# STUDYING SITUATED LEARNING IN A MULTI-USER VIRTUAL ENVIRONMENT

Diane Jass Ketelhut, Chris Dede, Jody Clarke, Brian Nelson, and Cassie Bowman,

Harvard Graduate School of Education

This material is based upon work supported by the National Science Foundation under Grant No. 0310188. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Multi-User Virtual Environments (MUVEs) enable multiple simultaneous participants to access virtual contexts (such as graphically represented buildings), interact with digital artifacts and tools (such as digitized images and virtual microscopes), represent themselves through "avatars" (graphical representations of participants), communicate with other participants and with "agents" (personalities simulated by a computer), and enact collaborative learning activities of various types (Dede et al, 2005). Most students now using multi-user virtual environments (MUVEs) do so in the context of online gaming (Dede, 2005) through commercial products such as the Sims™ (various lifestyles), Everquest™ (swords and sorcery), Grand Theft Auto™ (crime), or Halo™ (war). Our research group is studying the motivational power of MUVEs and the sophisticated learning processes they enable to increase educational outcomes for deep academic knowledge and higher order thinking skills, as opposed to the largely useless fantasy content and skills participants gain from these entertainment products. This chapter describes multiple ways in which the detailed record of student actions and utterances automatically collected in MUVEs offers great potential for assessment, both from a research perspective and in terms of formative, diagnostic information that could help tailor instruction to individual needs.

## Overview of "River City" Learning Environment

Using a MUVE as a pedagogical vehicle, our research team is exploring how a technology-intensive learning experience that immerses participants in a virtual "world" whose residents face chronic illnesses can help middle school students learn both deep inquiry skills (scientific processes) and science knowledge. In particular, we are working to dramatically

improve the educational outcomes of students performing in the bottom third of their class, pupils who—even by middle school—frequently have given up on themselves as learners. These students are disengaged from schooling and typically are difficult to motivate, even by good teachers using best-practice, inquiry-based pedagogies. We are investigating whether the use of educational MUVEs, which resemble the entertainment and communication media kids use outside of school, can both reengage these students in learning and aid them in mastering higher order thinking processes, as well as standards-based content in biology and ecology (National Research Council, 1996). This focus of our MUVE, "*River City*", grew out of conversations with science teachers about the subject areas they had the most difficulty teaching: problem finding, hypothesis formation, and experimental design.

The virtual "world" of *River City* consists of a city, set in the late 1800's, with a river running through it, different forms of terrain that influence water runoff, and various neighborhoods, industries, and institutions such as a hospital and a university. Upon entering the city, the students' avatars can interact with each other, computer-based agents, digital artifacts, and the avatars of instructors (Figure 1). In exploring, students also encounter various visual and auditory stimuli that provide tacit clues as to possible causes of illness. Content in the right-hand interface-window shifts based on what the participant encounters or activates in the virtual "world", such as a digital, interactive microscope that allows students to examine samples of water (Figure 2).

Place

Figure 1 here

Place

Figure 2 here

In *River City*, students work in teams to develop hypotheses regarding one of three strands of illness in the town (water-borne, air-borne, and insect-borne). These three disease strands are integrated with historical, social, and geographical content, allowing students to experience the inquiry skills involved in disentangling multi-causal problems embedded within a complex environment. At the end of the curriculum, student teams compare their research findings with other those from other groups of students in their class to delineate some of the many potential hypotheses and causal relationships embedded in the virtual "world".

In 2002, we conducted our pilot implementations of *River City* and a matched control curriculum in four public school classrooms in a large urban area in Massachusetts with high percentages of English language learners and students receiving free or reduced lunch. We examined usability, student motivation, student learning, and classroom implementation issues (Dede & Ketelhut, 2003; Dede, Ketelhut & Reuss, 2002). Whole classes were assigned to the treatment (using "*River City*") or to the control. The control condition utilized a curriculum in which the same content and skills were taught in equivalent time to comparable students via a guided social constructivist-based pedagogy (Vygotsky, 1978), but in a paper-based format without the use of computers. Activities were matched closely between the two curricula. This type of control curriculum enables us to focus on the strengths and limits of MUVEs, as well as the types of pedagogy best supported by this medium, as compared to best-case science teaching far superior to the typical content and pedagogy in our implementation sites.

Our results from the pilot study indicated this first-generation MUVE was motivating for all students, including lower ability students typically uninterested in classroom activities. The *River City* group, on average, had more positive changes in motivation than did the control group. Subscale averages for students' perceptions of global science self-efficacy also showed

significant differences between the two groups ($t=3.36$, $p< .05$), with the *River City* group showing an increase of one point out of five on average as opposed to the control group's decrease of .31 (Dede & Ketelhut, 2003).

We also found that students discovered multiple intriguing situations in the MUVE to investigate. In one classroom, five different hypotheses about the illnesses emerged, with posited causes ranging from population density to immigration to water pollution. Another finding is that the MUVE seemed to have the most positive effects for students with high perceptions of their own thoughtfulness of inquiry. These students, on average, scored the highest on the post content test, controlling for SES, science GPA, ethnicity, and content pre-test score (Dede & Ketelhut, 2003). Also, the data showed no differences between English language learning (ELL) and non-ELL students in performance, despite the reliance in this medium on reading and writing, indicating the success of our design strategy that ensured each team of students included someone who could read English.

Based on this pilot work, in our current research we have developed multiple, second-generation variations of the *River City* curriculum leading to three treatments along with a control. One variant centers on a guided social constructivist (GSC) model of learning-by-doing, in which inquiry experiences in the MUVE are supported by both virtual and physical lab notebooks and complemented by in-class interpretive sessions led by the teacher. Another variant shifts the learning experience to a situated pedagogy based on expert modeling and coaching (EMC), as described by Brown, Collins and Duguid (1989), in which students interact with expert avatars (played by college science majors) and computer-based agents embedded in the MUVE. The third variant, Legitimate Peripheral Participation (LPP), is based on Lave and Wenger's (1991) concept of a community of practice; students move from simple peripheral

roles to more complex tasks through tacit forms of learning such as internships that are supported by a computer-based agent and the lab notebook.

Through a series of implementations from 2003 to 2005, these *River City* variants were compared to a control condition similar to that in our pilot research. Based on studies of these variants, a "best-of-the-best" version of the *River City* curriculum has recently been developed to include the most successful features of each variant, as well as an individualized, computer-based guidance system (Nelson, 2005). We implemented this version in a variety of classrooms during May-October, 2005 and are just beginning our analysis of it.

In 2003-2004, to assess the effectiveness and value of the "*River City*" MUVE curriculum, we conducted four large-scale implementations involving more than 2500 students in major urban areas in New England, the Midwest, California and the Southeast, in schools with high proportions of English language learners and students receiving free or reduced lunch. Depending on the implementation, different computer-based variants were randomly assigned to students within each classroom, with teachers instructed to minimize cross-contamination of treatments. The paper-based control treatment was randomly assigned to whole classes, with each teacher offering both the computer-based treatments and the control. After two weeks of designing and conducting their experiments, students in both the control and *River City* treatments were asked to write letters to the Mayor of *River City* in which they discussed their hypothesis, experimental design, findings, and recommendations for solving the city's health problem. More detailed information about the River City curriculum and our research findings is available at http://muve.gse.harvard.edu/rivercity/.

## River City as a Database

From a technical standpoint, MUVEs are unique in their ability to keep minutely detailed records of the moment-by-moment movements, actions, and utterances of each participant in the environment. As part of the MUVEES project, Active Worlds Corporation (www.activeworlds.com) has developed a customized 'plug-in server' that provides additional functionality to the *River City* MUVE environment. The features most relevant to our research are the data-tracking system and log files. With the data-tracking system, we are able to collect, store, and retrieve information on the activities of each student as s/he explores the MUVE. These data form the basis of a personal MUVE history of each student that follows him or her from session to session, in the form of extensive log files—a feature impossible to replicate in a classroom-based experience. The level of detail in these records is extensive: the logs indicate exactly where students went, with whom they communicated, what they said in these interactions, what virtual artifacts they activated, and how long each of these activities took.

To create these personalized histories for each student in *River City*, all the items with which students could interact were programmatically tagged with identification codes. Every time a student clicks on an object or 'speaks' to a *River City* resident (a computer-based agent) or other avatars, a record of the event is stored in a server-side database. A wide array of objects with which students can interact is available in *River City*, and consequently a richly varied store of data exists from which to build a sophisticated analysis of student participation in the MUVE.

A more detailed description of River City will serve to illustrate the complexity of potential student actions. *River City* features a river running through the town (Figure 3) and is divided into several geographic zones. The river begins in the mountainous upper elevations of town, where the wealthy residents live and the university is housed. It then travels down past the

middle class homes, curving around south of the downtown shopping district, and slowly flowing through the tenement district along the dump. In addition to the river, there is a bog in which tenement residents often swim, bathe, and even wash their clothes. A centrally located train station is a beehive of activity with new residents arriving daily; nearby, *River City's* hospital cares for the town's sick.

Place Figure 3 here.

Within each area of town, there are a number of items with which students can interact to gain information about the illnesses in *River City*. These include:

- *Historical photos and accompanying text:* Buildings throughout town contain digitized historical images. Clicking on an in-world image causes a web page to appear in the right hand screen that contains information about the town, its people, environmental factors, and other clues about the illnesses in town.
- *Books and charts:* Scattered throughout the world are digital signs and books that link to web pages containing additional information. For example, the hospital has an admissions chart containing a list of newly admitted patients, their ages, addresses, and symptoms. The town library has virtual books that link to online dictionaries and encyclopedias.
- *Data stations:* Data collection stations appear in some sessions of the world. Clicking on these stations opens a virtual tool that provides a microscopic view of the water in specific place, allowing students to collect and analyze samples.

In addition, students can use the chat system in the MUVE to ask a limited range of questions addressed to the *River City* residents. There are 32 residents (computer-based agents) scattered throughout town. These residents offer short sets of information about happenings in town and the illnesses. As an illustration, the agents can respond with a short answer to the question "What's new?" Asking what's new of Nurse Patterson in the *River City* Hospital, for example, elicits a clue about the recurring stomach illness in *River City*, "*We have a couple of new patients with that stomach upset we saw last summer. I hope this isn't becoming an epidemic again! Most of them seem to be from the tenement homes."*

Through collecting data about how students interact with the various types of artifacts described above, we can analyze student movement through the world, looking at patterns of movement, interactions, chat, and questioning of *River City* residents. Combining these data with basic student-level predictors—demographic characteristics such as gender, SES, age, and ethnicity; pre- and post-test scores on quantitative measures of scientific knowledge; judgments along multiple dimensions about the quality of their "summative performance" letter to the mayor; and, for some students, in-class observations and pre/post interviews—allows for a sophisticated analysis of how participant characteristics interact with in-world activity over time, and how these interactions may influence student learning.

## Analysis of the River City Database Information

Analysis on this extensive *River City* database is in its early stages, so the following findings represent some preliminary trends in that data. Sample sizes are noted. As we discuss above, essentially all student interactions are recorded in this database, offering an opportunity to

see variations in patterns of student behavior nearly impossible to detect in any other type of learning context.

The early analysis of student-to-*River City* resident conversation for four classes having the same teacher (n=96) illustrates this potential power for detailed student assessment (Crusoe, 2005). In an attempt to provide a virtual world comfortable for girls, we intentionally created more female residents (twenty-one) than male (thirteen) in *River City*. Also, to supply role models in science for girls, four of the five resident (computer-based) experts are female. Therefore, one research question we are interested in exploring is whether or not students discriminate between the residents to whom they talk, based on the gender of the resident.

The students in these four classes logged in over 3,000 conversational gambits to various *River City* residents during the two-week implementation. 40% of the conversations that boys had with *River City* residents were directed towards male figures; similarly, 39% of the conversations that girls had with residents were directed toward male residents. These numbers roughly reflect the percent of male residents of *River City*. From this evidence, it does not appear as if gender of the residents is affecting conversational choices. However, upon further investigation, we discovered that, while boys constitute 46% of the students in this sample, they only account for 31% of these conversational entries. We are now conducting additional analyses of data to determine if boys differentially preferred other methods of gaining information (e.g., exploring, interacting with artifacts), rather than conversation with residents.

Additionally, there is a difference in what students said to residents. While the residents of *River City* can understand and respond to "hello," "thank you," "where am I," and "good-bye," "what's new" is the only phrase that will elicit a clue about the problems in *River City*. For the students in this subsample, "what's new" constituted 35% of the girls' comments to residents,

while this phase was uttered 48% of the time for boys. So, while boys talk less frequently than girls, their conversation is relatively more task-oriented; in contrast, girls' interactions with residents frequently first attempt to establish a social relationship. It is too early for us in this analysis to make claims or offer explanations for these trends, but we offer them here instead as an example of the power of multi-user virtual environments to help illuminate students' learning patterns.

## River City Individualized Guidance System

As another example of the rich analytic power possible through the use of MUVE log files, our project is the first to create and investigate the use of an embedded individualized guidance system (IGS) with an educational MUVE. The guidance system utilizes personalized interaction histories collected on each student's activities to offer real-time, customized support. The IGS offers reflective prompts about each student's learning in the world, with the content of the messages based on in-world events and basic event histories of that individual. To create the IGS, all the items with which students could interact were programmatically tagged with identification codes. Every time a particular student clicked on an object or 'spoke' to a *River City* resident, a record of the event was stored in a server-side database. The cumulative record of events resulted in a personalized history for each student.

A guidance model, operated by a back-end software agent, was triggered after each student interaction event in the MUVE. A subset of events was associated with guidance scripts, and the guidance model used these scripts to offer a specific selection of messages to each student. As a default, three links to guidance messages related to the current world object are displayed. The default messages are designed to help students process information on the content

or attributes of the object itself. For example, there are a large number of clickable pictures hanging on walls around River City. When students click on these, a webpage appears showing a brief paragraph providing information about the picture and/or the illnesses in the city. If a given picture includes guidance hints, the default hints will present students with guiding questions about the text associated with the picture.

In addition to the default messages, each tagged object could display up to three customized guidance messages. These customized messages were displayed when a predefined set of prerequisite objects or interactions existed within a given student's personal history. For example, a clickable admissions chart is located in the *River City* hospital. When a student clicked on this chart, the guidance model read a script to see if the given student had previously clicked on objects defined as prerequisites for showing customized messages. Up to three sets of prerequisite interactions could exist in the rules script for any tagged object (one set for each customized guidance message). In the example case of clicking on the admissions chart, a predefined rule stated that, if the student had previously visited the tenement district and talked to a resident there, then a customized guidance message would be shown reminding the student that he/she had previously visited the tenement district, and asking the student how many patients listed on the chart came from that part of town.

Place Figure 4 here

Multilevel multiple regression analysis findings show that use of this guidance system with our MUVE-based curriculum has a statistically significant, positive impact ($p<.05$) on student learning for both girls and boys (Nelson, 2005). This fitted relationship is plotted for

both boys and girls in Figure 4. Students who viewed more guidance earned higher GAIN scores than those who viewed fewer messages, and the benefit of guidance system use varied by gender, with boys doing worse, on average, at each level of guidance viewing. We are currently studying possible explanations for these outcomes.

In addition to using the log files to personalize the guidance provided to each student, we are able to make use of this data to conduct sophisticated analyses of guidance use. We know when and if students first choose to use the guidance system, which messages they view, where they are in the virtual world when they view them, and what actions they take subsequent to viewing a given guidance message. This provides diagnostic information that potentially teachers could access to gain formative assessment insights to guide their daily instruction.

## Case studies

The logfiles of student experiences also enable another type of assessment: tracking the learning trajectory of individual students to see what aspects of *River City* engaged them and were helpful in their learning. An example of this can be seen in the transformation of Kimmie (a pseudonym) to "Shorty" (Kimmie's *River City* persona), as she gains critical thinking skills (Partnership for 21st Century Skills, 2003) in the course of her longitudinal experiences in *River City*. Based on logfile analysis and interview transcripts, we can see Shorty move from the periphery of the decision-making process of her team to a more central role in her team's experimenting (Clarke, 2005).

Kimmie is a 12-year-old female in the 7$^{th}$ grade. She does not think she is good at science and claims "...it's boring when he (her teacher) talks a lot." Kimmie appears to view science differently from science class. As a pre-assessment, we ask each student to draw a

picture of herself doing science in science class. Kimmie drew herself in a science laboratory doing an experiment, and explained it as "once we did an experiment with a ruler to see if you could … catch it fast." However, when asked if the picture captures how she usually feels in science class, she replied: "sometimes he just teaches us like the things about like about science and then we study and we have tests on it."

Kimmie further describes her science class as, "We write notes and sometimes we do experiments…Well, when he usually just talks a lot, like I remember one day, we didn't do nothing except for write notes and he talked a lot." Not surprisingly, on a scale of one to five, five being her most favorite, she rated her science class as 2. If given a choice, Kimmy would not take science classes in high school "...because ...I am not really interested in science."

Both Kimmie and her teacher describe her stance in class as a follower rather than a leader. In the beginning of the project, as Shorty, she relies on her teammates, hype69 and rhia, to tell her what to do:

*Shorty:* *hype69 rhia what do we have to do*

*Shorty:* *what d we do*

*Rhia:* *Doctor Paterson Said they cant find a cure yet for the new problem.*

*Shorty:* *what page is it on*

*Rhia:* *17*

*Shorty:* *what else do we do*

As the project progresses, Shorty makes observations and inferences about her discoveries and shares them with her teammates. However, initially her low "self-efficacy" in science (belief that one can succeed in learning a particular subject) still causes her to defer to Rhia's discovery over her own observations:

*Shorty:* *miss howell said that there is a bad stomach problem at the tenements*

*Hype69:* *wathank you*

*Rhia:* *Mrs Lopas saiys that the sewage runs into the stream where the people get water*

*Shorty:* *is that the problem*

*Shorty:* *because if it is do we tell the teacher*

*Rhia* *yes*

Even though Shorty observed the trash in the water near the tenements and found information from Miss Howell, she refrains from pushing her point:

*We wanted to move the sewage pipe. I mean, I thought it was the dump but I think it was both so sewage was going down near the tenements but the dump was right there.*

At this point, we see that Shorty is beginning to move from the periphery of the team towards the center in terms of leadership, because she has begun to espouse her own theory without need for direction. However, at this stage of the *River City* experience, she still yields to her other teammates' ideas. Further logfile analysis indicates that, by the end of the project, Shorty is no longer asking her teammates what she should do, but has begun to tell them what to do. As one illustration, she takes the lead on sampling the water and tells Rhia to go to the hospital:

*Shorty:* *i'm goin to da water samplin sation in the wealthy homes*

*Shorty:* *rhia do da hospital im takin wata samples*

This shift indicates Shorty's growing engagement and self-efficacy.

Students collect data first in a control world and then in an experimental world. In order to compare their data, they need to conduct observations and tests in the same places in both worlds. We can see from the logfiles that Shorty understands this and is directing her teammates to ensure that this is accomplished successfully:

*Hype69: what do we do*

*Shorty: do the same thing u did on friday ask the same people and write down what the said this time*

This represents an important change for Shorty as she moves from the role of a novice to a stance more closely resembling that of an expert (Brown et al, 1989; Lave & Wenger, 1991).

In follow-up interviews, Shorty confirms what we are seeing in the logfiles and helps us make sense of why this shift happened for her. She tells us that she felt like a scientist "because we had to figure out things and ask questions and use our brains and really think hard." Actively solving the problem made it easier to "understand" and helped her comprehend concepts such as "hypothesis" and "procedure":

*...it helped me understand hypothesis and procedure better ... I knew what they were but then this project it was easy to write a hypothesis and procedure for. Ummm because there was more data than sometimes you don't have a lot data and there was a lot because you could ask a lot of people and talk to a lot of people.*

Shorty had learned about these concepts before, but solving a problem by visually seeing tacit information helped make abstract concepts like hypothesis and procedure more concrete:

*We actually got to see where everything is, where the dump is and the tenements and the scenic lookout where the pipe, the sewage pipe was going to the water. Yeah, because you got to do more and get more information and actually see the thing instead of just imagining it. You get to actually see it. It helps because we can ask people questions and they can tell us. Like they can talk to us and we got to walk around and see what the city looks like if it is dirty in some spots and really clean in other spots.*

Shorty learned immersively by walking around, talking to residents, and noting tacit visual and auditory clues, yet we would not have known this solely through examination of the summative affective and content measures. Examining the log files and talking to Shorty about her experience enables understanding the processes that led to her engagement and learning.

## Contrasting Conventional versus Performance-Based Methods of Assessing Learning

Up to this point, we have discussed the power of MUVEs for tracking students' learning trajectories to a degree not possible with other techniques. What are the difficulties in using these methods to measure the complex learning that happens in MUVEs, both for research purposes and for formative, diagnostic assessment to aid instruction?

Our students complete two pre/post measures. The first is an affective measure adapted from three different surveys, Self-Efficacy in Technology and Science (Ketelhut, 2005), Patterns for Adaptive Learning Survey (Midgley, C. 2000), and the Test of Science Related Attitudes (Fraser, 1981). This modified measure has scales to evaluate students' efficacy of technology use (videogame, computer, chat, etc), science efficacy, thoughtfulness of inquiry, science enjoyment, and career interest in science. Its individual subscales have reasonable internal consistency reliability estimates (ranging from .8 to .93), as well as validity evidence from prior research (Ketelhut, 2005).

Our second survey assesses understanding and content knowledge, such as science inquiry skills, science process skills, and biology. Previously published and validated instruments for assessing science process skills are available (Dillashaw and Okey 1979); however, these are not appropriate for the student developmental level and curricular units in our project. Therefore, we have developed our own measures to assess student learning of content

and scientific problem solving skills. Internal consistency reliability was estimated using Cronbach's alpha and Principal Components Analysis (PCA), indicating a reliability of .86. Content validity was established through analysis by a team of experts.

We are still in the early stages of analyzing data from these instruments, but interesting patterns are emerging about which students do best under our various pedagogical treatments. Results of a randomly chosen representative subgroup of students from 4 of the 11 teachers in the first implementation (n=330) were analyzed via multi-level modeling, using students' class assignment as the grouping variable. The examination of the results indicates that, on average, students in the guided social constructivist (GSC) experimental group achieved 16% higher scores on the posttest in biology than students learning with the control curriculum.

Similar results were seen from the affective measures. In the initial pilot implementation (n=81), over the course of the study experimental students raised their average score in *thoughtfulness of inquiry*, a measure of metacognition, significantly more than control students. This initial result was confirmed in this later implementation. Student scores for thoughtfulness of inquiry on the post-survey were significantly higher ($p<.01$) on average for two of the experimental groups, in comparison to the scores for students in the control group. For example, students scoring an average of 1 (strongly disagree) on the scale of 1-5 for the pretest were associated with scores of 1.8 on the posttest for GSC and 1.9 for EMC, nearly double their starting average score (as a result of a curricular intervention of approximately ten hours). Students in the control group also improved, on average, but only to 1.3.

However, when we looked for evidence of improved inquiry skills using our content survey questions, we found equivocal results. Improvements were seen across the board for knowledge and application of scientific processes, but there were no significant differences

between treatments. This result was replicated in all of our implementations to date for nearly 2,000 students. Is this because our project does not affect student learning in this area, or because inquiry is difficult to assess with close-ended survey questions?

This is an issue that has been in the forefront of assessment research for decades, even more so now that standardized national and state testing have taken such a prominent role in accountability sanctions. Interest in assessing student learning for accountability purposes grew in the 1980s, possibly as a reaction to the concerns voiced in *A Nation at Risk* (Kane et al, 1997). The current trend towards high-stakes standardized testing has its roots in this decade.

In reaction to the emphasis on high stakes tests, and based on influential reports calling for the inclusion of more inquiry in science curricula (AAAS, 1990; NRC, 1996), the 1990's saw an increased interest in alternative assessments, such as performance-based assessments in which a student's process in analyzing a problem was revealed (Baxter & Shavelson, 1994; Klein et al, 1997; Stecher & Klein, 1997). Published reviews are available that detail the debate between the proponents of each of these styles of assessments (Mehrens, 1998; Kane et al, 1997; Moore, 2003). To summarize: Proponents of alternative assessments view them as capturing student understanding better than standardized tests, which they feel measure decontextualized knowledge. Opponents argue that performance-based assessments are not cost effective, can not be compared from teacher to teacher due to individual grading differences, and are inconclusive about what the tasks are actually measuring (Stecher & Klein, 1997).

## MUVEs as a Means of Integrating Conventional and Performance-Based Assessments

This debate informs the assessment methods in our project. We have designed *River City* to see if MUVEs enable students to engage in scientific inquiry and make sense of complex data.

The National Research Council defines science inquiry as a multifaceted activity that involves students actively making observations, posing questions, planning and conducting experiments, and communicating results (NRC, 1996). This view of inquiry as an ongoing process involves higher order skills that are not easily measured by multiple-choice tests (Resnick & Resnick, 1992). We want to establish what kind of assessments will allow us to infer that students have learned how to engage in inquiry, particularly at the "front end" of using inquiry processes to make sense out of complexity: problem finding, hypothesis formation, and experimental design.

Because we are still in the early stages of establishing how MUVEs can aid assessment, we are interested in comparing (1) what we learn from traditional assessments, such as pre- and post-measures with (2) what we see in our logfile database and with (3) how students perform on performance-based assessments, such as the "letters to the Mayor."

To date, we have analyzed a random subsample of students' "letters to the mayor" in both the treatment and control groups (n = 224) for evidence of inquiry. In particular, we look for the following:

- Problem identification
- A testable hypothesis based on observations
- An experimental design appropriate to test the hypothesis
- Conclusions based on collected data
- Recommendations

Our scoring rubric based on these items evolved throughout our project, and the current version reflects input from the research team as well as from science teachers. In order to control for coding differences, the team coded the first letters separately and discovered an 80% agreement rate. Coding differences were discussed in order to improve accuracy.

Our preliminary results suggest students are developing an understanding of the process of inquiry that is not well captured in traditional pre/post-test measures (Ketelhut et al, 2005). Overall, students in the *River City* treatments as a group earned "letter-evaluation" scores more than double that of their paper-based control peers, ($p<.01$), a surprising result given the roughly equivalent score for these groups in inquiry gains on the pre/post measures. Further, although we found no differences by type of pedagogical treatment on the inquiry section of our pre/post measures, when we analyzed the letters to the mayor data using multi-level modeling, we discovered interesting differences by treatment.

Table 1 summarizes the results of that analysis. Column 1 of this table lists the letter-evaluation categories that showed differences across the treatments. The treatments with significantly higher scores for that category are denoted with a '✳' in columns 2-5; treatments which had worse scores are denoted with a '—' in those same columns.

Place Table 1 here

As can be seen in Table 1, students in the guided social constructivist treatment had higher scores in nearly every category, whereas students in the control treatment did not do significantly better on any aspect of the letters to the mayor than did the *River City* treatment students. Additionally, the letters written by students in the control curriculum often: were much shorter in length, did not demonstrate motivation or engagement, did not mention the experiment, and did not explicitly recognize the interconnectedness of the chosen problem with other possible causes of the larger problem.

Our point here is not to expound on the differences we found due to our treatments, but to show the richness of learning that is difficult to uncover with multiple choice tests. For example, students who scored low on the science inquiry post-test wrote letters that were of similar quality to those written by students who scored higher on the post-test. As one illustration, in their letters low-performing-content students matched the high-performing-content students around criteria of stating an opinion regarding the cause of the problem and/or the outcome of the experiment. Interestingly, more of the lower-performing test students met the criteria of providing potential interventions or suggesting further research than did students who scored higher on the inquiry test pre/post measures. This suggests that the complexity of the MUVE treatment creates intricate patterns of learning more appropriately measured with a performance-based activity, such as writing an experimental report.

These results, however, are still open to the criticisms levied against performance-based assessments stated above: potential lack of standardization of grading and questionable validity of the task. To address those concerns, we need to connect a student's letter to the mayor to other data that also show evidence of the process of inquiry, such as that in our MUVE database. As we have seen above with the case study of Kimmie, MUVE technologies capture and document students' strategies and allow us to gather a series of observations that sheds light on multiple aspects of a student's knowledge and higher order thinking skills.

## Developing Learning Trajectories for Complex Reasoning Processes

To understand our students' learning processes more fully, we are combining evidence from the logfiles that trace students' movements in the world and matching this to data from our pre-and post-measures and from the letters to the mayor. By connecting and triangulating these

sources of data, we can create rich cases of students' learning and produce evidence of validity. For example, the case of "Audrey/Princess" illustrates the process of a student engaging in learning about inquiry. Audrey [a pseudonym] is a 12-year-old female in the 7th grade whose teacher has below-average expectations of her ability to master science content. (Interestingly, this teacher has higher expectations for some boys in her class with worse academic histories.) In her pre-interview, Audrey indicated that she has low feelings of self-efficacy in science, and before the intervention she scored slightly below average for this sample on the "self-efficacy in scientific inquiry" subscale. She is reading below grade level and scored at the 10% level on the content pre-survey.

Yet, when Audrey enters *River City*, she transforms, taking on the avatar identity of "Princess" and leaving Audrey's damaged sense of self-efficacy behind. She starts out slowly, mostly engaged in organizational issues and exploring the world:

*princess: whose on my team*

*princess: james i have found alot steve u guys go to the wealthy homes me me there ok*

However, as Princess, she quickly becomes engaged in discovering "what is the problem." She wants to figure out why people are getting sick:

*princess: well at the hospital the doors are open and its right near the dump and there are a lot of people in the tentements and they really are sick so yea I think it is the mosiqutos cause they can carry things from the dump*

Audrey works at trying to make sense of the complex data in *River City*, using multiple sources of information:

*princess: It could be the horse poop…welll when the miosqutos are attrected to it the smeell so when the get so of it like taste it or somethinf like that they carry it to the tenements*

*princess: andrea wiggs said( simce they drained the bog, they havent seen any new cases of fever, just like in the winter!*

*princess: I am at the library to see if I can get any information*

On the last day in the world while collecting data about their experiment, one of her teammates thought their hypothesis was wrong, based on an interaction with one of the *River City* residents. Princess quickly investigated the matter by talking to one of the resident (avatar) scientists in the world, Dr. Richards, before drawing the correct conclusion:

*slikyste: whats causing it is that the pipe has lead and if the pipe has lead people would be drinking water and theyd get sick thats what i found out*

*princess: well i dont think that it is the water it was just a hypopthesis its just saying if the pipe was made of of lead. she just said if it was made out of lead she is just teaching her class?*

*princess: so..... we still could be right*

*slikyste: you just heard what he said you*

*jwrb27: yeah shes just showing how to do an experiment*

As can be seen here, the data from the logfiles shows Princess engaging in inquiry and growing throughout the project. Does this match with her survey gain scores and her letter to the mayor? According to the content post-survey, Audrey improved her disease knowledge by 20% and her inquiry skills by 10%. This supports to some degree what we are seeing in the logfiles, but nonetheless is less impressive than her logfiles, especially given her very low

starting score. However, when we coded her letter to the mayor (blindly, as all letters were coded without identification), we discovered that Princess had received the second highest score for her letter to the mayor—clearly, reinforcing the evidence in the logfiles!

## The Challenge of Automating Data Collection from Logfiles

MUVEs provide an emerging, exciting method for studying sophisticated types of learning and instruction under controlled conditions that in real world situations are clouded by the many confounds that inexorably occur in complex, authentic settings. However, this potential power is mitigated by the enormous amounts of data about student performance collected in a MUVE. To conduct the research described in this chapter, extensive analysis "by hand" of student logfiles was required, a laborious and time-consuming process. Thus, achieving the potential of MUVEs for assessment is dependent to some degree on the extent to which one can automate the collection of particular types of data from logfiles, reducing the analytic burden and also enabling real-time feedback for instructional purposes.

Experimenting with this type of automated collection is high on our list of priorities. Some types of logfile data appear relatively easy to aggregate. For example, with well structured logfile formats, writing a computer program to count the number of times a student "talks to" a computer-based agent during a particular MUVE session is not difficult. Determining what type of talk is occurring (e.g., "what's new?" "hello") is also quite feasible. Because of the conversational limits of the agents, this is a far simpler task than the challenges of automated analysis of person-to-person online dialogues (such as verbal interactions among team-mates).

Other types of automated data collection from logfiles are more complex (Baker, Corbett, & Koedinger, 2004; Levy & Wilensky, 2005). For example, is the pattern of a student engaging

in increasingly complex forms of inquiry over the time she spends in the MUVE amenable to collection by a computer program? Although we have not yet attempted this task, this may require sophisticated human judgments beyond what a computer program can accomplish. However, automation might accomplish partial-tasks within this overall effort, such as collecting student utterances that include words suggestive about various stages of inquiry (e.g., "hypothesis" and "because" as a possible example of causal inference).

Beyond conversational analysis, understanding how students move in the world is another form of logfile analysis. We can imagine creating an automated system that produces a map of each student's trajectory while exploring *River City*. Time spent in each location would be indicated by the width of the path. Since one of the advantages of using a graphical MUVE over other forms of technology is that students can actively explore the world, this "map" would allow us to evaluate the impact of this on student learning and engagement. Overall, the degree to which automated data collection could simplify logfile analysis in MUVEs is uncertain, but we believe further research on this topic may result in substantial progress.

## Conclusion

This analysis sketches multiple ways in which the detailed record of student activities and utterances automatically collected in MUVEs offers great potential for student assessment, both from a research perspective and in terms of formative, diagnostic information that could help to tailor instruction to individual needs. Although not discussed extensively in this chapter, MUVEs are also a powerful testbed for theories of learning and teaching, because the designer can shape every aspect of the participant's immersive experience, altering specific variables to conduct research experiments. While realizing this full potential is not an easy task, MUVEs are

likely to add a valuable resource to the spectrum of tools and methods available for assessing student learning.

References

American Association for the Advancement of Science. (1990). *Science for All Americans*. New York, N.Y. Oxford University Press.

Baker, R.S., Corbett, A.T., & Koedinger, K.R. (2004) Detecting Student Misuse of Intelligent Tutoring Systems. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems,* 531-540.

Baxter, G.P. & Shavelson, R.J. (1994). Science Performance Assessments: Benchmarks and surrogates. *International Journal of Educational Research, 21*, 279-297.

Brown, J.S., Collins, A. & Duguid, P. (1989), Situated Cognition and the Culture of Learning, *Education Researcher, 18 (1*), 32-42.

Clarke, J. (2005). *Making Learning Meaningful: An Exploratory Pilot Study of Using Multi User Virtual Environments in Middle School Science*. Unpublished Qualifying Paper, Harvard University Graduate School of Education, Cambridge, MA.

Crusoe, D. (2005). *Citizens' interaction in River City: Possible gender effects within citizen-construct and citizen-citizen conversation data.* Unpublished independent study paper, Harvard University Graduate School of Education, Cambridge.

Dede, C. (2005). Planning for "Neomillennial" Learning Styles: Implications for Investments in Technology and Faculty. In J. Oblinger and D. Oblinger (Eds.), *Educating the Net Generation*, Boulder, CO: EDUCAUSE Publishers. 226-247. Accessed at: http://www.educause.edu/educatingthenetgen/

Dede, C. & Ketelhut, D. (2003, April). *Designing for Motivation and Usability in a Museum-based Multi-User Virtual Environment.* Paper presented at the American Educational Research Association Conference, Chicago.

Dede, C., Ketelhut, D., & Ruess, K. (2002). Motivation, Usability, and Learning Outcomes in a Prototype Museum-based Multi-User Virtual Environment. In P. Bell, R. Stevens, & T. Satwicz (Eds.), *Keeping Learning Complex: The Proceedings of the Fifth International Conference of the Learning Sciences (ICLS).* Mahwah, NJ: Erlbaum.

Dede, C., Clarke, J., Ketelhut, D., Nelson, B., & Bowman, C. (2005). *Fostering Motivation, Learning, and Transfer in Multi-User Virtual Environments*. Paper presented at the American Educational Research Association Conference, Montreal.

Dillashaw, F. G. & Okey, J. R. (1980). Test of integrated process skills for secondary science students. *Science Education, 64*(5), 601-608.

Fraser, B. (1981). *TOSRA: Test of Science Related Attitudes*. Australian Council for Educational Research, Hawthorne, VIC.

Kane, M., Khattri, N., Reeve, A. & Adamson, R. (1997). *Studies of Education Reform: Assessment of Student Performance*. U.S. Education Department's Office of Educational Research and Improvement, Washington, D.C.

Ketelhut, D. (2005, April 4-8). *Assessing Science Self-Efficacy in a Virtual Environment: a Measurement Pilot*. Paper presented at the National Association of Research in Science Teaching Conference, Dallas.

Ketelhut, D. J., Clarke, J., Dede, C., Nelson, B. & Bowman, C. (2005, April 4-8). *Inquiry Teaching for Depth and Coverage via Multi-User Virtual Environments.* Paper presented at the National Association for Research in Science Teaching, Dallas.

Klein, S. P., Jovanovic, J., Stecher, B. M., McCaffrey, D., Shavelson, R. J., Haertel, E., Solano-Flores, G. & Comfort, K. (1997). Gender and racial/ethnic differences on performance assessments in science. *Educational Evaluation and Policy Analysis, 19(2)*, 83-97.

Lave, J. & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. New York, NY: Cambridge University Press.

Levy, S. & Wilensky, U. (2005). *An analysis of patterns of exploration found in logs of students' work with NetLogo models embedded in the Connected Chemistry environment*. Paper presented at the American Educational Research Association Conference, Montreal.

Mehrens, W. (1998). Consequences of Assessment: What is the Evidence? *Education Policy Analysis Archives, 6(13).* Online: http://epaa.asu.edu/epaa/v6n13.html.

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., Gheen, M., Kaplan, A., Kumar, R., Middleton, M. J., Nelson, J., Roeser, R. & Urdan, T. (2000*). Manual for the Patterns of Adaptive Learning Scales (PALS),* Ann Arbor, MI: University of Michigan.

Moore, Wayne. *(2003).* Facts and Assumptions of Assessment: Technology, The Missing Link. *T H E Journal 30 (6).* 20-26.

National Research Council (1996). *National science education standards*. Washington, DC: National Academy Press.

Nelson, B. (2005). *Investigating the Impact of Individualized, Reflective Guidance on Student Learning in an Educational Multi-User Virtual Environment*. Unpublished dissertation, Harvard University.

Partnership for 21st Century Skills. (2003). *Learning for the 21st Century*. Washington, D.C.: Author. Available online at http://www.21stcenturyskills.org

Resnick, L.B. & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. Gifford & M. O'Connor (Eds.), *Changing Assessments: Alternative*

*Views of Aptitude, Achievement, and Instruction.* Norwell, MA: Kluwer Academic Publishers, 37-75.

Stecher, B. M. & Klein, S. P. (1997). The cost of science performance assessments in large-scale testing programs. *Educational Evaluation and Policy Analysis, 19(1),* 1-14.

Vygotsky, L. (1978). *Mind in Society*. London: Harvard University Press.

FIGURES

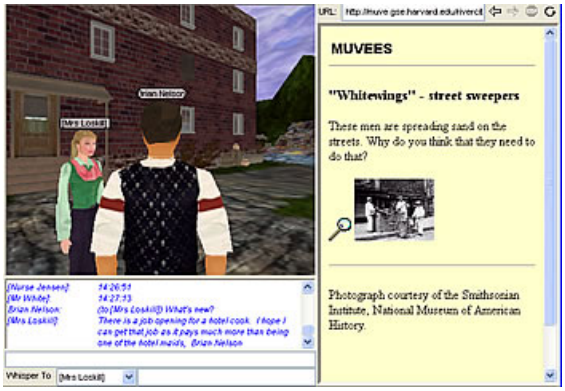


Figure 1: Avatar talking with agent;

Digital artifact displayed in right window.

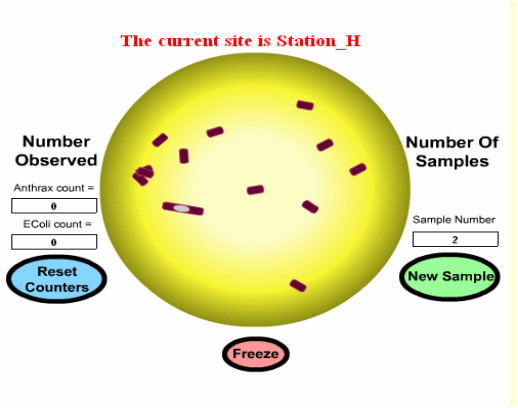


Figure 2: Tool for analyzing water data

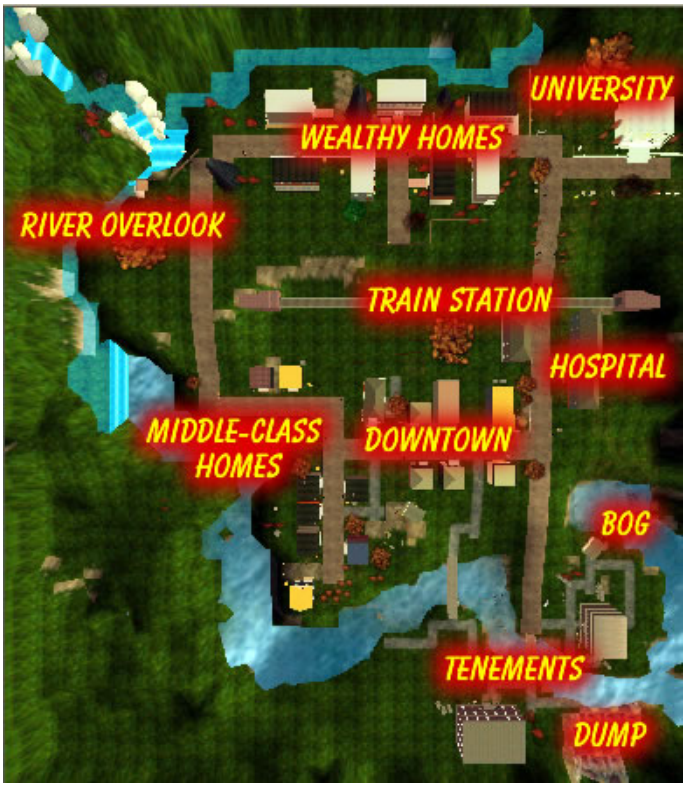


Figure 3: The *River City* layout

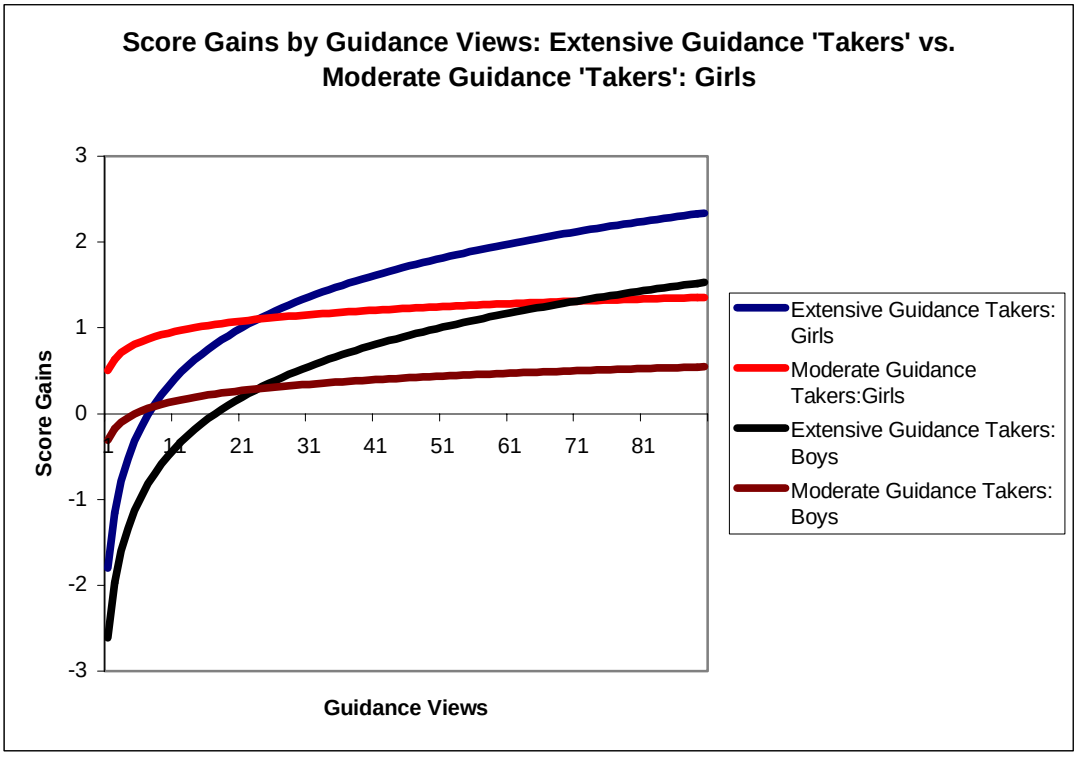


Figure 4: The fitted relationship between content test score gains and levels of guidance system use by students exposed to extensive or moderate levels of guidance who chose to "take up" the guidance at least one time in a MUVE-based curriculum, by gender. (n=272)

TABLES

Table 1. Coded sections of the "letters to the mayor" that showed significant differences ($p<.05$) by treatment in student scores relative to one or more of the other treatments (n=173) with ✳ indicating the treatment which had the highest scores.

| **Areas that differed significantly by treatment ($p<.05$)** | **GSC** | **EMC** | **LPP** | **Control** |
|---|---|---|---|---|
| Overall quality | ✳ | — | | — |
| Summarizing the problem | | | ✳ | — |
| Awareness that different symptoms were related to different diseases | ✳ | — | ✳ | — |
| Stating a testable hypothesis | ✳ | — | ✳ | |
| Collecting evidence to test hypothesis | ✳ | — | — | — |
| Understanding the vector of disease transmission | ✳ | ✳ | — | |
| Stating a conclusion | ✳ | | | — |

*Key:* ✳ = Treatment that on average had highest scores in this category

— = Treatments that on average had worse scores in this category relative to ✳ treatments

= Treatments that on average were not significantly different from the others in this category

GSC = Guided social constructivist treatment.

EMC = Expert modeling and coaching treatment.

LPP = Legitimate peripheral participation treatment.