continue constructing their explanations, conducting experiments, and using the evidence from their experiments to understand what is going on in the ecosystem. They are introduced to the Constructing Explanations move. It is used along with the Concept Mapping tool to support them in making sense of the "big picture" as they put all of their clues together.

Day Ten: Students begin compiling their evidence and preparing to present their work to others. Students focus on building the fullest explanation that they can with their concept maps. As they are working, the teacher circulates and helps them to find gaps in their explanation. They use confirming and disconfirming evidence to support their explanation. With help from the visual cues/codes in the concept maps, they reflect on the kinds of evidence they are using and determine if there may be information that is missing from their explanation.

Day Eleven: For the first third of class, students continue preparing their concept maps to present to the class. They make sure that all of their evidence is listed and that there are no gaps in their explanations. They include confirming and disconfirming evidence in their concept maps. The teacher then stops them and asks them to carefully review their evidence and concept maps. Students write an individual essay explaining what they think happened to the fish.

Day Twelve: Students share their findings for what happened at the pond. They listen carefully to each other's presentations to help their classmates discover what is well-supported in their arguments and where evidence for claims may be missing. If conducted as a whole class discussion, it is facilitated so that all of the students are able to contribute aspects of the complex causal scenario underlying what happened in the ecosystem. The session underscores that a good explanation is a well-supported, well-reasoned one in which the mechanisms for the causal connections are explained.

Day Thirteen: This is a day of reflection on the big lessons from EcoXPT. It is not about the explanation that they came up with but about the messages that they learned about science, ecosystems science, and coming up with an explanation. Students have an opportunity to reflect upon their own ideas and then the class discusses it.

## Appendix C: Teacher Interview Protocol

- (Remind teachers about newly added features) Did you see students using these new tools during their investigations?
- Did they ever mention them being helpful or confusing?
- Did students mention preferring automatic help or help on request?
- Did you use the reflection prompts with your classes? Did you find them helpful?
- How did you make use of the several teacher reports you received during the curriculum?
- Did the visualization at the top help direct your attention?
- Did the notes at the bottom help direct your attention?
- What behaviors did you notice when students were stuck?
- How do you distinguish between being stuck and productively struggling?
- What kinds of supports would you find most helpful in teaching EcoXPT?
- (new teachers) How does this type of instruction differ from what you're used to?
- (veteran teachers) How has your use of these kinds of curricula changed over time?

## Bibliography


Adams, D. M., & Clark, D. B. (2014). Integrating self-explanation functionality into a complex game environment: Keeping gaming in motion. *Computers & Education*, *73*, 149–159. https://doi.org/10.1016/j.compedu.2014.01.002

Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2016). Help Helps, But Only So Much: Research on Help Seeking with Intelligent Tutoring Systems. *International Journal of Artificial Intelligence in Education*, *26*(1), 205–223. https://doi.org/10.1007/s40593-015-0089-1

Andersen, E., Liu, Y.-E., Apter, E., Boucher-Genesse, F., & Popović, Z. (2010). Gameplay analysis through state projection. *Proceedings of the Fifth International Conference on the Foundations of Digital Games - FDG '10*, 1–8. https://doi.org/10.1145/1822348.1822349

Andersen, E., O'Rourke, E., Liu, Y.-E., Snider, R., Lowdermilk, J., Truong, D., Cooper, S., & Popovic, Z. (2012). The impact of tutorials on games of varying complexity. *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems - CHI '12*, 59. https://doi.org/10.1145/2207676.2207687

Anderson, T., & Shattuck, J. (2012). Design-Based Research: A Decade of Progress in Education Research? *Educational Researcher*, *41*(1), 16–25. https://doi.org/10.3102/0013189X11428813

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*(3), 603–617.

Baker, R. S. J. D., & Clarke-Midura, J. (2013). Predicting Successful Inquiry Learning in a Virtual Performance Assessment for Science. In S. Carberry, S. Weibelzahl, A. Micarelli, & G. Semeraro (Eds.), *User Modeling, Adaptation, and Personalization* (Vol. 7899, pp. 203–214). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-38844-6_17

Baker, R. S. J. d., D'Mello, S. K., Rodrigo, Ma. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, *68*(4), 223–241. https://doi.org/10.1016/j.ijhcs.2009.12.003

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, *19*(2), 185–224.

Baker, Ryan S. (2019). *Challenges for the Future of Educational Data Mining: The Baker Learning Analytics Prizes*. *11*(1), 17.

Baker, Ryan S., Ocumpaugh, J., Gowda, S. M., Kamarainen, A. M., & Metcalf, S. J. (2014). Extending Log-Based Affect Detection to a Multi-User Virtual Environment for Science. In V. Dimitrova, T. Kuflik, D. Chin, F. Ricci, P. Dolog, & G.-J. Houben (Eds.), *User Modeling, Adaptation, and Personalization* (Vol. 8538, pp. 290–300). Springer International Publishing. https://doi.org/10.1007/978-3-319-08786-3_25

Baker, Ryan Shaun, & Inventado, P. S. (2014). Educational Data Mining and Learning Analytics. In J. A. Larusson & B. White (Eds.), *Learning Analytics: From Research to Practice* (pp. 61–75). Springer New York. https://doi.org/10.1007/978-1-4614-3305-7_4

Barnes, T., & Stamper, J. (2008). Toward Automatic Hint Generation for Logic Proof Tutoring Using Historical Student Data. In B. P. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie (Eds.), *Intelligent Tutoring Systems* (Vol. 5091, pp. 373–382). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-69132-7_41

Bauer, A., & Popović, Z. (2017). Collaborative Problem Solving in an Open-Ended Scientific Discovery Game. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 1–21. https://doi.org/10.1145/3134657

Beals, K., & Willard, C. (2001). *Environmental Detectives. Grades 5-8. Teacher's Guide*. University of California, Berkeley, GEMS, Lawrence Hall of Science #5200, Berkeley, CA.

Behrens, J. T., DiCerbo, K. E., & Foltz, P. W. (2019). Assessment of Complex Performances in Digital Environments. *The ANNALS of the American Academy of Political and Social Science*, *683*(1), 217–232. https://doi.org/10.1177/0002716219846850

Belland, B. R., & Drake, J. (2013). Toward a framework on how affordances and motives can drive different uses of scaffolds: Theory, evidence, and design implications. *Educational*

*Technology Research and Development*, *61*(6), 903–925. https://doi.org/10.1007/s11423-013-9313-6

Ben-Naim, D., Bain, M., & Marcus, N. (2009). A User-Driven and Data-Driven Approach for Supporting Teachers in Reflection and Adaptation of Adaptive Tutorials. In *Proceedings for the 2nd International Conference on Educational Data Mining*, 21–30.

Bergstrom, K., Jenson, J., & de Castell, S. (2012). What's' choice'got to do with it? Avatar selection differences between novice and expert players of World of Warcraft and Rift. In *Proceedings of the International Conference on the Foundations of Digital Games*, 97–104.

Berland, M., Martin, T., Benton, T., Petrick Smith, C., & Davis, D. (2013). Using Learning Analytics to Understand the Learning Pathways of Novice Programmers. *Journal of the Learning Sciences*, *22*(4), 564–599. https://doi.org/10.1080/10508406.2013.836655

Bertling, M., Jackson, G. T., Oranje, A., & Owen, V. E. (2015). Measuring Argumentation Skills with Game-Based Assessments: Evidence for Incremental Validity and Learning. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Artificial Intelligence in Education* (pp. 545–549). Springer International Publishing.

Bjork, E. L., & Bjork, R. A. (2014). Making Things Hard on Yourself, But in a Good Way: Creating Desirable Difficulties to Enhance Learning. In M. A. Gernsbacher & J. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (2nd ed., pp. 59–68). Worth.

Brown, A. L., & Campione, J. C. (1994). Guided discovery in a community of learners. In *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 229–270). The MIT Press.

Bruckman, A. (1996). Finding One's Own Space in Cyberspace. *Technology Review*, *99*(1), 48–54.

Butler, D. L., & Winne, P. H. (1995). Feedback and Self-Regulated Learning: A Theoretical Synthesis. *Review of Educational Research*, *65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Calvo, R. A., & D'Mello, S. (2010). Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transactions on Affective Computing*, *1*(1), 18–37. https://doi.org/10.1109/T-AFFC.2010.1

Chen, J. A., Metcalf, S. J., & Tutwiler, M. S. (2014). Motivation and beliefs about the nature of scientific knowledge within an immersive virtual ecosystems environment. *Contemporary Educational Psychology*, *39*(2), 112–123. https://doi.org/10.1016/j.cedpsych.2014.02.004

Chen, J. A., Tutwiler, M. S., Metcalf, S. J., Kamarainen, A., Grotzer, T., & Dede, C. (2016). A multi-user virtual environment to support students' self-efficacy and interest in science: A latent growth model analysis. *Learning and Instruction*, *41*, 11–22. https://doi.org/10.1016/j.learninstruc.2015.09.007

Cheng, M.-T., Rosenheck, L., Lin, C.-Y., & Klopfer, E. (2017). Analyzing gameplay data to inform feedback loops in The Radix Endeavor. *Computers & Education*, *111*, 60–73. https://doi.org/10.1016/j.compedu.2017.03.015

Chi, M. T. H., & Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist*, *49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education*, *86*(2), 175–218. https://doi.org/10.1002/sce.10001

Clark, D. B., & Martinez-Garza, M. (2012). Prediction and Explanation as Design Mechanics in Conceptually Integrated Digital Games to Help Players Articulate the Tacit Understandings They Build through Game Play. In C. Steinkuehler, K. Squire, & S. Barab (Eds.), *Games, Learning, and Society: Learning and Meaning in the Digital Age* (pp. 279–305). Cambridge University Press; Cambridge Core. https://doi.org/10.1017/CBO9781139031127.023

Clark, D. B., Virk, S. S., Barnes, J., & Adams, D. M. (2016). Self-explanation and digital games: Adaptively increasing abstraction. *Computers & Education*, *103*, 28–43. https://doi.org/10.1016/j.compedu.2016.09.010

Clarke-Midura, J., & Dede, C. (2010). Assessment, Technology, and Change. *Journal of Research on Technology in Education*, *42*(3), 309–328. https://doi.org/10.1080/15391523.2010.10782553

Clarke-Midura, J., & Yudelson, M. V. (2013). Towards Identifying Students' Causal Reasoning Using Machine Learning. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial Intelligence in Education* (Vol. 7926, pp. 704–707). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-39112-5_93

Code, J., & Zap, N. (2017). Assessment in Immersive Virtual Environments: Cases for Learning, of Learning, and as Learning. *Journal of Interactive Learning Research*, *28*(3), 235–248.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Second). Lawrence Erlblaum Associates.

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*(4), 253–278. https://doi.org/10.1007/BF01099821

Cottom, T. (2015). Intersectionality and Critical Engagement with the Internet. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2568956

Cutumisu, M., Pilner Blair, K., Chin, D. B., & Lewis Schwartz, D. (2015). Posterlet: A Game-Based Assessment of Children's Choices to Seek Feedback and to Revise. *Journal of Learning Analytics*, *2*(1), 49–71. https://doi.org/10.18608/jla.2015.21.4

Dalgarno, B., & Lee, M. J. W. (2010). What are the learning affordances of 3-D virtual environments?: Learning affordances of 3-D virtual environments. *British Journal of Educational Technology*, *41*(1), 10–32. https://doi.org/10.1111/j.1467-8535.2009.01038.x

Datnow, A., & Hubbard, L. (2015). Teachers' Use of Assessment Data to Inform Instruction: Lessons from the Past and Prospects for the Future. *Teachers College Record*, *117*(4).

De Jong, T., & Lazonder, A. W. (2014). The Guided Discovery Learning Principle in Multimedia Learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 371–390). Cambridge University Press.

Deci, E. L., Eghrari, H., Patrick, B. C., & Leone, D. R. (1994). Facilitating internalization: The self-determination theory perspective. *Journal of Personality*, *62*(1), 119–142.

Dede, C., Grotzer, T. A., Kamarainen, A., & Metcalf, S. (2017). EcoXPT: Designing for Deeper Learning through Experimentation in an Immersive Virtual Ecosystem. *Journal of Educational Technology & Society*, *20*(4), 166–178. JSTOR.

Dede, C., Grotzer, T., Kamarainen, A., & Metcalf, S. (2019). Designing immersive authentic simulations that enhance motivation and learning: EcoLearn. In R. Feldman (Ed.), *Learning science: Theory, research, practice* (pp. 229–259). McGraw-Hill.

Dever, D. A., & Azevedo, R. (2019). Autonomy and Types of Informational Text Presentations in Game-Based Learning Environments. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Artificial Intelligence in Education* (Vol. 11625, pp. 110–120). Springer International Publishing. https://doi.org/10.1007/978-3-030-23204-7_10

Dillenbourg, P., & Jermann, P. (2010). Technology for Classroom Orchestration. In M. S. Khine & I. M. Saleh (Eds.), *New Science of Learning* (pp. 525–552). Springer New York. https://doi.org/10.1007/978-1-4419-5716-0_26

Dzikovska, M., Steinhauser, N., Farrow, E., Moore, J., & Campbell, G. (2014). BEETLE II: Deep Natural Language Understanding and Automatic Feedback Generation for Intelligent Tutoring in Basic Electricity and Electronics. *International Journal of Artificial Intelligence in Education*, *24*(3), 284–332. https://doi.org/10.1007/s40593-014-0017-9

Elby, A., Frederiksen, J., Schwarz, C., & White, B. (1997). *EBAPS: epistemological beliefs assessment for physical sciences*. 24–28.

Fayer, S., Lacey, A., & Watson, A. (2017). *STEM Occupations: Past, Present, And Future*. U.S. Bureau of Labor Statistics.

Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smythe, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining Big Data in Education: Affordances and Challenges. *Review of Research in Education*.

Fishman, B., & Dede, C. (2016). Teaching and technology: New tools for new times. In D. Gitomer & C. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp. 1269–1334). Springer.

Freitas, S. de, & Neumann, T. (2009). The use of 'exploratory learning' for supporting immersive learning in virtual environments. *Computers & Education*, *52*(2), 343–352. https://doi.org/10.1016/j.compedu.2008.09.010

Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge - LAK19*, 225–234. https://doi.org/10.1145/3303772.3303791

Gobert, J. D., & Sao Pedro, M. (2016). Digital assessment environments for scientific inquiry practices. In *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications.* (pp. 508–534). Wiley-Blackwell.

Gross, S., Zhu, X., Hammer, B., & Pinkwart, N. (2012). Cluster Based Feedback Provision Strategies in Intelligent Tutoring Systems. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Intelligent Tutoring Systems* (pp. 699–700). Springer Berlin Heidelberg.

Großmann, N., & Wilde, M. (2019). Experimentation in biology lessons: Guided discovery through incremental scaffolds. *International Journal of Science Education*, *41*(6), 759–781. https://doi.org/10.1080/09500693.2019.1579392

Grotzer, T. A., Kamarainen, A. M., Tutwiler, M. S., Metcalf, S., & Dede, C. (2013). Learning to Reason about Ecosystems Dynamics over Time: The Challenges of an Event-Based Causal Focus. *BioScience*, *63*(4), 288–296. https://doi.org/10.1525/bio.2013.63.4.9

Grotzer, T. A., Metcalf, S. J., Tutwiler, M. S., Kamarainen, A. M., Thompson, M., & Dede, C. (2017, April). *Teaching the systems aspects of epistemologically authentic*

*experimentation in ecosystems through immersive virtual worlds*. National Association for Research in Science Teaching (NARST), San Antonio, TX.

Grotzer, T. A., Powell, M. M., M. Derbiszewska, K., Courter, C. J., Kamarainen, A. M., Metcalf, S. J., & Dede, C. J. (2015). Turning Transfer Inside Out: The Affordances of Virtual Worlds and Mobile Devices in Real World Contexts for Teaching About Causality Across Time and Distance in Ecosystems. *Technology, Knowledge and Learning*, *20*(1), 43–69. https://doi.org/10.1007/s10758-014-9241-5

Grotzer, TA, Tutwiler, M. S., Dede, C., Kamarainen, A., & Metcalf, S. (2011). Helping students learn more expert framing of complex causal dynamics in ecosystems using EcoMUVE. *National Association of Research in Science Teaching Conference*, *4*.

Grotzer, Tina, Basca, B., & Donis, K. (2002). *Causal patterns in ecosystems: Lessons to infuse into ecosystems units.* President and Fellows of Harvard College. https://pz.harvard.edu/sites/default/files/revised_ecosystem.pdf

Grotzer, Tina, Gonzalez, E., Kamarainen, A., Metcalf, S., & Dede, C. (2018, March). *Moving from Exploring Patterns to Causal Explanations in Ecosystems Science Reasoning*. National Association for Research in Science Teaching (NARST), Atlanta, GA.

Habgood, M. P. J., & Ainsworth, S. E. (2011). Motivating Children to Learn Effectively: Exploring the Value of Intrinsic Integration in Educational Games. *Journal of the Learning Sciences*, *20*(2), 169–206. https://doi.org/10.1080/10508406.2010.508029

Hadwin, A. F., & Winne, P. H. (2001). CoNoteS2: A Software Tool for Promoting Self-Regulation. *Educational Research and Evaluation*, *7*(2–3), 313–334. https://doi.org/10.1076/edre.7.2.313.3868

Halverson, R., & Owen, V. E. (2014). Game-based Assessment: An Integrated Model for Capturing Evidence of Learning in Play. *Int. J. Learn. Technol.*, *9*(2), 111–138. https://doi.org/10.1504/IJLT.2014.064489

Hamari, J., Shernoff, D. J., Rowe, E., Coller, B., Asbell-Clarke, J., & Edwards, T. (2016). Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Computers in Human Behavior*, *54*, 170–179. https://doi.org/10.1016/j.chb.2015.07.045

Hansen, J. D., & Reich, J. (2015). Democratizing education? Examining access and usage patterns in massive open online courses. *Science*, *350*(6265), 1245–1248.

Hao, J., Smith, L., Mislevy, R., von Davier, A., & Bauer, M. (2016). Taming Log Files From Game/Simulation-Based Assessments: Data Models and Data Analysis Tools: Taming Log Files From Game/Simulation-Based Assessments. *ETS Research Report Series*, *2016*(1), 1–17. https://doi.org/10.1002/ets2.12096

Henderson, N. L., Rowe, J. P., Mott, B. W., Brawner, K., Baker, R., & Lester, J. C. (2019). 4D Affect Detection: Improving Frustration Detection in Game-Based Learning with Posture-Based Temporal Data Fusion. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Artificial Intelligence in Education* (Vol. 11625, pp. 144–156). Springer International Publishing. https://doi.org/10.1007/978-3-030-23204-7_13

Hiebert, J., & Grouws, D. A. (2006). *The Effects of Classroom Mathematics Teaching On Students ' Learning*. In F. Lester (Ed.), Second Handbook of Research on Mathematics Teaching and Learning (pp. 371-404). Charlotte, NC: Information Age.

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and Achievement in Problem-Based and Inquiry Learning: A Response to Kirschner, Sweller, and. *Educational Psychologist*, *42*(2), 99–107. https://doi.org/10.1080/00461520701263368

Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. *Science Education*, *88*(1), 28–54. https://doi.org/10.1002/sce.10106

Holstein, K., McLaren, B. M., & Aleven, V. (2019). Co-Designing a Real-Time Classroom Orchestration Tool to Support Teacher–AI Complementarity. *Journal of Learning Analytics*, *6*(2), 27–52. https://doi.org/10.18608/jla.2019.62.3

Holstein, K., McLaren, B. M., & Aleven, V. (2017). Intelligent tutors as teachers' aides: Exploring teacher needs for real-time analytics in blended classrooms. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17*, 257–266. https://doi.org/10.1145/3027385.3027451

Huizenga, J., Admiraal, W., Akkerman, S., & Dam, G. ten. (2009). Mobile game-based learning in secondary education: Engagement, motivation and learning in a mobile city game. *Journal of Computer Assisted Learning*, *25*(4), 332–344.

Jimerson, J. B., Cho, V., & Wayman, J. C. (2016). Student-involved data use: Teacher practices and considerations for professional learning. *Teaching and Teacher Education*, *60*, 413–424.

Johnson, W. L., & Lester, J. C. (2018). Pedagogical Agents: Back to the Future. *AI Magazine*, *39*(2), 33–44. https://doi.org/10.1609/aimag.v39i2.2793

Kamarainen, A. M., & Grotzer, T. A. (2019). Constructing Causal Understanding in Complex Systems: Epistemic Strategies Used by Ecosystem Scientists. *BioScience*, *69*(7), 533–543. https://doi.org/10.1093/biosci/biz053

Kamarainen, A. M., Metcalf, S., Grotzer, T., Browne, A., Mazzuca, D., Tutwiler, M. S., & Dede, C. (2013). EcoMOBILE: Integrating augmented reality and probeware with environmental education field trips. *Computers & Education*, *68*, 545–556. https://doi.org/10.1016/j.compedu.2013.02.018

Kamarainen, A. M., Metcalf, S., Grotzer, T., & Dede, C. (2015). Exploring Ecosystems from the Inside: How Immersive Multi-user Virtual Environments Can Support Development of Epistemologically Grounded Modeling Practices in Ecosystem Science Instruction. *Journal of Science Education and Technology*, *24*(2–3), 148–167. https://doi.org/10.1007/s10956-014-9531-7

Kamphorst, B. A. (2017). E-coaching systems: What they are, and what they aren't. *Personal and Ubiquitous Computing*, *21*(4), 625–632. https://doi.org/10.1007/s00779-017-1020-6

Kaplinsky, R. (2015). *Productive Struggle*. 54th Northwest Mathematics Conference, Whistler, Canada.

Kapur, M. (2016). Examining Productive Failure, Productive Success, Unproductive Failure, and Unproductive Success in Learning. *Educational Psychologist*, *51*(2), 289–299. https://doi.org/10.1080/00461520.2016.1155457

Ke, F., Xie, K., & Xie, Y. (2016). Game-based learning engagement: A theory- and data-driven exploration: Game-based learning engagement. *British Journal of Educational Technology*, *47*(6), 1183–1201. https://doi.org/10.1111/bjet.12314

Kennedy, B. L., & Datnow, A. (2011). Student involvement and data-driven decision making: Developing a new typology. *Youth & Society*, *43*(4), 1246–1271.

Ketelhut, D. J., Nelson, B. C., Clarke, J., & Dede, C. (2010). A multi-user virtual environment for building and assessing higher order inquiry skills in science: Building and assessing higher order inquiry skills in science. *British Journal of Educational Technology*, *41*(1), 56–68. https://doi.org/10.1111/j.1467-8535.2009.01036.x

Koedinger, K. R., Brunskill, E., Baker, R. S. J. d., McLaughlin, E. A., & Stamper, J. (2013). New Potentials for Data-Driven Intelligent Tutoring System Development and Optimization. *AI Magazine*, *34*(3), 27. https://doi.org/10.1609/aimag.v34i3.2484

Kolodner, J. L., Said, T., Wright, K., & Pallant, A. (2017). Drawn into Science Through Authentic Virtual Practice. *Proceedings of the 2017 Conference on Interaction Design and Children - IDC '17*, 385–391. https://doi.org/10.1145/3078072.3079751

Krajcik, J. S., & Mun, K. (2014). Promises and challenges of using learning technologies to promote student learning of science. In N. G. Lederman & S. K. Abell (Eds.), *Handbook of Research on Science Education* (pp. 337–360). Routledge.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863

Laney, D. (2001, February 6). *3D Data Management: Controlling Data Volume, Velocity, and Variety*. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Langdon, D., McKittrick, G., Beede, D., Khan, B., & Doms, M. (2011). *STEM: Good Jobs Now and for the Future. ESA Issue Brief #03-11*. US Department of Commerce. https://eric.ed.gov/?id=ED522129

Le, N.-T., Strickroth, S., Gross, S., & Pinkwart, N. (2013). A Review of AI-Supported Tutoring Approaches for Learning Programming. In N. T. Nguyen, T. van Do, & H. A. le Thi (Eds.), *Advanced Computational Methods for Knowledge Engineering* (Vol. 479, pp. 267–279). Springer International Publishing. https://doi.org/10.1007/978-3-319-00293-4_20

Lee, H., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., & Liu, O. L. (2019). Automated text scoring and real-time adjustable feedback: Supporting revision of scientific arguments involving uncertainty. *Science Education*, *103*(3), 590–622. https://doi.org/10.1002/sce.21504

Lee, S. J., Liu, Y.-E., & Popovic, Z. (2014). Learning Individual Behavior in an Educational Game: A Data-Driven Approach. *Proceedings of the 7th International Conference on Educational Data Mining*, 114–121.

Lee, S. Y., Mott, B. W., & Lester, J. C. (2012). Real-Time Narrative-Centered Tutorial Planning for Story-Based Learning. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Intelligent Tutoring Systems* (pp. 476–481). Springer Berlin Heidelberg.

Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., & Bhogal, R. S. (1997). *The persona effect: Affective impact of animated pedagogical agents*. 359–366.

Levy, S. T., & Wilensky, U. (2011). Mining students' inquiry actions for understanding of complex systems. *Computers & Education*, *56*(3), 556–573. https://doi.org/10.1016/j.compedu.2010.09.015

Li, H., Gobert, J., & Dickler, R. (2017a). Dusting Off the Messy Middle: Assessing Students' Inquiry Skills Through Doing and Writing. In E. André, R. Baker, X. Hu, Ma. M. T. Rodrigo, & B. du Boulay (Eds.), *Artificial Intelligence in Education* (Vol. 10331, pp. 175–187). Springer International Publishing. https://doi.org/10.1007/978-3-319-61425-0_15

Li, H., Gobert, J., & Dickler, R. (2017b). Automated Assessment for Scientific Explanations in On-line Science Inquiry. *Proceedings of the 10th International Conference on Educational Data Mining*, 214–219.

Lynch, C. F. (2017). Who prophets from big data in education? New insights and new challenges. *Theory and Research in Education*, *15*(3), 249–271. https://doi.org/10.1177/1477878517738448

Lynch, C. F., Merceron, A., Desmarais, M., & Nkambou, R. (Eds.). (2019). *Proceedings of the 12th International Conference on Educational Data Mining*.

Maltese, A. V., Simpson, A., & Anderson, A. (2018). Failing to learn: The impact of failures during making activities. *Thinking Skills and Creativity*, *30*, 116–124. https://doi.org/10.1016/j.tsc.2018.01.003

Mao, Y., Zhi, R., & Khoshnevisan, F. (2019). *One minute is enough: Early Prediction of Student Success and Event-level Difficulty during a Novice Programming Task*. 11.

McBroom, J., Yacef, K., Koprinska, I., & Curran, J. R. (2018). A Data-Driven Method for Helping Teachers Improve Feedback in Computer Programming Automated Tutors. In C. Penstein Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, & B. du Boulay (Eds.), *Artificial Intelligence in Education* (Vol. 10947, pp. 324–337). Springer International Publishing. https://doi.org/10.1007/978-3-319-93843-1_24

McGivney, E., Gonzalez, E., De Los Santos, S., Kamarainen, A., & Grotzer, T. (2019, April 2). *Improving Understanding of Teaching Practice for Student Learning: A Holistic Measure of Fidelity of Implementation*. National Association for Research in Science Teaching, Baltimore, MD. https://clic.gse.harvard.edu/files/clic/files/narst-ecoxptfidelity-4.2.2019.pdf

McLaren, B. M., van Gog, T., Ganoe, C., Yaron, D., & Karabinos, M. (2015). Worked Examples are More Efficient for Learning than High-Assistance Instructional Software. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Artificial Intelligence in Education* (Vol. 9112, pp. 710–713). Springer International Publishing. https://doi.org/10.1007/978-3-319-19773-9_98

McNeill, K., & Krajcik, J. (2008). Assessing middle school students' content knowledge and reasoning through written scientific explanations. *Assessing Science Learning: Perspectives from Research and Practice*, 101–116.

McNeill, K. L., & Krajcik, J. S. (2011). *Supporting Grade 5-8 Students in Constructing Explanations in Science: The Claim, Evidence, and Reasoning Framework for Talk and Writing*. Pearson.

Metcalf, S., Chen, J., Kamarainen, A., Frumin, K., Vickrey, T., Grotzer, T., & Dede, C. (2014). Shifts in Student Motivation during Usage of a Multi-User Virtual Environment for Ecosystem Science: *International Journal of Virtual and Personal Learning Environments*, *5*(4), 1–16. https://doi.org/10.4018/IJVPLE.2014100101

Metcalf, S. J., Kamarainen, A. M., Grotzer, T., & Dede, C. (2013). Teacher Perceptions of the Practicality and Effectiveness of Immersive Ecological Simulations as Classroom Curricula: *International Journal of Virtual and Personal Learning Environments*, *4*(3), 66–77. https://doi.org/10.4018/jvple.2013070105

Metcalf, S. J., Reilly, J. M., Kamarainen, A. M., King, J., Grotzer, T. A., & Dede, C. (2018). Supports for deeper learning of inquiry-based ecosystem science in virtual environments—Comparing virtual and physical concept mapping. *Computers in Human Behavior*, *87*, 459–469. https://doi.org/10.1016/j.chb.2018.03.018

Metcalf, S., Kamarainen, A. M., Grotzer, T., & Dede, C. (2013, June). Collaborative learning in virtual environments: Role-based exploration of causality in ecosystems over time and scale. *Computer-Supported Collaborative Learning*. 10th International Conference on Computer-Supported Collaborative Learning.

Metcalf, S., Kamarainen, A., Tutwiler, M. S., Grotzer, T., & Dede, C. (2011). Ecosystem Science Learning via Multi-User Virtual Environments. *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS)*, *3*(1), 86–90. https://doi.org/10.4018/jgcms.2011010107

Metcalf, S., Reilly, J., Dede, C., & Grotzer, T. (2019). *Linking Evidence and Concept Maps in Virtual Environments for Ecosystems Science Learning*. 92nd Annual International Conference of the National Association for Research in Science Teaching, Baltimore, MD.

Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Wiebe, E., Boyer, K. E., & Lester, J. C. (2015). DeepStealth: Leveraging Deep Learning Models for Stealth Assessment in Game-Based Learning Environments. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Artificial Intelligence in Education* (Vol. 9112, pp. 277–286). Springer International Publishing. https://doi.org/10.1007/978-3-319-19773-9_28

Min, W., Mott, B., Rowe, J., Liu, B., & Lester, J. (2016). Player Goal Recognition in Open-World Digital Games with Long Short-Term Memory Networks. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2590–2596.

Mott, B. W., Taylor, R. G., Lee, S. Y., Rowe, J. P., Saleh, A., Glazewski, K. D., Hmelo-Silver, C. E., & Lester, J. C. (2019). Designing and Developing Interactive Narratives for Collaborative Problem-Based Learning. In R. E. Cardona-Rivera, A. Sullivan, & R. M.

Young (Eds.), *Interactive Storytelling* (Vol. 11869, pp. 86–100). Springer International Publishing. https://doi.org/10.1007/978-3-030-33894-7_10

Muñoz-Cristóbal, J. A., Rodriguez Triana, M., Bote-Lorenzo, M. L., Villagrá-Sobrino, S. L., Asensio-Pérez, J. I., & Martínez-Monés, A. (2017). Toward multimodal analytics in ubiquitous learning environments. *Joint Proceedings of the Sixth Multimodal Learning Analytics (MMLA) Workshop and the Second Cross-LAK Workshop Co-Located with 7th International Learning Analytics and Knowledge Conference*, *1828*(CONF), 60–67.

National Research Council. (1996). *National Science Education Standards*. National Academy Press, 2101 Constitution Avenue, N.

National Research Council. (2011). *Successful K-12 STEM Education: Identifying Effective Approaches in Science, Technology, Engineering, and Mathematics*. National Academies Press. https://doi.org/10.17226/13158

Nelson, B. C. (2007). Exploring the Use of Individualized, Reflective Guidance In an Educational Multi-User Virtual Environment. *Journal of Science Education and Technology*, *16*(1), 83–97. https://doi.org/10.1007/s10956-006-9039-x

Nelson, B. C., & Ketelhut, D. J. (2007). Scientific Inquiry in Educational Multi-user Virtual Environments. *Educational Psychology Review*, *19*(3), 265–283. https://doi.org/10.1007/s10648-007-9048-1

Nelson, B. C., Ketelhut, D. J., Clarke, J., Dieterle, E., Dede, C., & Erlandson, B. (2007). Robust Design Strategies for Scaling Educational Innovations: The River City Case Study. *The Design and Use of Simulation Computer Games in Education*, 217–239. https://doi.org/10.1163/9789087903121_013

Nuzzo, R. (2014). Statistical Errors. *Nature*, *506*(7487), 150–152.

O'Neil, H. F., Chung, G. K. W. K., Kerr, D., Vendlinski, T. P., Buschang, R. E., & Mayer, R. E. (2014). Adding self-explanation prompts to an educational computer game. *Computers in Human Behavior*, *30*, 23–28. https://doi.org/10.1016/j.chb.2013.07.025

Paaßen, B., Hammer, B., Price, T. W., Barnes, T., Gross, S., & Pinkwart, N. (2017). The Continuous Hint Factory—Providing Hints in Vast and Sparsely Populated Edit Distance Spaces. *ArXiv:1708.06564 [Cs]*. http://arxiv.org/abs/1708.06564

Papendieck, A. (2018). *Technology for Equity and Social Justice in Education: A Critical Issue Overview*. https://doi.org/10.15781/T2891278V

Peffer, M., Quigley, D., & Mostowfi, M. (2019). Clustering Analysis Reveals Authentic Science Inquiry Trajectories Among Undergraduates. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge - LAK19*, 96–100. https://doi.org/10.1145/3303772.3303831

Pezzullo, L. G., Wiggins, J. B., Frankosky, M. H., Min, W., Boyer, K. E., Mott, B. W., Wiebe, E. N., & Lester, J. C. (2017). "Thanks Alisha, Keep in Touch": Gender Effects and Engagement with Virtual Learning Companions. In E. André, R. Baker, X. Hu, Ma. M. T. Rodrigo, & B. du Boulay (Eds.), *Artificial Intelligence in Education* (Vol. 10331, pp. 299–310). Springer International Publishing. https://doi.org/10.1007/978-3-319-61425-0_25

Price, T. W., Zhi, R., & Barnes, T. (2017). Hint Generation Under Uncertainty: The Effect of Hint Quality on Help-Seeking Behavior. In E. André, R. Baker, X. Hu, Ma. M. T. Rodrigo, & B. du Boulay (Eds.), *Artificial Intelligence in Education* (Vol. 10331, pp. 311–322). Springer International Publishing. https://doi.org/10.1007/978-3-319-61425-0_26

Pugh, K. J., Linnenbrink-Garcia, L., Koskey, K. L., Stewart, V. C., & Manzey, C. (2010). Motivation, learning, and transformative experience: A study of deep engagement in science. *Science Education*, *94*(1), 1–28.

Puntambekar, S., & Hubscher, R. (2005). Tools for Scaffolding Students in a Complex Learning Environment: What Have We Gained and What Have We Missed? *Educational Psychologist*, *40*(1), 1–12. https://doi.org/10.1207/s15326985ep4001_1

Reilly, J., & Dede, C. (2019). Exploring Stealth Assessment via Deep Learning in an Open-Ended Virtual Environment. *The 12th International Conference on Educational Data Mining*, 343–346.

Reilly, J., Kamarainen, A., Metcalf, S., Dede, C., & Grotzer, T. (2019). *The Importance of Time and Sequence on Learning in Mobile Augmented Reality*. 92nd Annual International Conference of the National Association for Research in Science Teaching, Baltimore, MD.

Reilly, J. M., & Dede, C. (2019). Differences in Student Trajectories via Filtered Time Series Analysis in an Immersive Virtual World. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 130–134. https://doi.org/10.1145/3303772.3303832

Reilly, J. M., Kumar, V., Berland, M., & Dede, C. (2018). Learning Analytics in a Teacher Dashboard to Facilitate Inquiry-Based Instruction. *Proceedings of the 2018 Connected Learning Summit*, 377–378.

Reilly, J., Metcalf, S., Studwell, J., Dede, C., & Grotzer, T. (2019). *Automatic Evaluation of Concept Maps from an Immersive Virtual World for Ecosystem Science Learning*. Annual Meeting of the American Educational Research Association, Toronto, Canada.

Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *The Journal of the Learning Sciences*, *13*(3), 273–304.

Renkl, A. (2014). The Worked Examples Principle in Multimedia Learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 391–412). Cambridge University Press.

Rivers, K., & Koedinger, K. R. (2013). Automatic Generation of Programming Feedback: A Data-Driven Approach. *The First Workshop on AI-Supported Education for Computer Science (AIEDCS 2013)*. AI in Education, Memphis, TN.

Rivers, K., & Koedinger, K. R. (2017). Data-Driven Hint Generation in Vast Solution Spaces: A Self-Improving Python Programming Tutor. *International Journal of Artificial Intelligence in Education*, *27*(1), 37–64. https://doi.org/10.1007/s40593-015-0070-z

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, *21*(2), 267–280. https://doi.org/10.1016/j.learninstruc.2010.07.004

Rowe, J. P., Mott, B. W., Mcquiggan, S. W., Robison, J. L., Lee, S., & Lester, J. C. (2009). Crystal Island: A Narrative-Centered Learning Environment for Eighth Grade Microbiology. *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 11–20.

Roy, M., & Chi, M. T. (2005). The self-explanation principle in multimedia learning. *The Cambridge Handbook of Multimedia Learning*, 271–286.

Sabourin, J., Rowe, J., Mott, B. W., & Lester, J. C. (2012). Exploring Inquiry-Based Problem-Solving Strategies in Game-Based Learning Environments. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Intelligent Tutoring Systems* (Vol. 7315, pp. 470–475). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30950-2_60

Sao Pedro, Baker, R. S., & Gobert, J. D. (2013). Incorporating Scaffolding and Tutor Context into Bayesian Knowledge Tracing to Predict Inquiry Skill Acquisition. *Proceedings of the 6th International Conference on Educational Data Mining*, 185–192.

Sao Pedro, M., Baker, R. S., Montalvo, O., Nakama, A., & Gobert, J. D. (2010). Using Text Replay Tagging to Produce Detectors of Systematic Experimentation Behavior Patterns. *Proceedings of the 3rd International Conference on Educational Data Mining*, 181–190.

Scalise, K. (2017). Hybrid Measurement Models for Technology-Enhanced Assessments Through mIRT-bayes. *International Journal of Statistics and Probability*, *6*(3), 168. https://doi.org/10.5539/ijsp.v6n3p168

Schwartz, R. S., & Crawford, B. A. (2006). Authentic Scientific Inquiry As Context For Teaching Nature Of Science: Identifying Critical Element. In L. B. Flick & N. G. Lederman (Eds.), *Scientific Inquiry and Nature of Science: Implications for Teaching, Learning, and Teacher Education* (pp. 331–355). Springer Netherlands. https://doi.org/10.1007/978-1-4020-5814-1_16

Sengupta-Irving, T., & Agarwal, P. (2017). Conceptualizing Perseverance in Problem Solving as Collective Enterprise. *Mathematical Thinking and Learning*, *19*(2), 115–138. https://doi.org/10.1080/10986065.2017.1295417

Shaer, O., Strait, M., Valdes, C., Feng, T., Lintz, M., & Wang, H. (2011). *Enhancing genomic learning through tabletop interaction*. 2817–2826.

Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Shute, V. J. (2011). Stealth Assessment in Computer-Based Games to Support Learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer Games and Instruction* (pp. 503–523). Information Age Publishing.

Shute, V. J., & Moore, G. R. (2017). Consistency and Validity in Game-Based Stealth Assessment. In H. Jiao & R. W. Lissitz (Eds.), *Technology Enhanced Innovative Assessment: Development, Modeling, and Scoring From an Interdisciplinary Perspective* (pp. 31–51). Information Age Publishing.

Siemens, G., & Baker, R. S. d. (2012). *Learning analytics and educational data mining: Towards communication and collaboration*. 252–254.

Stamper, J., Barnes, T., Lehmann, L., & Croy, M. (2008). The Hint Factory: Automatic Generation of Contextualized Help for Existing Computer Aided Instruction. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems Young Researchers Track*, 71–78.

Stamper, J. C., Eagle, M., Barnes, T., & Croy, M. (2011). Experimental Evaluation of Automatic Hint Generation for a Logic Tutor. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Artificial Intelligence in Education* (Vol. 6738, pp. 345–352). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-21869-9_45

Stamper, J., Eagle, M., Barnes, T., & Croy, M. (2013). Experimental evaluation of automatic hint generation for a logic tutor. *International Journal of Artificial Intelligence in Education*, *22*(1–2), 3–17.

Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S. (2020). Towards a generalized competency model of collaborative problem solving. *Computers & Education*, *143*, 103672. https://doi.org/10.1016/j.compedu.2019.103672

Sutton, R. S., & Barto, A. G. (1999). Reinforcement learning. *Journal of Cognitive Neuroscience*, *11*(1), 126–134.

Thompson, M., Tutwiler, M., Kamarainen, A., Metcalf, S., Grotzer, T., & Dede, C. (2016). A Blended assessment strategy for EcoXPT: An Experimentation-driven ecosystems science-based multi-user virtual environment. *American Educational Research Association (AERA), Washington DC*.

Tsai, F.-H., Tsai, C.-C., & Lin, K.-Y. (2015). The evaluation of different gaming modes and feedback types on game-based formative assessment in an online learning environment. *Computers & Education*, *81*, 259–269. https://doi.org/10.1016/j.compedu.2014.10.013

Tutwiler, M. Shane, Grotzer, T., Thompson, M., Kamarainen, A., Metcalf, S., & Dede, C. (2016). *Validation of an instrument measuring student complex causal assumptions*. 1.

Tutwiler, Michael Shane. (2014). *Trends in the Salience of Data Collected in a Multi User Virtual Environment: An Exploratory Study*. https://dash.harvard.edu/handle/1/13383549

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of Feedback in a Computer-Based Learning Environment on Students' Learning Outcomes: A Meta-Analysis. *Review of Educational Research*, *85*(4), 475–511. https://doi.org/10.3102/0034654314564881

Van Leeuwen, A., & Rummel, N. (2017). Teacher regulation of collaborative learning: Research directions for learning analytics dashboards. *Making a Difference: Prioritizing Equity and Access in CSCL*, *2*, 805–806.

Verbert, K., Govaerts, S., Duval, E., Santos, J. L., Van Assche, F., Parra, G., & Klerkx, J. (2013). Learning dashboards: An overview and future research opportunities. *Personal and Ubiquitous Computing*. https://doi.org/10.1007/s00779-013-0751-2

Wade, W. Y., Rasmussen, K. L., & Fox-Turnbull, W. (2013). Can Technology Be a Transformative Force in Education? *Preventing School Failure: Alternative Education for Children and Youth*, *57*(3), 162–170. https://doi.org/10.1080/1045988X.2013.795790

Wang, M, Derry, S., & Ge, X. (2017). Fostering Deep Learning in Problem-Solving Contexts with the Support of Technology. *Educational Technology & Society*, *20*(4), 162–165.

Wang, Minhong, Wu, B., Kirschner, P. A., & Michael Spector, J. (2018). Using cognitive mapping to foster deeper learning with complex problems in a computer-based environment. *Computers in Human Behavior*, *87*, 450–458. https://doi.org/10.1016/j.chb.2018.01.024

Warshauer, H. K. (2015). Productive struggle in middle school mathematics classrooms. *Journal of Mathematics Teacher Education*, *18*(4), 375–400. https://doi.org/10.1007/s10857-014-9286-3

Watters, A. (2019, December 31). *The 100 Worst Ed-Tech Debacles of the Decade*. Hack Education. http://hackeducation.com/2019/12/31/what-a-shitshow

Wijesooriya, C., Heales, J., & Clutterbuck, P. (2015). *Forms of Formative Assessment in Virtual Learning Environments*. Twenty-first Americas Conference on Information Systems, Puerto Rico.

Wolters, C. A. (2010). *Self-regulated learning and the 21st century competencies*. http://www.hewlett.org/uploads/Self_Regulated_Learning_21st_Century_Competencies.pdf

Wood, H., & Wood, D. (1999). Help seeking, learning and contingent tutoring. *Computers & Education*, *33*(2–3), 153–169. https://doi.org/10.1016/S0360-1315(99)00030-5

Worsley, M., & Blikstein, P. (2015). *Leveraging multimodal learning analytics to differentiate student learning strategies*. 360–367.

Xhakaj, F., Aleven, V., & McLaren, B. M. (2016). How Teachers Use Data to Help Students Learn: Contextual Inquiry for the Design of a Dashboard. In K. Verbert, M. Sharples, & T. Klobučar (Eds.), *Adaptive and Adaptable Learning* (Vol. 9891, pp. 340–354). Springer International Publishing. https://doi.org/10.1007/978-3-319-45153-4_26

Xhakaj, F., Aleven, V., & McLaren, B. M. (2017). Effects of a Teacher Dashboard for an Intelligent Tutoring System on Teacher Knowledge, Lesson Planning, Lessons and Student Learning. In É. Lavoué, H. Drachsler, K. Verbert, J. Broisin, & M. Pérez-Sanagustín (Eds.), *Data Driven Approaches in Digital Education* (Vol. 10474, pp. 315–329). Springer International Publishing. https://doi.org/10.1007/978-3-319-66610-5_23

Zacharia, Z. C., Manoli, C., Xenofontos, N., de Jong, T., Pedaste, M., van Riesen, S. A. N., Kamp, E. T., Mäeots, M., Siiman, L., & Tsourlidaki, E. (2015). Identifying potential types of guidance for supporting student inquiry when using virtual and remote labs in science: A literature review. *Educational Technology Research and Development*, *63*(2), 257–302. https://doi.org/10.1007/s11423-015-9370-0