Developing Computational Methods to Measure and Track Learners' Spatial Reasoning in an Open-Ended Simulation

Aditi Mallavarapu
University of Illinois at Chicago
amalla5@uic.edu

Leilah Lyons
University of Illinois at Chicago
llyons@uic.edu

Tia Shelley
University of Illinois at Chicago
tshell2@uic.edu

Emily Minor
University of Illinois at Chicago
eminor@uic.edu

Brian Slattery
University of Illinois at Chicago
bslatt2@uic.edu

Moria Zellner
University of Illinois at Chicago
mzellner@uic.edu

Interactive learning environments can provide learners with opportunities to explore rich, real-world problem spaces, but the nature of these problem spaces can make assessing learner progress difficult. Such assessment can be useful for providing formative and summative feedback to the learners, to educators, and to the designers of the environments. This work adds to a growing body of research that is applying EDM techniques to more open-ended problem spaces.

The open-ended problem space under study here was an environmental science simulation. Learners were confronted with the real-world challenge of effectively placing green infrastructure in an urban neighborhood to reduce surface flooding. Learners could try out different 2D spatial arrangements of green infrastructure and use the simulation to test each solution's impact on flooding. The learners' solutions and the solutions' performances were logged during a controlled experiment with different user interface designs for the simulation. As with many open-problem spaces, analyzing this data was difficult due to the large state space, many good solutions, and many alternate paths to those good solutions.

This work proposes a procedure for reducing the state space of solutions defined by 2D spatial patterns while maintaining their critical spatial properties. Spatial reasoning problems are a problem class not extensively examined by EDM, so this work sets the stage for further research in this area. This work also details a procedure for discovering effective 2D spatial strategies and solution paths, demonstrates how this information can be used to give formative feedback to the designers of the interactive learning environment, and speculates about how it could be used to provide formative feedback to learners.

1. INTRODUCTION

The increasing prevalence of technology in classrooms and other learning environments has the potential to both affect the way students learn things, as well as to help educators and educational designers get a better window onto the *processes* by which students learn things. This latter capacity, being able to track how students act within technological learning environments, has become increasingly important as our ability to create rich interactive learning experiences has outstripped our ability to design assessments. Teachers most often formatively assess learners' progress via observation or via strategies like pop quizzes, and summatively assess learners' performance via written paper tests. These formats don't easily cover the wide range of learning possible within an interactive learning environment. For example, learners can exhibit a range of skills and epistemic knowledge while engaged in a task that they could seldom learn from reading a textbook passage or express on a written test.

"Stealth assessment", a term usually applied to interactive simulations or games, is one approach to automate and embed assessment (Shute, 2011). In the "stealth assessment" approach, the design of the technological learning environment affects the types of observations of learner performances that are available for analysis, and good design of the learning environment can allow for a rich image of learner capabilities to be built through use. While there have been many examples of using data mining to track students' progress through interactive learning environments using log files (e.g., Harpstead, MacLellan, Koedinger, Aleven, Dow, & Myers, 2013; Andersen, Yun-En Liu, Apter, Boucher-Genesse, & Popovic´, 2010; Martinez-Maldonado, Yacef, & Kay, 2013; Gobert, Sao Pedro, Raziuddin, & Baker, 2013; Rafferty, Davenport, & Brunskill, 2013; Biswas, Loretz, & Segedy, 2013; Desmarais & Lemieux, 2013; Jarusek, Klusacek, & Pelanek, 2013; DiCerbo & Kidwai, 2013; Muller, Kretzschmar, & Greiff, 2013) most of these learning experiences are intentionally highly constrained so as to maximize the informational value of the logged observations. For example, learners may be given a well-defined, fixed goal where there is a known optimal number of steps to reach this goal, and there are a known, fixed number of choices that can be made by the learner. In such circumstances any user action can easily be judged as taking them closer to or farther away from the goal. This clarity often underpins the structure of Intelligent Tutoring Systems (ITS), which typically combine exhaustive, *a priori* models of the content domain and prior learner performance with models of the student's current progress to generate guidance (Van Lehn, 2011). These well-constrained problem spaces have successfully been used by data miners, who rely on *a priori* models and on *post hoc* analysis to provide formative feedback to the students (Gobert, Sao Pedro, Raziuddin, & Baker, 2013; Biswas, Loretz, & Segedy, 2013) or to their teachers (Martinez-Maldonado, Yacef, & Kay, 2013), to provide formative feedback to the environment designers (Harpstead, MacLellan, Koedinger, Aleven, Dow, & Myers, 2013; Martinez-Maldonado, Yacef, & Kay, 2013), or to provide evaluative feedback on the nature and scope of mistakes made by learners in the environment (Andersen, Yun-En Liu, Apter, Boucher-Genesse, & Popovic´, 2010; Rafferty, Davenport, & Brunskill, 2013). However, these constrained problem spaces often do not reflect problems found in the real world.

Real world problems often have many different solutions, which can be reached via many different paths. While presenting learners with simplified and constrained problems can be a good way to help them come to understand the core properties of a domain, exposing learners to less constrained, more open-ended problems can help them get experience with disciplinary

processes and dispositions. Many educational standards now recommend that learners be given opportunities to practice disciplinary processes and develop disciplinary dispositions which can be provided by project based student centered approaches (Schweingruber, Keller, & Quinn, 2012). One barrier to giving learners these experiences is that a lot more work is required to assess student progress within open-ended problem spaces. The main challenge for the researchers lies in how to represent the problem space so that meaningful relationships between the environment state and user actions can be mined to discover models of learner behavior.

The work presented here addresses a learning environment that is decidedly open-ended, with many degrees of freedom, a lack of path dependence in learner actions, and no clear prescriptions for good and bad solutions: an environmental simulation where the relative, not absolute, spatial placements of elements matter. The problem space we are confronted in our work has a very large state space – there are 324! different possible solutions (2.28899746 E+674), arising from the ability of learners to place items in 18x18 allowable spaces on a 22x22 grid. A large number of those solutions are likely to be fairly equivalent in terms of their performance, even though they may not be structurally similar at all. Our work, then, extends the current work on EDM for open-ended problems by 1) devising an approach for reducing the state space of a spatial problem with a genuinely large and non-path-dependent set of possible learner actions so that it would be tractable for analysis, (2) using this state-space reduction approach to discover the spatial strategies learners apply in the open-ended problem space, and (3) using the results of this strategy discovery to compare how different user interfaces can impact learners' use of those spatial strategies.

1.1. Context: a Collaborative Game to Support Spatial Pattern Reasoning in an Urban Planning Domain

The learners' challenge, drawn from urban planning and environmental science, is to integrate green infrastructure into an existing urban infrastructure in order to reduce surface flooding in urban areas. Most learners naively assume that the problem is a matter of matching green infrastructure capacity to the anticipated rainfall amount, but in reality, *where* the infrastructure is placed spatially dramatically affects its effectiveness. Ecologists have used two-dimensional spatial patterns to track environmental phenomena for decades (Dale, 1999). A problem-solving orientation that attends to the role of 2D spatial patterns in producing emergent environmental effects is an important disciplinary disposition for learners to develop, and urban planning researchers suggest that spatially-sensitive simulations may be a good platform for learners to acquire the disposition (Zellner, 2008). We thus developed a collaborative game (see Section 2.1 for more details) based on a simulated model of this problem to help learners develop this disciplinary disposition towards 2D spatial reasoning. We next turned our attention to what kind of user interface we should construct for the game.

Tangible User Interfaces (TUIs) are theorized to provide benefits for both spatial reasoning tasks (Kim & Maher, 2008; Antle, Droumeva, & Ha, 2009; Marshall, 2007) and for collaboration (Marshall, 2007; Schneider, Jermann, Zufferey, & Dillenbourg, 2011). Thus, we constructed a TUI for use with the simulation (Shelley, Lyons, Shi, Minor, & Zellner, 2010; Shelley, Lyons, Minor, & Zellner, 2011), but we wanted to experimentally determine if the TUI provided the theorized benefits for spatial reasoning, so we also built two control user interfaces to isolate the benefits (see Sections 2.1 and 3.2 for more details) and conducted an experiment. The work here analyzed log data collected during the experiment to see if the interface design impacted the

spatial strategies used by learners (we leave an examination of collaboration for another paper). EDM allowed us to go beyond summative evaluation (i.e., did one interface allow groups to produce better-performing solutions) and have a better window onto the specific two-dimensional spatial patterns the different interface designs subtly encouraged or discouraged, and how their use of patterns changed over time. As others have noted, it is sometimes more useful (for researchers, designers, educators, and learners) to see the evolution of a learner's strategy than to see only the learner's end solution (Blikstein, 2011). By employing EDM, we could better assess not just *whether* the interface design impacted spatial reasoning, but *how* the interface impacted spatial reasoning, as expressed by the evolving spatial patterns used by learners.

1.2. Approach and Contribution

This work explored a method to meaningfully and efficiently characterize the spatial patterns created by learners. A multivariate linear regression approach was then used to determine which patterns at which spatial scales were associated with improvements in rainwater capture. This marriage of a 2-dimensional spatial pattern and specific level of spatial scale is effectively what defines a "strategy" in this problem space, as a pattern that might have a large positive impact on outcomes at one level of scale might be ineffective or worse at another level of scale. It should be noted that we use the term "strategy" in a precise way to refer to spatial patterns that meaningfully affect the simulation outcomes, and that we do not use the term to refer to the conceptions that the learners themselves bring to the spatial patterns they're employing. While the learners' conceptions would certainly be of interest to us, we suspect that reliably eliciting these ideas would be difficult, as people can recognize and respond to spatial phenomena long before they acquire the vocabulary to describe it. Cognitive psychology has been studying how humans quickly recognize two-dimensional visual patterns since Max Wertheimer's landmark Gestalt psychology paper in 1923 (Wertheimer, 1923), and while more modern theories posit a role for "top-down" processes (like past experience) in affecting pattern detection, by and large it is still considered to be an automatic, low-level (and certainly pre-verbal) cognitive process. Thus, our approach may be able to detect learners' use of spatial strategies before they are able to articulate what they are doing, opening a role for formative feedback that we will address in this paper's conclusion.

Essentially, we used the data generated by learners interacting with the problem space to bootstrap the development of a model of how these novice learners engage with the problem space (i.e., the combinations of spatial strategies they found to be effective at capturing rainwater). We then used these results to examine if the user interface design affected the way in which learners approached exploring the problem space: did they use different spatial strategies, or discover them more quickly or more slowly, when using different user interfaces? We found that this was indeed the case - certain spatial strategies were more often present in some user interface conditions than others. We also used the results to examine if the patterns of spatial strategy exploration differed across user interface conditions, and found that certain interface designs did seem to promote earlier discovery of spatial strategies.

This work adds to the body of EDM papers that tackle open-ended problem spaces with both multiple solutions and multiple solution paths (Andersen, Yun-En Liu, Apter, Boucher-Genesse, & Popovic´, 2010; Blikstein, 2011; DiCerbo & Kidwai, 2013; Eagle & Barnes, 2014; Harpstead, et al., 2013; Johnson, Eagle, & Barnes, 2013; Lee, Yun-En, & Popovic, 2014; Liu, et al., 2013;

Muller, Kretzschmar, & Greiff, 2013; Siswono, 2008, Smith, Wiebe, Mott, & Lester, 2014), and is the among the very few to use educational data mining to approach the challenge of engaging learners in spatial reasoning (Wiederrecht & Ulinski, 2012; Fournier-Viger P. , Nkambou, Nguifo, Mayers, & Faghihi, 2013). It demonstrates the potential for educational data mining approaches to help educators develop models of effective solution strategies in rich but under-explored problem spaces. We also demonstrate that such "discovered" models can be used to evaluate and compare different learning environment designs. Additionally, these results suggest that we may be able to examine the meta-strategies (i.e., the ordering of strategy exploration) to determine which patterns of exploration may be more or less effective in this complex spatial problem space. This is a finding that could be used in future work to develop dynamic formative feedback to help learners engage with complex spatial problem spaces.

2. BACKGROUND

2.1. *ECOCOLLAGE*: A PLATFORM FOR PROMOTING AND STUDYING SPATIAL REASONING AROUND URBAN PLANNING PROBLEMS

According to the Next Generation Science Standards (NGSS), learners should engage with simulated models both to deepen their content knowledge of systems within the content domain and to acquire and practice skills (Schweingruber, Keller, & Quinn, 2012). In environmental science, which includes the disciplines of ecology and urban planning, system functions are dependent on the relative spatial positions of elements (e.g., buildings, permeable surfaces, habitats) (Minor & Urban, 2008). The planning challenge of installing "green infrastructure" (e.g., garden swales, green roofs, permeable pavement) to capture storm water involves this kind of sensitivity to spatial patterns. For example, a garden swale may be more or less effective at capturing storm water depending on how close it is to other green infrastructure elements, or to "grey" infrastructure elements like sewers, as each element affects the path and depth of the stormwater as it moves across the landscape. The *EcoCollage* game we created was adapted from a storm water simulation created for the Illinois Environmental Protection Agency (EPA). It is intended to be a multi-player experience, where learners jointly make decisions about where to place green infrastructure elements in an urban landscape. More details on the game can be found in Section 3.2.

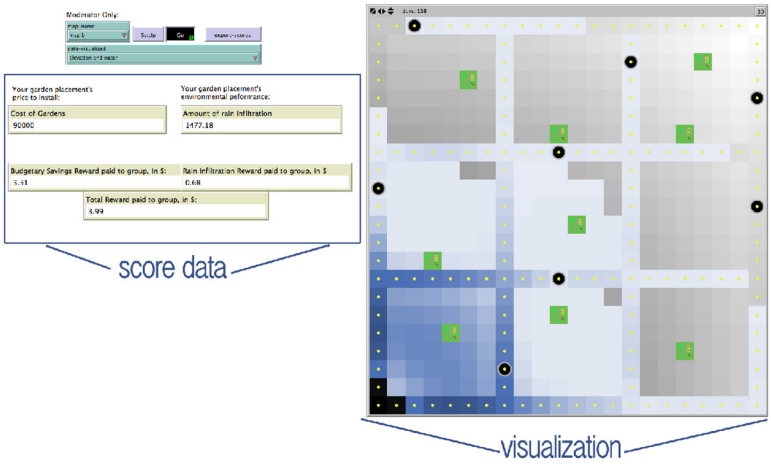

Figure 1: The simulation display: The left region provides numeric scores, and in the visualization on the right the green squares represent swales. The blue coloration shows the depth of water at the moment when this screen shot was taken, where darker blues represent deeper water. The black circles represent sewers, which also capture water, and the channels with small yellow dots are streets (where players cannot place swales).

The *EcoCollage* simulation's output (Figure 1) allows learners to see the effect of spatial patterns on storm water capture, but we suspected that the simulation's input could also play a role in meaningfully engaging leaners in the problem of discovering good patterns. Theories of embodied reasoning claim that abstract visual and spatial concepts are acquired from embodied sensorimotor experiences (Lakoff & Johnson, 1980), which led us to design a Tangible User Interface (TUI) for our simulation. Moreover, cognitive psychology experiments have shown that humans' perception of spatial patterns incorporates stereoscopic depth information (Rock & Brosgole, 1964), which suggests that a 3D tangible interface would not impede (and may even assist with) users' 2D pattern recognition. Thus, if TUIs better align with how humans innately perceive and reason about spatial relationships, then in theory a TUI would allow novice users' spatial problem solving to be less cognitively taxing.

Surprisingly, we could find little experimental work that verified the theorized benefits of TUIs. Our initial pilots demonstrated that our TUI was more efficient than the standard programming interface (Shelley, Lyons, Shi, Minor, & Zellner, 2010) but that the TUI was not very different from a multi-mouse interface in terms of usability and collaboration support of dyads (Shelley, Lyons, Minor, & Zellner, 2011). However, in these pilots we had no good way to examine the spatial reasoning of participants. We were interested in investigating how the interface design impacted users' choices during spatial problem solving, both in terms of their exploration of the problem space (i.e. breadth of problem solving) and in terms of their optimization of proposed solutions (i.e. depth of problem solving). The lack of a nuanced way to track learner's spatial manipulations motivated us to conduct the work presented in this paper. The user data analyzed here was collected from a within-subject with-rotation experiment where we had triads of users solve three equivalent but non-identical problems across three conditions: a paper-based TUI, a multiple mouse interface, and a single mouse interface (see Figure 2). See Section 3 for more information on the experimental setup.

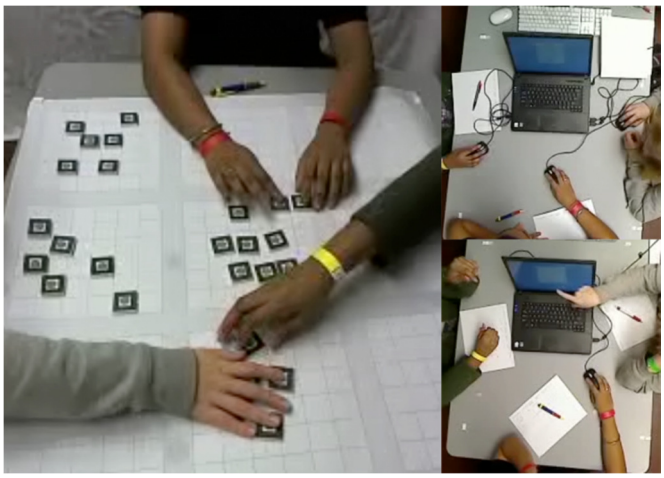

Figure 2: The three experimental conditions: the Paper TUI (left), Multi-mouse (top right), Single-mouse (bottom right). In the TUI condition, a webcam and computer vision software detected where users placed swale tiles on a paper map. In the two mouse conditions, a custom interface allowed players to drag-and-drop swales onto a digital map. In all conditions, players saw the same simulation output screen (Figure 1) when testing their arrangements.

2.2. EDUCATIONAL DATA MINING FOR OPEN-ENDED PROBLEMS

Compared to other problems, open ended problems lack structure: they are not well-bounded (the distance between the current state and the goal is difficult to determine), and the problem's solution set may comprise large number of candidates which makes it impossible to practically enumerate or individually evaluate each of the possible solutions (Biswas, Loretz, & Segedy, 2013). This property prevents us from using methodologies like Bayesian knowledge tracing and Markov models, which when applied to better-defined problems spaces give us detailed insight into learners' behavior, as done by (Jarusek, Klusacek, & Pelanek, 2013), and (Falakmasir, Pardos, Gordon, & Brusilovsky, 2013).

However, in recent years more researchers have begun to explore open-ended problems in EDM (Amershi & Conati, 2009; Berland, Baker, & Blikstein, 2014; Blikstein, 2011; DiCerbo & Kidwai, 2013; Eagle & Barnes, 2014; Gobert, Sao Pedro, Raziuddin, & Baker, 2013; Harpstead, et al., 2013; Johnson, Eagle, & Barnes, 2013; Lee, Yun-En, & Popovic, 2014; Liu, et al., 2013, Lynch, Ashley, Pinkwart, & Aleven, 2008; Smith, Wiebe, Mott, & Lester, 2014). Often the first task that confronts the researchers is the need to select a way to represent the solution space such that the analysis is tractable.

For example, (Kardan & Conati, 2013) focused on providing adaptive support based on user interaction patterns with a simulation built to teach the students to solve constraint satisfaction problems. In (Blikstein, 2011), learning analytics is used to assess students' behavior in open-ended programming tasks. Snapshots of code during the assignment were used to extract student behavior and categorize them in terms of programming experience. In related work, Berland and colleagues used manually-derived feature sets to describe novice programmers' code, and then used those descriptions to assemble probabilistic state transition diagrams describing how novice programmers would proceed when assembling their programs (Berland, Martin, Benton,

Smith, & Davis, 2013). In these scenarios, the problem space is similar to ours considering that the learners could solve the problems using different solution paths, however these differed in the sense that the problem spaces had a concretely-structured solution or end goal to which the learners solutions could be compared, which our problem space lacked (while we have reward functions, there are many possible structures that could be used to attain similar rewards).

One open-ended problem EDM approach tackled the problem of multiple possible solutions by using a combination of data mining and automaton theory to extract features of the learners' solutions and compare them to designers' solutions (Harpstead, MacLellan, Koedinger, Aleven, Dow, & Myers, 2013). The context was a game called *RumbleBlocks* where children 5- 8 years of age had to place blocks to build a stable structure. Their understanding of center of mass and stability was being assessed through the game. For each student solution a decision tree was constructed; once they had these trees researchers tried to extract features to create a vector that would describe how and what students were doing. The features of those structures were then matched to the features from the solutions produced by the designer of the game. The study focuses on helping the designers and researchers redesign aspects of the learning experience that seemed to produce discrepancies between how the players used it and how the designers had envisioned its use. This approach works very well when (1) there are "expert solution paths" available for comparison, and (2) there is "path dependence" in user actions – meaning that later user actions are constrained by earlier actions, thus reducing the size of the problem space. For many open-ended problems, like our ecology simulation, these two conditions may not hold.

2.3. Prior Work using Technology to Promote Spatial Reasoning

Hegarty and colleagues define spatial reasoning as: "Ability which is concerned with individual differences in how people mentally represent and manipulate spatial information to perform cognitive tasks" (Hegarty & Waller, 2005). We first reviewed factor analytic studies of spatial abilities. This research tradition provided strong evidence that spatial ability is differentiated from general intelligence and that it is not a single, undifferentiated construct, but instead is composed of several somewhat separate abilities.

"Spatial reasoning" is a term that covers a wide variety of mental tasks. An entire branch of cognitive psychology, visuospatial cognition, has been devoted to studying and measuring spatial reasoning. The term "visuospatial" is used in this domain because it clarifies the fact that the information being reasoned about is "visual in nature (initiated by stimulation of the retina by light) and has spatial properties (involving the representation of space including relationships between objects within that space)" (Halpern & Collear, 2005). Those researching visuospatial cognition have long acknowledged that there may be different "types" of visuospatial cognition, or factors, which individuals may be better or worse at performing. Many of these factors are strongly associated with the tests used to measure them - for example, "spatial orientation" involves imagining how one's perspective might shift the appearance of a visual array, while "spatial relations" involves demonstrating that one understands how the parts of a 3D object relate to one another by correctly identifying it in rotated views. What is often lost in discussion is the fact that the field of visuospatial cognition arose from the early 20th-century need to design tests to identify people with mechanical skills (Hegarty & Waller, 2005). Thus, there has always been a strong focus on the ability of people to comprehend the structure of 3D objects. Later on, another common research focus was on navigation, i.e., how people could move through 3-dimensional spaces. Thus, two-dimensional pattern recognition, which is what our work requires

of learners, has no validated visuospatial metrics. Existing visuospatial tests like the Factor-Referenced Tests described in (Ekstrom, French, & Harmon, 1976) often ignore reasoning about allocentric representations, as one performs with 2D patterns. There are two main categories of spatial knowledge representations: allocentric spatial reasoning represents the ability to reason about spatial arrangements between objects independent of a first-person perspective, and egocentric spatial arrangements describe the position of the object from a persons' perspective (Fournier-Viger P. , Nkambou, Nguifo, Mayers, & Faghihi, 2013). Tests often ask learners to infer egocentric spatial information from allocentric spatial representations, or vice versa, but do not test them on allocentric reasoning alone.

Egocentric spatial reasoning has been identified as a very important especially in the field of engineering where the learners should be able to visualize 3D objects and their 2D projections through rotation tasks. For this reason, much of the prior work on supporting spatial reasoning via technology had focused on improving learners' performance on more traditional egocentric, 3-dimensional tasks, like object rotation or navigation. There have been several attempts at creating tutors which could help the learners improve their visuospatial skills by solving problems like the missing views problem (Wang & Kim, 2005; Connell & Stevens, 2002; Bravo, Hernandez, Saorin, & Contero, 2010; Nesbitt, Sutton, Wilson, & Hookham, 2009; Mengshoel, Chauhan, & Yong, 1996; Hubbard, Mengshoel, Moon, & Yong, 1996). Egocentric skills have also been explored from the perspective of improving navigation skills among learners through computer games (Bravo, Hernandez, Saorin, & Contero, 2010).

The ability of people to reason about two-dimensional allocentric spatial phenomena, like the relationships between objects in 2D spatial patterns, has received relatively less attention (Fournier-Viger P. , Nkambou, Nguifo, Mayers, & Faghihi, 2013) than these highly-egocentric 3-dimensional reasoning challenges. However, our problem domain deals with complex 2D allocentric spatial reasoning, which is more concerned about the relationship between the elements than the elements in themselves. Our study involved closely observing how the allocentric representations affected the other aspects like the infiltration, as we tried to understand the learners' perspective of the arrangements and the changes in the arrangements.

2.4. METRICS FOR SPATIAL PHENOMENA

Spatial reasoning, which is a catch-all term for the ability to mentally visualize and manipulate two- and three-dimensional objects, is a known predictor of success in Science, Technology, Engineering and Math (STEM) fields (Wai, Lubinski, & Benbow, 2009). However, if we want to go beyond just noting this correlation to improving the success learners have in STEM, more must be known about how to support and measure spatial learning in real time. Though research has explored how different training (Uttal, et al., 2013) and pedagogical strategies in classrooms (Stieff, Dixon, Kumi, & Hegarty, 2014) can improve spatial skills on post-tests of spatial abilities, there is a lack of methods for studying how spatial reasoning evidences itself *during* the learning process, which is what is needed to provide formative feedback to learners. Our work contributes to this endeavor, although we only focus on one area of spatial reasoning: two-dimensional spatial patterns.

2.4.1. Two-Dimensional Spatial Pattern Characterization Methods

Before we can hope to study how people reason about spatial problems, we need some way of measuring the spatial properties of their proposed solutions. The literature on statistics to

measure spatial patterns is extensive, and is most often found fields like plant ecology, animal ecology, geography, mining, and engineering. These fields make use of spatial statistics for either explorative or inference purposes, and thus employ approaches like counting methods, covariance, variance, etc. (Fortin, Dale, & Hoef, 2002). The selection of the spatial statistic is influenced by research objective, measurement types and sample data (Fortin, Dale, & Hoef, 2002). For this problem space, we are concerned with the relative placement of items: are they near one another, or spread apart? The relative distances between swales and other swales, and between swales and other water-capturing elements (like sewers) meaningfully affects the patterns of flooding that emerge in urban settings.

The spatial metric we settled on for this work is the Ripley's *K* metric. A refinement of the nearest-*K* neighbors metric, it calculates spatial metrics on varying scales of distances (Dixon, 1995). The Ripley's *K* function quantifies the density of points for various sizes of circular windows. The Ripley's *K* metric can thus successfully detect combinations of effects like clustering at large scales while simultaneously being sensitive to regularity at smaller scales. It can also compute these properties for both univariate (where the arrangement of an item is observed with respect the other items of the same type) and multivariate (the arrangement of an item with respect to one or more other item types) patterns. By using a circular window Ripley's *K* gives an isotropic (i.e., non-direction-sensitive) measure of point density (Fortin, Dale, & Hoef, 2002). We used the normalized Ripley's *K* metric (the Ripley's *L* measure) to convert our problem's state space from one of absolute Cartesian placements of swales to the much smaller (and more meaningful) state space of the relative spatial arrangements of swales.

2.4.2. Details on the Ripley's *K(t)* function

The Ripley's function *K(t)* is defined to give the probability of finding the elements of interest in the specified window size given the overall density of elements in that area. The general definition of the Ripley's *K*-function for a certain distance t is

$$\boldsymbol{K(t) = \frac{1}{\lambda[E(t)]}}$$

Equation 1: Theoretical *K(t)* function

Where:

- λ is the density of the study plot, measured as $\frac{n}{A}$, where *n* is the number of points in the study plot, and *A* is the area of the study plot
- *E(t)* is the expected number of points within distance *t* of an arbitrary point

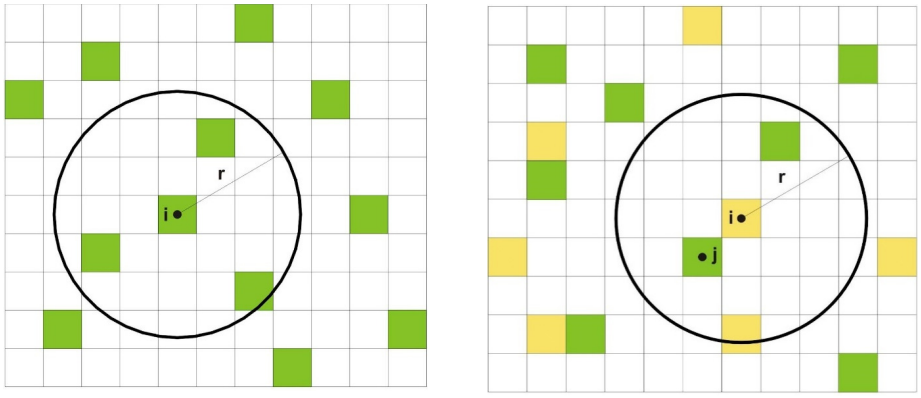

Figure 3. Illustration showing Ripley's univariate calculation (left) and bivariate calculation (right).

Figure 3 shows an illustration of how Ripley's univariate and bivariate calculation works for the radius *r*. The univariate calculation, sweeps a region of size *r* around each item of interest, *i*, counting the number of other items of that type within the region. The approach repeats this calculation for radii of size 1 to *t*, producing t different *K* statistics. Whereas the Ripley's *K* bivariate sweeps the study space while tallying the number of items of a second type *j* found within radius r of each item *i* of the first type. For our analysis, we wanted to track how learners placed swales in relation to other swales (a univariate calculation), and how they placed swales in relation to existing sewers (a bivariate calculation), which we will now describe.

For univariate computations:

$$K(t) = \frac{1}{\lambda}\sum_{i=1}^{n}\sum_{j=1}^{m} w(i,j)I(i,j)$$

Equation 2: Ripley's *K* Univariate

Where:

$i \neq j$

λ is the density of the study plot, measured as $\frac{n}{A}$, where n is the number of points in the study plot, and A is the area of the study plot

$w(i,j)$ is the edge correction factor, which is 1 if the search circle centered at i and passing through j is completely inside the study area, otherwise it is the proportion of the search circle in the study area

$I(i,j)$ is the indicator function, which is 1 if point j is within distance t of point i, 0 otherwise

For bivariate computations:

$$K(t) = \frac{1}{\lambda_1 * \lambda_2} \sum_{i=1}^{n} \sum_{j=1}^{m} w(i,j) I(i,j)$$

Equation 3: Ripley's *K* for Bivariate

Where:

$i \neq j$

λ_1 is the density of elements of type 1 within the study plot

λ_2 is the density of elements of type 2 within the study plot

m is the number of elements of type 1 within the study plot

n is the number of elements of type 2 within the study plot

Because of its hyperbolic behavior, the interpretation of *K*-function is not straightforward, especially if one wishes to compare the spatial characteristics of one map against another. For this reason, a modification called *L*-function has been proposed to normalize it:

$$L(t) = \sqrt{K(t)/\Pi} - t$$

Equation 4: Ripley's *L* Equation

The expected value of the univariate L-function under CSR (complete spatial randomness) is 0 for all *t*. Complete spatial randomness (CSR) describes a point process whereby point events occur within a given study area in a completely random fashion. Such a process is modeled using only one parameter λ, i.e. the density of points within the defined area (Maimon & Rokach, 2010). Poisson distribution is used to express the probability of given number of events occurring in a fixed interval of space and/or time independently of the last event. Thus, when the *L* value is positive, indicating that the pattern is more tightly-packed than one would expect to see by chance, we know that the pattern tends to be clustered, and when the *L*-value is negative the pattern is tending towards being overdispersed or regular (Dixon, 1995).

The accuracy of the *K* value highly depends on the size and shape of the study area and the edge effects, which need be considered when the search circle intersects the edge of the study plot. The edge effect, if uncorrected, would overestimate how much "empty space" surrounds points of interest especially at the boundaries of the study plot as compared to those in the center of the study plot. As shown in Figure 4 the search circle consists of two distinctive parts: one inside the study plot, $A(r)^+$, and another outside the study plot, $A(r)^-$. If *A(r)* includes the portion of the search circle denoted by $A(r)^-$ the area would have fewer points than expected (Protázio, Pereira, & Elayne Jesus de Castro, 1999). Edge correction calculates the proportion of the search circle inside the study plot, $A(r)^+$, and utilizes the area of this proportion in the calculations. The mechanism becomes complicated when the shape of the study plot is irregular, but in our case, we are using rectangular maps and this method suffices. Common practice while considering edge corrections is to cap the maximum search circle radius to be about one half of the shortest dimension of the study area (Protázio, Pereira, & Elayne Jesus de Castro, 1999), as this reduces the number of assumptions being made about the pattern, and so we capped our maximum radius to be 11.

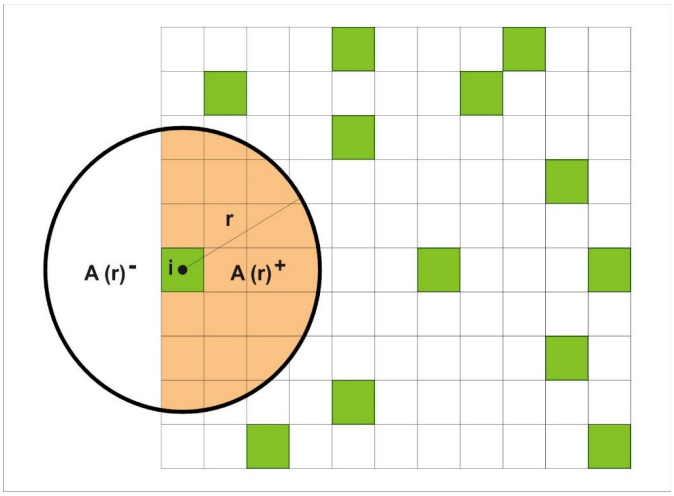

Figure 4. Study area and a search circle.

3. METHODS

3.1. EXPERIMENTAL PARTICIPANTS

The target demographic for *EcoCollage* was high school and college undergraduate students. In total we had 90 participants, divided into 30 triads. There were 22 triads of college students and 8 triads of high school students. (Post-hoc analysis found no significant differences in the performance of the high school students and college students, and so all data was considered together in the remaining analyses). The college students were recruited through fliers, classroom visits, and email lists. The high school students were recruited through a summer science program at our institution. We wanted to have a mix of student groups who already knew each other and groups who did not know each other, to reflect the varying familiarity students would have with one another in real classroom settings. This is important because familiarity can affect how small groups collaborate. The 8 triads of high school students all knew one another. To recruit college students who knew one another, we asked each respondent if they wanted to suggest a friend or classmate who could join them in the activity. We had 10 collegiate triads where the participants knew one another to some degree, and 12 collegiate triads where all three participants were strangers. A discussion of the potential interaction of familiarity and interface will be left for a future paper focusing on collaboration.

3.2. EXPERIMENTAL MATERIALS

The version of the *EcoCollage* game used in this experiment allowed players to choose where on a 22x22 grid representing an urban area (see Figure 1) they wished to install green infrastructure. For the experiment here, we limited users to a single type of green infrastructure element: swales (gardens that capture stormwater). "Installation" of the swales in the interfaces involved taking a swale symbol (either a digital icon or a tangible fiducial symbol) and placing it on the map (either a digital map or a physical paper map). The swales could be placed on any non-street grid square, resulting in 18x18 or 324 allowable spaces to install swales on the map. As in real life, the streets were of slightly lower elevation than nearby spaces on the map, and contain sewers, which can also capture storm water. The grid also included a "sink" – the point

of lowest elevation on the map, which in our experiments was always on the map edge or corner. Thus, if players do not capture stormwater with swales, it either runs into the sewers (which is undesirable, as in many cities it overwhelms water-treatment facilities), or "off the map" into the sink (which is also undesirable, as it represents stormwater that flows into adjoining neighborhoods). To avoid learners just reusing solutions from one condition to the next, we designed three maps which had similar characteristics, but different Cartesian locations of the sewers and the sink. In the paper TUI condition, the maps were printed on paper and the swales were cardboard tokens, while in the multi-mouse and single-mouse conditions, the maps were presented via a custom graphical user interface, where users could drag and drop swales.

Participants could test their swale arrangement at any time by clicking a "test" button (located onscreen in the multi-mouse and single-mouse interface conditions, and on a nearby desktop in the paper condition). The test button would trigger the generation of a text file containing the Cartesian coordinates of the swales on the map. The swale placement text file was read into the simulation (created using NetLogo), and then the storm water simulation was run. Participants could see the effect of their placement in two ways: via score data (in the experiment they received three scores: one based on the monetary cost to install the swales, one based on the amount of storm water captured by the swales, and an aggregate score that combined the cost and capture scores) and via an animated visualization that depicts the flow and depth of water on the landscape in the hours after the rainstorm (see Figure 1). See Section 3.4 for more details on the scores.

3.3. EXPERIMENTAL DESIGN

This study used a 3 x 3 (interface x map design) within-subject with-rotation design. The independent variables were interface (paper TUI, multi-mouse, single mouse; see Figure 2) and map design (map a, map b, map c). These conditions were both rotated to counterbalance any learning effects or inherent differences in the maps. Thus, each triad experienced each of the three interface conditions, and each of the 3 different map designs, but the order of exposure and interface-map pairings rotated from triad to triad – one triad might experience paper/map a, multi-mouse/map b, single mouse/map c, while another triad might experience multi-mouse/map, c paper/map b, single mouse/map a. The dependent variables were the trials produced by the triads (a trial is the arrangement of swales tested in the simulation by the triad, obtained from the Cartesian coordinate text files used by the simulation), and the component scores associated with those trials (more details on scoring is in Section 3.4). Videos, interviews, surveys and post-test data were also collected but were not used in the analysis presented in this paper.

3.4. EXPERIMENTAL PROCEDURE

Participants were asked to place tokens representing swales within a gridded map of an urban landscape. They were free to use as many or as few swales as they chose. The effectiveness of the swales is dependent on their proximity to the sewers, the elevation gradient, and the arrangement of other swales. Users must be sensitive to these spatial patterns to configure efficient swale arrangements.

Their challenge was to balance two competing objectives: to maximize the amount of groundwater infiltration by capturing more stormwater with swales (as opposed to allowing it to run into the sewers), and to minimize the cost of added infrastructure. To motivate participant

performance, we offered a financial incentive that was additively computed from the two component scores: infiltration and cost. The infiltration score ranged from $0 – $3.50, with 0 corresponding to no infiltration (no stormwater is captured) and $3.50 corresponding to the maximum amount of infiltration (all 324 possible lots converted to swales). The cost score also ranged from $0 – $3.50, with 0 representing the maximum cost (all 324 lots converted to swales) and $3.50 representing the minimum cost (no swales). Because the component scores are in contradiction (adding swales improves infiltration but increases infrastructure cost), participants were incentivized to maximize the efficiency of each swale to maximize the summed payout. Participants received payments for each of the three conditions, corresponding to their triad's top-scoring trial within that condition.

Participants had twelve minutes in each condition to use the condition's interface to try to complete this challenge. Each time the participants tested an arrangement using the simulation, we dubbed the arrangement a "trial". The participants were informed that they were allowed to run as many trials as they wished within those twelve minutes, with their payment dependent on the best score attained, so they would feel free to explore how their placements modified their simulation outcomes. Across all the experimental conditions, the average number of trials produced within each of the 12-minute segments was 20.

3.5. ANALYTIC PROCEDURE

Here we describe how we approached reducing the state space of the triads' trials to the point where we could begin to meaningfully analyze their exploration of 2D spatial patterns. We needed to explore ways to "bin" solutions into classes or categories to even begin to attempt to apply educational data mining techniques to this problem space. Whenever educational data mining researchers confront the question of state space reduction, they must decide if they will follow a top-down approach informed by knowledge of the learning domain, or if they will follow a bottom-up approach, using the learner performance data itself to bootstrap a reduced state space. An initial, unsuccessful top-down approach is detailed in Section 3.5.1, and our eventual bottom-up solution is covered in Section 3.5.2.

3.5.1. Initial Top-Down State Space Reduction Approach: Condensing State Space to "Change" Space

Our first state space reduction involved using our understanding of the problem space to design spatial metrics to reflect micro-genetic changes learners could make to the maps (Lyons, Dasgupta, Shelley, Slattery, Minor, & Zellner, 2012). We designed 4 different metrics based on what we thought would constitute features of learner exploration. These metrics were designed to track trial-to-trial spatial changes in placements, so we could understand how learners were exploring the problem space. These metrics were not intended to represent strategies, but rather were selected on the basis of their ability to reveal learners' exploration patterns: for example, slow-and-steady changes (analogous to "hill climbing") might be more effective at producing good outcomes than sharp changes (analogous to "random restarting"). We thus chose two metrics that could indicate how similar or dissimilar each solution seemed to one another, in terms of the placement "actions" taken by the users to produce the solutions:

Placement Dissimilarity (*PD*): was designed to note the changes in the placements of swales. We designed this metric using the Hamming distance metric from information theory (i.e., the number of edits needed to make two identical-length strings match). Assuming the map divided

into grid of $l * b$ blocks (l being the *length* of the map and b being the *breadth*), the "string" would be comprised of $l * b$ binary numbers, where 1 indicates the presence of a swale, and a 0 indicates an absence of a swale. The PD metric counts the number of points on the map that do not match if the maps are aligned with each other, normalized by the number of swales in both maps.

Mathematically, *PD* was denoted as:

$$PD = \frac{\sum_{i=0}^{l*b} |M_a^i - M_b^i|}{(N_a + N_{b)}}$$

Equation 5: Placement Dissimilarity

Where:

M_a^i is 1 if a swale is present at location i in Map a, and 0 if no swale is present at location i

M_b^i is 1 if a swale is present at location i in Map b, and 0 if no swale is present at location i

N_a is the number of swales in Map a

N_b is the number of swales in Map b

Abundance Dissimilarity (*AD*): was designed to track changes in the number of swales the learners used. We normalized this metric by the maximum number of swales of the two maps that were being considered. *AD* was calculated as:

$$AD = \frac{|N_a - N_b|}{MAX(N_a, N_b)}$$

Equation 6: Abundance Dissimilarity

We also wanted to be sure to capture nuance in how (perhaps even slight) changes in placements of elements relative to one another might cause large impacts to spatial patterns, and so devised two relative-placement-dependent metrics that relied on the Ripley's metrics described in Section 2.4.2:

Spatial Dispersion Dissimilarity (SDD_U, SDD_B): To compute the Spatial Dispersion Dissimilarity (*SDD*) across two maps, a and b, we first computed the univariate and bivariate Ripley's L values for the different sweep radii, r, and converted them into strings that would indicate if the map was clumped, random, or overdispersed at a given radius r. Then we computed an edit distance:

$$SDD = \frac{1}{r}\sum_{i=1}^{r} \left|L_a^i - L_b^i\right|$$

Equation 7: Spatial Dispersion Dissimilarity

Where:

r is the maxium radius of the Ripley's sweeps (11 in our case)

L_a^i is 1 if the Ripley's L value indicates statistically significant clumping at radius *i* in map *a*, 0 if it indicates random placement, and -1 if it indicates overdispersion

L_b^i is 1 if the Ripley's L value indicates statistically significant clumping at radius *i* in map *b*, 0 if it indicates random placement, and -1 if it indicates overdispersion

SDD_U computes the SDD for the univariate Ripley's L values (swale vs. swale placements)

SDD_B computes the SDD for the bivariate Ripley's L values (swale vs. sewer placements)

Though these metrics were able to tell us about the changes the learners made, it appeared our assumptions about which spatial changes would serve as a window onto spatial reasoning processes were wrong. We were not able to significantly and reliably correlate these micro-genetic exploration changes to changes in the outcome metrics like infiltration, so using these metrics as the basis for data mining was a non-starter. We needed to re-examine our approach – rather than taking a top-down approach of assuming what might constitute a spatial manipulation worth attending to and manually creating feature sets, we needed a bottom-up approach that would help us *discover* spatial strategies that actually had meaningful impact on the simulation outcomes.

3.5.2. Final Bottom-Up State Space Reduction Approach: Data Mining Ripley's Values to Reduce the State Space by Discovering "Strategies"

Our goal was to infer the spatial strategy or strategies the learners were using to improve their scores. In our second state space reduction attempt, we decided to use the Ripley's metric to construct spatial "profiles" for each solution, and use these profiles to discover which 2D spatial strategies are effective. Each "profile" initially consisted of two vectors of length 11, representing the univariate (swales placed relative to swales) and bivariate (swales placed relative to sewers) Ripley's *L* values calculated at each radii varying from 1 to 11 (half of the length of the study area).

So, for example, for the map depicted in Figure 5, the univariate vector, L_u, would be:

$$L_u = \{1.21, 10.18, 21.34, 26.32, 29.86, 33.92, 37{,}26, 42.01, 45.81, 49.49, 51.84\}$$

Figure 5. A visualization of one of the trials. The green patches represent the swales the triad placed on the map.

These L_u values are also plotted in Figure 6 with the lower and the upper bounds of the confidence intervals. In order to test the deviation from randomness (dispersion or clustering) of the point patterns using the univariate or the bivariate functions, we computed a 99% confidence interval of *L(t)* using the Monte Carlo method from 500 simulated CSR patterns with the same number of points contained inside a region with the same geometry (Dixon, 1995). The points above the confidence interval displayed clumped patterns whereas the points below the lower confidence interval displayed an overdispersed pattern (i.e., a regular pattern).

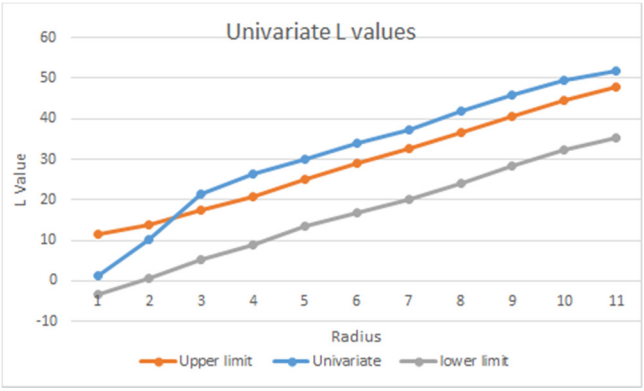

Figure 6. Plot of Univariate L values for the visualization in Figure 5. Note that for radii of size 3 and above, the swales are placed closer to one another than one would expect by chance (i.e., they are clumped)

Based on the confidence intervals we denote the radii which have clumped arrangements by 1, overdispersed arrangements by -1 and random arrangements by 0, much as we did to compute the SDD_U and SDD_B metrics described in Section 3.5.1. We call this notation the normalized *L* value notation, L^{norm}. We hypothesized that we could use these normalized values to compare the spatial patterns used across trials. For the visualization map shown in Figure 5 then, the normalized univariate L-values can be written as:

$$L_U^{norm} = [0,0,1,1,1,1,1,1,1,1,1]$$

For ease of later analysis, we split each 11-tuple into a binary tuple of length 22, L^{strat}, where the first 11 elements indicated the presence of clumping at each of the 11 radii with a 1, and the last 11 elements indicated the presence of overdispersion. The univariate L^{strat} values for the map in Figure 5 would then look like:

$$L_U^{Strat} = \{(0,0,1,1,1,1,1,1,1,1,1), (0,0,0,0,0,0,0,0,0,0,0)\}$$

Because we were computing these metrics for both univariate and bivariate distributions of swales, we ended up combining L_U^{strat} and L_B^{strat} to form 44-tuple "profile" to represent each solution. Nonetheless, our state space has a theoretical max of 22!, as a distribution can never be both clumped and overdispersed at the same radius at the same time. While large, this is still a great deal smaller than our original 324! state space.

We used multivariate stepwise regression to identify which of the spatial patterns the learners adopted helped them to improve infiltration (no matter where they place the swales some infiltration is bound to happen, although some arrangements are superior to others). In other words, which of these patterns actually had a meaningful impact on infiltration and could be considered "strategies." We felt comfortable using the experimental trial data for discovering effective spatial patterns (as opposed to a more complex but completely impractical approach where we would, say, generate all possible 324! solutions, test them in the simulation, and derive optimal spatial solutions) because we were more interested in examining the portion of the problem space actually explored by learners, not in exploring the problem space itself. Put in educational terms, we were interested in examining the strategies within the learners' "Zone of Proximal Development" (Vygotsky, 1978) – meaning, the space of strategies within reach of their current (novice-level) understanding of the problem. For these reasons, we decided to use learner-generated solutions to help us flag potentially beneficial spatial patterns. We also felt comfortable calling these discovered-to-be-effective patterns "strategies" because they are generalizable and repurposeable – recall that the Ripley's metric is isotropic, meaning that the pattern is not tied to any fixed location or map. We utilized the regression to identify the strategies which had a positive impact on the outcome metrics and which had a negative impact, as this could be determined by the signs on the parameters of the model.

4. RESULTS

We needed a mechanism to compare the strategies across interface designs transparently. For this purpose, we regressed all the trials, irrespective of the interface that they were attempted in. The regression model would give us the significant variables that are found to be more generally effective at producing positive infiltration, and we could then use these parameters to compare across the conditions. We regressed both the univariate and bivariate spatial metrics against the infiltration measure. The coefficients we got from the regression expressed how the clumped and overdispersed strategies at a particular radii compared against instances where the swales were effectively placed randomly (see Table 1).

Table 1. The coefficients of the significant variables for all the trials

	Significant Variables	Estimate	tStat	pValue
1	clumped Univariate at radius 1	1256.3	4.54	6.65E-06
2	overdispersed Univariate at radius 1	2866.2	8.88	8.33E-18
3	overdispersed Univariate at radius 2	977.01	2.60	9.62E-03
4	clumped Univariate at radius 3	1050.5	3.29	1.07E-03
5	clumped Univariate at radius 8	1332.5	3.12	1.87E-03
6	overdispersion Univariate at radius 10	-1921.2	-3.19	1.50E-03
7	clumped Univariate at radius 11	1211	3.00	2.77E-03
8	overdispersed Bivariate at radius 1	2635.2	8.22	1.28E-15
9	clumped Bivariate at radius 2	2809.5	6.28	6.49E-10
10	overdispersed Bivariate at radius 3	541.62	1.67	9.53E-02
11	clumped Bivariate at radius 4	1119.7	1.89	5.92E-02
12	overdispersed Bivariate at radius 5	726.19	2.21	2.76E-02
13	clumped Bivariate at radius 6	1691.7	2.95	3.31E-03
14	overdispersed Bivariate at radius 6	1565.1	4.50	8.36E-06
15	clumped Bivariate at radius 7	3488.8	8.86	9.66E-18
16	clumped Bivariate at radius 9	1002.1	2.22	2.68E-02
17	clumped Bivariate at radius 10	-1390.3	-3.22	1.37E-03
18	overdispersed Bivariate at radius 10	891.09	2.96	3.24E-03

The coefficients can be interpreted as comparisons against a random placement at a given radius; the random placement may be regarded as an absence of a spatially-sensitive strategy at that level of scale. We identified the strategies with positive coefficients as "good strategies" and strategies with negative coefficients as "bad strategies," because the positive coefficients terms add up to give a better infiltration whereas the negative coefficient terms reduce infiltration with respect to random arrangement. For example, the clumped arrangement for univariate spatial metrics at radius 1 has a positive coefficient (1256.3) and would be termed as "good strategy," whereas the overdispersed arrangement for univariate spatial metrics at radius 10 has a negative coefficient (-1390.0), and would be affecting the infiltration negatively so would be termed as "bad strategy". We observed that if the participants employed more good strategies they would get better outcomes metrics than when they would employ more of the bad strategies. Figure 7 illustrates two bivariate "good" strategies in use at the same time.

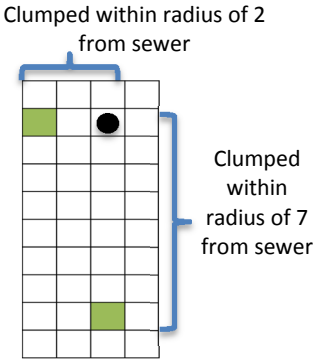

Figure 7. Example of two "good" bivariate strategies.

4.1. Comparing and Tracking Learner's Use of Spatial Strategies across User Interfaces

The model obtained from the regression included 18 (of 44 possible) different spatial strategies as significant contributors to infiltration outcomes, and had a coefficient of determination (R^2) value to be 0.802 ($p < 0.05$). This model is a fairly good fit for the data at hand. Table 2 is an alternate representation of the coefficients in Table 1. The coefficients of the significant variables for all the trials above highlighting their magnitude and polarity. False coloring indicates the degree of positive (green shading to orange) to negative (red) impact each strategy had on infiltration, and the numbers on the left indicate the radius of that arrangement. In total, 16 "good" and 2 "bad" strategies were found.

Table 2. An alternate representation of the coefficients in Table 1 above

	Overdispersed Univariate	**Overdispersed Bivariate**	**Clumped Univariate**	**Clumped Bivariate**
1	2866.2	2635.2	1256.3	
2	977.01			2809.5
3		541.62	1050.5	
4				1119.7
5		726.19		
6		1565.1		1691.7
7				3488.8
8			1332.5	
9				1002.1
10	-1921.2	891.09		-1390.3
11			1211	

We used this model to examine the learners' exploration of the problem space. This model allowed us to ask if the different interface designs might affect things, like:

- Are some of the interfaces more likely to encourage the use of good strategies than others? (As measured by the total number of good spatial strategies used in the different conditions),
- Did the interfaces affect the participants went about exploring the space of strategies identified as good strategies? (As measured by trial-to-trial changes in the applications of good strategies)
- Did the interfaces seem to influence whether participants were more likely to *explore* good strategies or to *exploit* good strategies? (As measured by the number of good strategies participants discovered within a condition)
- Did the interfaces affect how long it took participants to discover good strategies? (As measured by the number of iterations it took before a strategy was identified/employed)

Let's examine each of these in turn.

4.1.1. Does the interface type affect the total number of effective spatial strategies used?

We compared the total number of good strategies across the interface trials, and it seemed that, in the multi-mouse condition, participants employed slightly more total good strategies, followed by the paper condition, with the single-mouse condition showing the smallest total number of good strategies used, although we found that none of these differences were significant (see Table 3). This indicates that none of the interfaces predisposed learners to employ significantly higher numbers of shown-to-be-effective spatial strategies (in terms of infiltration), which might be expected, since the spatial problems were effectively the same in all three conditions.

Table 3. The total number of good strategies, by condition

	Paper	Multi-mouse	Single mouse
Total good strategies	572	676	530
Average per group	19.07	22.53	17.67
(STDEV)	(19.63)	(20.58)	(19.16)
Average per Trial	**2.95**	**3.25**	**2.40**
(STDEV)	**(2.63)**	**(2.89)**	**(2.35)**

4.1.2. Does the interface type affect how learners explored the space of effective spatial strategies?

We compared the average of trial-to-trial changes (deltas) in the specific good strategies used (Table 4). We found that in the paper condition, participants showed more change in the specific "good" strategies they employed from one trial to the next, followed by the multi-mouse condition, and with the single-mouse condition showing the smallest amount of change in the strategies used. This difference was significant according to a within-subject ANOVA ($F=4.43$, $p=0.0162$). A post-hoc Bonferroni-Holm correction revealed that the only significant pairwise difference was between the paper and the single mouse conditions. A higher trial-to-trial delta in good strategies employed indicates that a given trial is less similar to the trial that preceded it. Because we are only tracking the change in the application of strategies known to positively impact the outcome, the presence of a higher delta indicates that the participants are more active in exploring the space of good solutions, an activity that is more likely to yield meaningful outcomes. It can be seen as a marker of productive exploration of the *strategy* space (which is different from exploring the *problem* space). The problem space contains all 324! possible solutions, whether they are effective or not: with such an open-ended problem space, wide exploration can all too easily be non-productive. The strategy space is the smaller 16! space of good solutions, where exploration is more likely to be productive for optimizing infiltration. Thus, participants explored the strategy space significantly more effectively in the paper condition than in the single mouse condition. The fact that the strategy exploration of the participants was middling for the multi-mouse condition suggests that distributed control, regardless of whether it is accomplished with a TUI or with mice, also seems to promote more strategy space exploration.

Table 4. The average delta in number of good strategies used, trial-to-trial

	paper	multi-mouse	single mouse
Average Δ good strategies	1.52	1.17	0.75
(STDEV)	(1.56)	(1.10)	(0.68)

4.1.3. Does the interface type affect if learners were more likely to *explore* good strategies or to *exploit* good strategies?

In Artificial Intelligence, a common way of categorizing how intelligent agents respond to problem spaces is to determine whether they are likely to "explore" (where the agent makes large changes in the solution approach, favoring the discovery of multiple, wildly different types of solutions but risking discovering nothing but bad solutions) or to "exploit" (where the agent iterates on a decently good solution to try to maximize the outcome, favoring a good outcome but risking missing out on a much better solution by becoming trapped in a "local maximum" of the problem space). This model of exploration-versus-exploitation often gets applied either implicitly or explicitly to learners confronting rich problem spaces, for example, in the "oscillating" versus "inching" explorations of simulation users in (Levy & Wilensky, 2007). While each approach can be productive for learning, in this context it is arguably better to explore the space of good strategies, as learners are intended to use the simulation to build an

understanding of how different spatial patterns can affect stormwater infiltration. Thus, we used the count of unique strategies – i.e., how many of the 16 known "good" strategies a group attempted within a given condition – as a proxy for exploration. Groups which discovered a larger array of good strategies were considered explorative and the groups which discovered relatively fewer types of good strategies were exploitative. For example, groups sometimes made very minute changes in the patterns, essentially using the same strategy set as they had employed in the previous trial and thus exploiting a known solution, as shown in Figure 8.

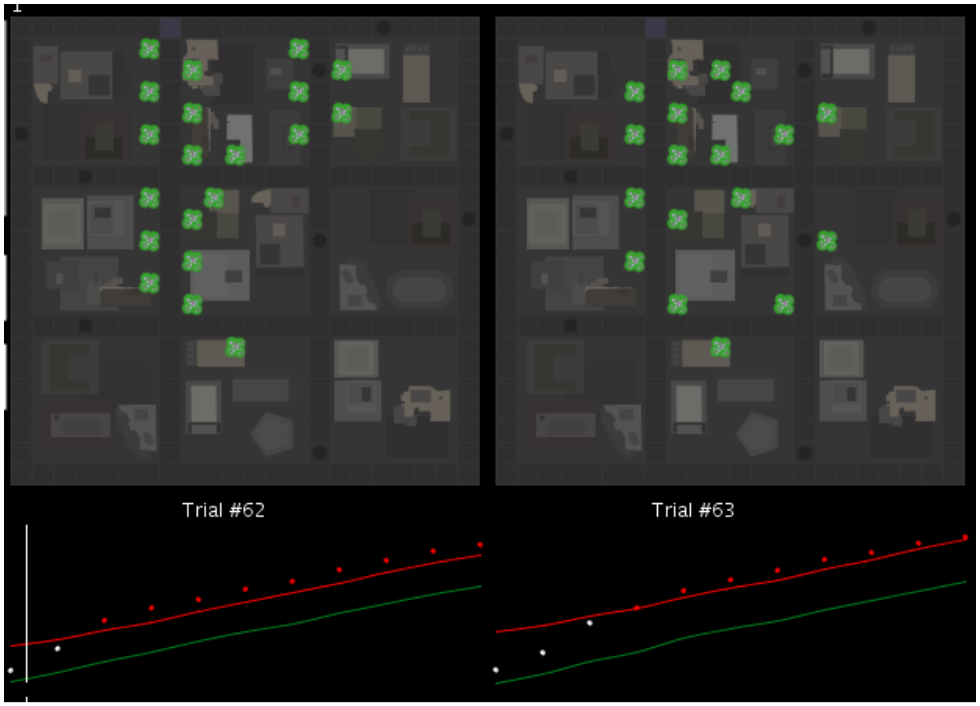

Figure 8. Consecutive trials of group 1 exhibiting exploitation (relying on the same set of strategies in generating solutions). The plot below the visual maps indicates the univariate L values and confidence levels.

The interpretation that the interface condition seems to affect strategy exploration seems to be supported when we examine the number of unique strategies discovered (Table 5). In the multi-mouse condition, when summed across all of their trials participants discovered more unique good strategies (5.83). This is followed by the paper condition (5.30) with the single-mouse condition showing the smallest total number of unique strategies used (4.03). This difference was significant according to a within-subject ANOVA ($F=3.81$, $p=0.0278$). A post-hoc Bonferroni-Holm correction revealed that the only significant pairwise difference was between the multi and the single mouse conditions. This suggests that participants were least likely to discover new strategies in the single-mouse condition, which again reinforces the idea that participants were not exploring the strategy space as thoroughly in the single mouse condition as they were in the multi-mouse condition.

One possible explanation for this marker of a lack of exploration is the lack of multi-user control - it can be hard to make meaningful change to a proposed solution when only one person is working on doing so, as individuals may be more prone to prematurely committing to a strategy, and thus more likely to explore variations around that strategy rather than exploring the strategy space more broadly. When we look at the averages of the number of good strategies discovered

on a per-trial basis, however, the differences become even more apparent. (This actually might be a more fair comparison, since the speed of use of the interfaces differed, as seen in the differences in the number of trials participants could complete in each condition). Although in the multi-mouse condition, participants discovered more unique good strategies in total, when the number of unique strategies discovered is averaged by trial, we see that the paper condition averages the largest number of unique good strategies discovered per trial (1.01), with multi-mouse not far behind (0.94), and the single mouse condition showing once again the smallest average of unique strategies discovered (0.61). This difference was significant according to a within-subject ANOVA ($F=4.67$, $p=0.01316$). A post-hoc Bonferroni-Holm correction revealed that there were two significant pairwise differences, between the paper and the single mouse conditions, and between the multi and the single mouse conditions. This suggests that participants explored the solution space more in the paper and multi-mouse control conditions, and did so more efficiently, discovering an average of 1.04 and 0.94 strategies in each trial, respectively. The much lower ratio of 0.61 discovered strategies per trial in the single mouse case further suggests that participants did not explore as broadly in that condition.

Table 5. The number of unique good strategies discovered by groups

	Paper	Multi-mouse	Single mouse
Average Unique Strategies, per Group	5.30	5.83	4.03
(STDEV)	(4.40)	(4.56)	(3.39)
Average Unique Strategies, per Trial	1.01	0.94	0.61
(STDEV)	0.93	0.88	0.48

4.1.4. Does the interface type affect how long it took learners to discover good strategies?

We also analyzed the data to determine the latency of discovery for the good strategies. We wanted to observe how the interface designs influenced this discovery process. By "latency of discovery" we refer to how many trials were needed before a group employed a given strategy – a strategy used during trial 1 would have a discovery latency of 1, and a strategy used first in the third trial would have a discovery latency of 3. To obtain these numbers, we first converted each trial into a binary 16-tuple indicating the presence or absence of each of the 16 significantly "good" strategies highlighted by the regression model (the 16 positive strategies out of the 18 strategies affecting infiltration). Then we multiplied each tuple by the order of that trial within its condition – so if a given trial was the fourth attempted, any 1s in the 16-tuple would be converted to 4s. Then, for each of the 3 conditions within the 30 experiments, we created another 16-tuple that recorded the earliest occurrence of each of the 16 strategies.

Table 6. Illustration of Group 9's latency of strategy discovery.

	Effective Spatial Strategies															
Trial	**1**	**4**	**5**	**7**	**2**	**3**	**9**	**11**	**13**	**15**	**16**	**8**	**10**	**12**	**14**	**18**
1																
2																
3																
4																
5																
6																
7																
8																
9																
	3	**2**	**2**	**3**							**6**		**5**	**5**		

Table 6 is an illustration of how Group 9's exploration of the strategy space proceeded in the multi-mouse condition, and how we distilled the latency of their strategy discovery. They uncovered the clumped univariate strategies (strategy 1 is clumped univariate at radius 1, 4 is clumped univariate at radius 3, 5 is clumped univariate at radius 8, and 7 is clumped univariate at radius 11 respectively – see Tables 1 and 2 for reference) fairly early during their 9-trial exploration, and took a bit longer to uncover the clumped bivariate strategy at radius 9 (strategy 16) and the overdispersed bivariate strategies at radii 3 and 5 (strategies 10 and 12). To compute the average discovery delay, we would average the discovery delay for each of these employed strategies: (3+2+2+3+6+5+5)/ 7 = 3.71. The following table (Table 7) was constructed with the average of the sums of these trials.

Table 7. Summary of the latency of appearance of good strategies (as identified by the multi-linear regression model) in trials generated in the three conditions.

Conditions	**Paper**	**Multi Mouse**	**Single Mouse**
Good strategy appearance count	159	175	128
Average first appearance of good strategies (STDEV)	2.77 (2.16)	2.82 (2.21)	3.70 (2.59)

Table 7 suggests that the interface designs had some impact on the latency of discovery of the good strategies. Moreover, in the paper and multi-mouse interface the learners were found to be discovering the good strategies faster than single mouse trials (see Table 8).

Table 8. ANOVA results for latency of discovery of good strategies

Order Discovery	Sum of Squares	Df	Mean Square	F	Sig.
Between Groups	74.931	2	37.465	7.05	0.001
Within Groups	2439.115	459	5.314		
Total	2514.045	461			

Table 9. Post-hoc tests for the order discovery of good strategies

(I) condition	(J) condition	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
paper	**multi**	-0.044	0.253	0.984	-0.64	0.55
	single	-.922*	0.274	**0.002***	-1.57	-0.28
multi	**paper**	0.044	0.253	0.984	-0.55	0.64
	single	-.878*	0.268	**0.003***	-1.51	-0.25
single	**paper**	.922*	0.274	**0.002***	0.28	1.57
	multi	.878*	0.268	**0.003***	0.25	1.51

An analysis of variance (ANOVA) on these scores (see Table 9) again yielded significant variation among conditions, $F(2, 259) = 7.05$, $p < 0.05$. A post hoc Tukey test showed that the paper and multi-mouse conditions differed significantly from the single mouse condition at $p < 0.05$; indicating that the interface designs were in fact influencing the strategy discovery latency. Again, the fact that there were no significant differences between the paper and multimouse conditions indicates that permitting all participants to contribute to the solution generation is perhaps more important of a factor than the modality of the interface (i.e., tangible versus mouse), although there was a slight non-significant improvement in strategy discovery for the tangible paper interface.

5. CONCLUSION AND FUTURE WORK

Our work demonstrates the potential for educational data mining approaches to help educators develop models of effective solution strategies in rich, open-ended problem spaces. To perform this work, we attempted two different state-space reduction approaches, one that was top-down and another that was bottom-up. The top-down state reduction approach used our assumptions about what might be markers of meaningful spatial manipulations, but we failed to find any relationship between learners' problem-space exploration behaviors and their ability to attain good outcomes. The bottom-up state space reduction approach used spatial metrics to construct isotropic "profiles" of the 2D patterns present in each solution constructed by the learners, and

then applied multilinear regression to identify which of these 2D patterns played a role in producing good outcomes (i.e., which patterns counted as "good" spatial strategies for this problem space). This bottom-up approach allowed us to focus our analysis on how learners explored the solution space (as opposed to the much larger problem space).

We subsequently demonstrated that "discovered" models of learner strategies can be used to evaluate and compare different learning environment designs. We originally set out to explore if a Tangible User Interface (TUI) could offer special affordances for reasoning about 2D spatial patterns, as embodied cognition and Gestalt perception theories might predict. The results we uncovered here provide evidence that user interface design can in fact impact how learners explore 2D spatial problem spaces, but the tangible nature of the UI may be less important a factor than the collaborative nature of the interface. The recommendation to designers seems to be that if one wants to promote more extensive and earlier exploration (as opposed to exploitation) of productive 2D spatial strategies, providing an interface that allows multiple users to all contribute to the solution is a good approach.

This work also enables a number of subsequent analyses to be performed as future work: examining if learners "get stuck" exploring patterns at certain radii (which would indicate that they might need guidance to help them consider incorporating new spatial scales into their conceptions of what "counts" as a solution), if learners struggle to perceive certain types of spatial patterns (an initial examination of the data, not reported on here, suggests that learners might take longer to realize that overdispersion is a strategy that can be employed), if certain types of explorations of the strategy space (for example, more systematically combining good strategies) is more effective at discovering good outcomes, and, ultimately, if any of these observable patterns of behavior result in a greater understanding of spatial phenomena (for which we would need to conduct another experiment where we interview participants before and after their use of the software in a more ecologically valid setting than a 40-minute lab experiment).

There are whole class of problems involving allocentric spatial reasoning that are appearing in national educational standards, but which are not currently instructionally supported, owing to a lack of teaching tools and a lack of assessment approaches. While traditional spatial ability tests may correlate with general visuospatial skills, they tend to stress egocentric skills and thus aren't particularly relevant to the 2D allocentric spatial problems faced by learners, meaning that they are useless for studying how learners acquire such knowledge or for giving learners formative feedback. We used a particular way of characterizing 2D spatial relationships (the Ripley's L metric) that had special relevance to our problem space, but there are a welter of other methods for characterizing 2D point patterns. (While many ecological, biological, and anthropological processes can be represented as 2D spatial patterns, not all of these can be properly summarized using the Ripley's L metric). We urge other researchers interested in characterizing 2D patterns to select a spatial metric which leaves the special spatial characteristics of their data set intact when reducing the solution space from a purely Cartesian representation. Ideally, researchers should have some *a priori* idea, as we did concerning the Ripley's L metric, that a given 2D point pattern characteristic is relevant to their problem space. Even if researchers are uncertain which patterns constitute good strategies, they can use a bootstrapping approach, as we did, to discover effective strategies, or if their problem space is small enough they can construct a testing set from first principles using their 2D point pattern metric as a guide.

We had devised our analytical approach with the initial goal to give feedback to educational software designers, but it could also be adapted to provide learners with formative feedback on their progress reasoning about 2D patterns. Other researchers have shown that learners tend to rehearse 3D spatial understandings via physical actions (gestures) before they acquire the vocabulary to describe spatial phenomena (Singer, Radinsky, & Goldman, 2008). It seems that the same may be true for 2D spatial representations: although we did not present results from our dialogue analysis here, few participants were able to verbally articulate their spatial strategies, and those that did, used imprecise and non-disciplinary vocabulary ("a lot over here" to indicate clumping, or "spread out" to describe overdispersion). We argued in the introduction to this paper that one of the values of highly-interactive digital learning environments is that learners can get a chance to get experience disciplinary processes and dispositions. By helping learners realize that (a) they are in fact employing a spatial strategy (many participants seemed to be doing so unconsciously, as Gestalt theories of perception might predict), and (b) that strategy can be precisely described (for example, that they are avoiding clumping at a radius of 1 while employing clumping at a radius of 4), we can empower learners to both acquire the language of the discipline as well as orient them towards thinking explicitly about how 2D patterns can affect ecological processes. More learning sciences research is needed to better understand how to productively communicate such formative feedback to learners.

In future work we wish to follow up on this nascent exploration of the model space underpinning spatial reasoning in this domain, and to use our findings to help guide learners to more productively and methodically explore spatial problem spaces. We see great potential in this, as it is too often the case that learners can get "lost" exploring open-ended problems, which could result in them not being able to get adequate exposure to comparing and contrasting effective strategies. As we further refine our model of the spatial reasoning learners may exhibit in this domain, we would be able to devise software-based interventions (termed "scaffolding" in the education literature) to guide the learner towards better explorations even in a complex solution space environment.

6. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation (NSF) under Grant Nos. 1020065 and 1135572. Any opinions, findings, interpretations, conclusions or recommendations expressed in this material are those of its authors and do not represent the views of the NSF.

7. REFERENCES

AMERSHI, S., & CONATI, C. (2009). Combining Unsupervised and Supervised Classification to build user models for Exploratory Learning Environments. *Journal of Educational Data Mining, 1* (1), 18-71.

ANDERSEN, E., YUN-EN LIU, APTER, E., BOUCHER-GENESSE, F., & POPOVIC´, Z. (2010). Gameplay Analysis through State Projection. *International Conference on the Foundations of Digital Games.* Monterey, California.

ANTLE, A. N., DROUMEVA, M., & HA, D. (2009). Hands on what? Comparing children's mouse-based and tangible-based interaction. *Proceedings of the 8th International Conference on Interaction Design and Children* (pp. 80-88). ACM.

BAKER, R. S., & YACEF, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining , 1* (1), 3-17.

BERLAND, M., BAKER, R. S., & BLIKSTEIN, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. *Technology,Knowledge, and Learning , 19* (1-2), 205-220.

BERLAND, M., MARTIN, T., BENTON, T., SMITH, C., & DAVIS, D. (2013). Using Learning Analytics to Understand the Learing Pathwaysof Novice Programmers. *Journal of Learing Sciences* , 564-599.

BISWAS, G., LORETZ, K. M., & SEGEDY, J. R. (2013). Model-Driven Assessment of Learners in an Open-Ended Learning Environment. *Third International Conference on Learning Analytics and Knowledge.* New York, NY.

BLIKSTEIN, P. (2011). Using learning analytics to assess students' behavior in open-ended Programming tasks. *Learning Analytics and Knowledge.* New York.

BRAVO, M., HERNANDEZ, J., SAORIN, J., & CONTERO, M. (2010). A 3D Educational Mobile Game to Enhance Student's Spatial Skills. *IEEE 10th International Conference on Advanced Learning Technologies (ICALT).* Sousse,Tunisia: IEEE Computer Society.

CONNELL, M., & STEVENS, D. (2002). A computer-based tutoring system for visual-spatial skills: dynamically adapting to the user's developmental range. *Proceedings of the The 2nd International Conference on Development and Learning.* Cambridge: IEEE COmputer Society.

DALE, M. (1999). *Spatial pattern analysis in plant ecology.* Cambridge, UK: Cambridge University Press.

DESMARAIS, M., & LEMIEUX, F. (2013). Clustering and Visualizing Study State Sequences. *6th International Conference on Educational Data Mining.* Memphis, Tennessee.

DICERBO, K. E., & KIDWAI, K. (2013). Detecting Player Goals from Game Log Files. *6th International Conference on Educational Data Mining.* Memphis, Tennessee.

DIXON, P. M. (1995). Ripley's K function. In *Encyclopedia Environmetrics* (pp. 1796-1803). NJ: Wiley.

EAGLE, M., & BARNES, T. (2014). Exploring Differences in Problem solving with Data-Driven Approach Maps. *Educational Data Mining.* Indianapolis.

EKSTROM, R. B., FRENCH, J. W., & HARMON, H. H. (1976). *Manual for the Kit of Factor-Referenced Cognitive Tests.* ETS.

FALAKMASIR, M. H., PARDOS, Z. A., GORDON, G. J., & BRUSILOVSKY, P. (2013). A Spectral Learning Approach to Knowledge Tracing. *6th International Conference on Educational Data Mining.* Memphis, Tennessee.

FALCÃO, T., & PRICE, S. (2009). What have you done! The role of 'interference' in tangible environments for supporting collabarative learning. *8th International Conference on Computer Supported Collaborative Learning.* Rhodes, Greece.

FORTIN, M. J., DALE, M. R., & HOEF, J. V. (2002). Spatial Analysis in Ecology. In *Encyclopedia of Environmetrics* (pp. 2051-2058). NJ: Wiley.

FOURNIER-VIGER, P., NKAMBOU, R., NGUIFO, E. M., MAYERS, A., & FAGHIHI, U. (2013). A multiparadigm intelligent tutoring system for robotic arm training. *IEEE Transactions on Learning Technologies , 6* (4), 364-377.

FOURNIER-VIGER, P., NKAMBOU, R., NGUIFO, E., MAYERS, A., & FAGHIHI, U. (2013). A Mutilparadigm intelligent tutoring system for robotic arm training. *Learning Technologies, IEEE Trasactions* , 364-377.

GOBERT, J., SAO PEDRO, M., RAZIUDDIN, J., & BAKER, R. (2013). From Log Files to Assessment Metrics for Science Inquiry Using Educational Data Mining. *Journal of the Learning Sciences, 22* (4), 521-563.

HALPERN, D., & COLLEAR, M. (2005). Sex Differences in Visuospatial Abilities. In P. Shah, & A. Miyake, *Cambridge Handbook of Visuospatial Thinking* (pp. 170-212). New York: Cambridge University Press.

HARPSTEAD, E., MACLELLAN, C. J., KOEDINGER, K. R., ALEVEN, V., DOW, S. P., & MYERS, B. A. (2013). Investigating the Solution Space of an Open-Ended Educational Game Using Conceptual Feature Extraction. *6th International Conference on Educational Data Mining.* Memphis, Tennessee.

HEGARTY, M., & WALLER, D. (2005). Individual and age related differences in visuospatial abilities. In P. Shah, & A. Miyake, *Cambridge handbook of Visuospatial Thinking* (pp. 121-169). New York: Cambridge University Press.

HUBBARD, C., MENGSHOEL, O., MOON, C., & YONG, S. (1996). Multimedia instructional software for visual reasoning: Visual Reasoning Tutor (VRT). *Proceedings of the Third IEEE International Conference on Multimedia Computing and Systems,*. Hiroshima, Japan: IEEE Computer Society.

JARUSEK, P., KLUSACEK, M., & PELANEK, R. (2013). Modeling Students' Learning and Variability of Performance in Problem Solving. *6th International Conference on Educational Data Mining.* Memphis, Tennessee.

JOHNSON, M. W., EAGLE, M., & BARNES, T. (2013). InVis: An Interactive Visualization Tool for Exploring Interaction Networks. *Educational Data Mining.* Memphis.

KARDAN, S., & CONATI, C. (2013). Evaluation of a Data Mining Approach to Providing Adaptive Support in an Open-Ended Learning Environment: A Pilot Study. *Artificial Intelligence In Education , 2*, 41-48.

KIM, M. J., & MAHER, M. (2008). The impact of tangible user interfaces on spatial cognition during collaborative design. *Design Studies , 29* (3), 222-253.

KUTNER, M., NACHTSHEIM, C., & NETER, J. (2004). *Applied Linear Regression Models.* McGraw Hill.

LAKOFF, G., & JOHNSON, M. (1980). *Metaphors We Live By.* Chicago: University of Chicago Press.

LEE, S. J., YUN-EN, L., & POPOVIC, Z. (2014). Learning Individual Behavior in an Educational Game: A Data- Driven Approach. *Educational data Mining.* Indianapolis.

LEVY, S., & WILENSKY, U. (2007). How do I get there...straight, oscillate or inch? High-school students' exploration patterns of Connected Chemistry. *2007 meeting of the American Educational Research Association.* Chicago: AERA.

LIU, Y.-E., MANDEL, T., BUTLER, E., ANDERSON, E., O'ROURKE, E., EMMA, B., ZORAN, P. (2013). Predicting Player moves in Educational Game: A Hybrid Approach. *Educational Data Mining.* Memphis.

LYNCH, C., ASHLEY, K., PINKWART, N., & ALEVEN, V. (2008). Argument graph classification with Genetic programming. *Educational Data Mining.* Montreal, Quebec.

LYONS, L., DASGUPTA, C., SHELLEY, T., SLATTERY, B., MINOR, E., & ZELLNER, M. (2012). Parsing Patterns: Developing Metrics to Characterize Spatial Problem Solving Strategies within an Environmental Science Simulation. *AREA*, (p. 19).

MAIMON, O., & ROKACH, L. (2010). *Data Mining and Knowledge Discovery Handbook.* Springer.

MARSHALL, P. (2007). Do tangible interfaces enhance learning? *Proceedings of the 1st international conference on Tangible and embedded interaction (TEI '07)* (pp. 163-170). ACM.

MARTINEZ-MALDONADO, R., YACEF, K., & KAY, J. (2013). Data Mining in the Classroom: Discovering Groups Strategies at a Multi-tabletop Environment . *6th International Conference on Educational Data Mining.* Memphis,Tennessee.

MASSEY, D., ZELLNER, M. L., COTNER, L., MINOR, E., & GONZALEZ-MELER, M. (2010). *Landscape Green Infrastructure Design Model (L-GrID) User's Manual. Report and software to the Illinois Environmental Protection Agency. Chicago, IL: University of Illinois at Chicago.* Chicago.

MENGSHOEL, O., CHAUHAN, S., & YONG, S. (1996). Intelligent critiquing and tutoring of spatial reasoning skills. *Artificial Intelligence for Engineering, Design, Analysis and Manufacturing* , 235-249.

MINOR, E. S., & URBAN, D. L. (2008). A Graph□Theory Framework for Evaluating Landscape Connectivity and Conservation Planning. *Conservation Biology , 22* (2), 297-307.

MULLER, J., KRETZSCHMAR, A., & GREIFF, S. (2013). Exploring Exploration: Inquiries into Exploration Behavior in Complex Problem Solving. *6th International Conference on Educational Data Mining.* Memphis, Tennessee.

NESBITT, K., SUTTON, K., WILSON, J., & HOOKHAM, G. (2009). Improving player spatial abilities for 3D challenges. *The 6th Australasian Conference on Interactive Entertainment.* Sidney,Australia.

PROTÁZIO, J., PEREIRA, W., & ELAYNE JESUS DE CASTRO, F. (1999). Explicit Formulas for an Area Based Edge Effect Correction Method and their Application to Ripley's K-Function. *Journal of Vegetation Science , 10* (3), 433-438.

RAFFERTY, A. N., DAVENPORT, J., & BRUNSKILL, E. (2013). Estimating Student Knowledge from Paired Interaction Data. *6th International Conference on Educational Data Mining*, (pp. 260-263). Memphis, Tennessee.

ROCK, I., & BROSGOLE, L. (1964). Grouping based on phenomenal proximity. *Journal of Experimental Psychology , 67* (6), 531-538.

ROMERO, C., & VENTURA, S. (2010). Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on , 40* (6), 601-618.

SCHNEIDER, B., JERMANN, P., ZUFFEREY, G., & DILLENBOURG, P. (2011). Benefits of a Tangible Interface for Collaborative Learning and Interaction. *IEEE Transactions on Learning Technologies , 4* (3), 222-232.

SCHWEINGRUBER, H., KELLER, T., & QUINN, H. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas.* National Academies Press.

SHELLEY, T., LYONS, L., MINOR, E., & ZELLNER, M. (2011). Evaluating the embodiment benefits of a paper- based TUI for educational simulations. *29th International Conference on Human Factors in Computing Systems.* New York, NY.

SHELLEY, T., LYONS, L., SHI, J., MINOR, E., & ZELLNER, M. (2010). Paper to parameters: designing tangible simulation input. *12th ACM International Conference adjunct papers on Ubiquitous computing.* New York, NY.

SHUTE, V. J. (2011). Stealth Assessment in computer based games to support learning. In *Computer Games and Instruction* (pp. 503-524). Charlotte, NC: Information Age Publishers.

SINGER, M., RADINSKY, J., & GOLDMAN, S. R. (2008). The Role of Gesture in Meaning Construction. *Discourse Processes: A Multidisciplinary Journal , 45* (4-5), 365-386.

SISWONO, T. Y. (2008). Promoting creativity in learning mathematics using open-ended problems. *The 3rd International Conference on Mathematics and Statistics (ICoMS-3).* Indonesia.

SMITH, A., WIEBE, E., MOTT, B., & LESTER, J. (2014). SKETCHMINER: Mining Learner-Generated Science Drawings with Topological Abstraction. *Educational Data Mining.* Indianapolis.

STIEFF, M., DIXON, B., KUMI, B., & HEGARTY, M. (2014). Strategy Training Eliminates Sex Differences in Spatial Problem Solving in a STEM Domain. *Journal of Educational Psychology, 106* (2), 390-402.

UTTAL, D., MEADOW, N., TIPTON, E., HAND , L., ALDEN, A., WARREN, C., NEWCOMBE, N. S (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin* , 352-402.

VAN LEHN, K. (2011). The relative effectiveness of Human Tutoring, Intelligent Tutoring Systems and other Tutoring System. *Educational Psychologist* , 197-221.

VYGOTSKY, L. (1978). *Mind in Society: The development of higher mental processes.* Cambridge, MA: Harvard University Press.

WAI, J., LUBINSKI, D., & BENBOW, C. (2009). Spatial ability for STEMdomains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology* , 817-835.

WANG, E., & KIM, Y. (2005). Intelligent Visual Reasoning Tutor. *Proceedings of the Fifth IEEE International Conference on Advanced Learning Technologies* (pp. 511-515). Washington,D.C: IEEE Computer Society.

WERTHEIMER, M. (1923). Untersuchungen zur Lehre von der Gestalt II. *Psycologische Forschung, 4*, 301-350.

WIEDERRECHT, M. A., & ULINSKI, A. C. (2012, January). Developmentally appropriate intelligent spatial tutoring for mobile devices. *Intelligent Tutoring Systems* , 594-596.

WILENSKY, U. J. (1993). *Connected Mathematics-Building Concrete Relationship with Mathematical Knowledge. Dissertation of Doctor of Philosophy, Massachusetts Institute of Technology.*

ZELLNER, M. L. (2008). Embracing Complexity and Uncertainty: The Potential of Agent-Based Modeling for Environmental Planning and Policy. *Planning Theory and Practice , 9* (4), 237-457.