EDUCATIONAL DESIGNER
JOURNAL OF THE INTERNATIONAL SOCIETY FOR DESIGN AND DEVELOPMENT IN EDUCATION

Opportunities and Challenges in Embedding Diagnostic Assessments into Immersive Interfaces

Chris Dede
Harvard Graduate School of Education

Introduction

Note: This paper is excerpted and extended from a longer White Paper commissioned by the Educational Testing Service: (Dede, C. 2012b).

Some types of technology-based learning-by-doing environments are tightly structured in several respects. The phenomenon to be studied is clearly defined, without extraneous or unrelated information clouding the situation. Also, at each decision point, the range of possible actions is limited to a few alternatives, all directly related to the phenomenon rather than possibly off-task. Often, the understandings and performances to be learned are relatively simple and can be encapsulated in factual statements and procedural recipes. These tightly structured characteristics typically apply to nonimmersive simulations (Quellmalz & Haertel, 2004) and to tutoring systems (Koedinger & Corbett, 2006; VanLehn, 2006). Many computer-based games also fit this description.

In contrast, internships in workplace contexts, immersive simulations, field trips to real-world settings, and some computer-based games (e.g., role-playing adventure games set in virtual environments, such as Quest Atlantis (Barab et al., 2010) are more open-ended learning experiences. Students encounter many phenomena, only some of which are related to what they seek to learn; and participants' range of possible actions at any given point is quite broad. The understandings to be developed are complex (e.g., problem finding in unstructured situations) and nonformulaic.

In tightly structured simulations, well-understood methods are available for collecting data about student behaviors, analyzing these performances to determine what learners understand and do not understand at a particular stage of instruction, and providing individualized feedback to students and teachers that is formative for next steps in learning and instruction (Brown, Hinze, & Pellegrino, 2008). Tutoring systems and related assistance environments also

have mature methods for collecting data about learner actions, interpreting this information, and providing diagnostic feedback (Heffernan, Heffernan, Decoteau, & Militello, 2012). Other technology enhanced assessments white-papers (Behrens, DiCerbo, & Ferrara, 2012; Heffernan & Koedinger, 2012; Levy, 2012) discussed ways in which embedding diagnostic assessments in these tightly structured learning environments is complicated and challenging. However, well-understood and mature strategies are available for formatively improving learning and instruction for the types of educational content and outcomes that these systems can support. In contrast, this paper focuses on less well understood, immature methods for the more problematic challenge of embedding diagnostic assessments in open-ended learning experiences created by immersive interfaces.

Immersive Interfaces

Immersion is "being there," the subjective impression that one is in a real place (Slator, 2009, p. 3549). For example, a well-designed movie draws viewers into the world portrayed on the screen, and they feel caught up in that virtual environment. Technologies can induce immersion via the sensory stimuli, participants' abilities to influence what happens in the environment, and the use of narrative and symbolism (Dawley & Dede, in press). Two types of immersive interfaces underlie a growing number of formal and informal learning experiences (Dede, 2009):

- Multiuser virtual environment (MUVE) interfaces offer students an engaging "Alice in Wonderland" experience in which their digital avatars in a graphical, virtual context actively participate in experiences with the avatars of other participants and with computerized agents. MUVEs provide rich environments in which participants interact with digital objects and tools, such as historical photographs or virtual microscopes (Ketelhut et al., 2010).

- Augmented reality (AR) interfaces enable *ubiquitous computing* models. Students carrying mobile wireless devices through real-world contexts interact with virtual information, visualizations, and simulations superimposed on physical landscapes (such as a tree describing its botanical characteristics, a historic photograph offering a contrast with the present scene, or a cloaked alien spaceship visible only through the mobile device). This type of immersion infuses digital resources throughout the real world, augmenting students' experiences and interactions (Klopfer, 2008).

The military and the entertainment industry have expended substantial resources in developing these immersive media, which have many applications in precollege and higher education, as well as for training and professional development. My research team's current work in immersive interfaces for learning centers on EcoMUVE (http://ecomuve.gse.harvard.edu), a middle school science curriculum utilizing two virtual ecosystems, and EcoMOBILE (http://ecomobile.gse.harvard.edu), an augmented reality approach to field trips in real ecosystems (Dede et al, in press).

Immersive interfaces offer unique potential for interwoven assessment because of three Es: engagement, evocation, and evidence. In our research, we find that immersive authentic simulations are very engaging for students, even though we deliberately avoid some types of motivation common in games (Dede, 2009). Students try hard to succeed, and the learning experiences promote their self-efficacy as a scientist. This means that, unlike many types of testing, all students are putting forth their best efforts. Further, immersive interfaces can evoke a wide spectrum of performances.. This means that a very broad palette of learning experiences—and assessment situations by which a mentor would assess progress towards mastery—can create opportunities for students to reveal their degrees of engagement, self-efficacy, understandings, and performances.

Drawing on a broad spectrum of performances is important in determining the true extent of what a student knows and does not know. For a decade, we developed and studied the River City curriculum (http://muve.gse.harvard.edu/rivercityproject/); middle grades students learned epidemiology and biological principles, as well as collaborative scientific inquiry, by traveling back about 130 years to an immersive virtual town plagued by diseases (Ketelhut et al., 2010). In our detailed analysis of students' ongoing activities and interactions, we found evidence of learning that was not captured by pre/post-tests or by a scientific-conference presentation student teams gave at the end of the unit (Ketelhut, Dede, Clarke, Nelson, & Bowman, 2007). The evidentiary trail of learning trajectories afforded by interwoven diagnostic assessments is richer and often more valid than a snapshot summative measure, even a rich artifact like a synthesis presentation.

Finally, immersive interfaces can collect an impressive array of evidence about what a learner knows (and does not know), what he or she can do (and cannot do), and whether he or she knows when and how to apply disciplinary frames and prior knowledge to a novel problem. Immersive environments—because of their situated nature and because they generate log files—make it easy to design for eliciting performances, to collect continuous data, and to interpret structures of evidence. In a virtual world, the server documents and timestamps actions by

each student: movements, interactions, utterances, saved data, and so on. In an AR, the mobile device can save moderately detailed information about movements and actions, and using the device to record learners' voices as their team interacts could provide another resource for analysis. Given engagement, evocation, and evidence, immersive learning interfaces potentially are the most powerful and valid assessment medium available—but can we realize this potential?

Difficulties of interwoven diagnostic assessment in immersive interfaces

Quellmalz, Timms, and Schneider (2009) examined issues of embedding assessments into games and simulations in science education. Their analysis included both tightly-structured and open-ended learning experiences. After studying several immersive games and simulations related to learning science, including River City, they noted that the complex tasks in simulations and games cannot be adequately modeled using only classical test theory and item response theory. This shortfall arises because these complex tasks have four characteristics (Williamson, Bejar, & Mislevy, 2006). First, completion of the task requires the student to undergo multiple, nontrivial, domain-relevant steps and/or cognitive processes. Second, multiple elements, or features, of each task performance are captured and considered in the determination of summaries of ability and/or diagnostic feedback. Third, the data vectors for each task have a high degree of potential variability, reflecting relatively unconstrained work product production. Fourth and finally, evaluation of the adequacy of task solutions requires the task features to be considered as an interdependent set, for which assumptions of conditional independence typically do not hold.

Quellmalz et al. (2009) concluded that, given the challenges of complex tasks, more appropriate measurement models for simulations and games—particularly those that are open-ended—include Bayes nets, artificial neural networks, and model tracing. They added that new psychometric methods beyond these will likely be needed. Beal and Stevens (2007) used various types of probabilistic models in studying students' performance in simulations of scientific problem solving. Bennett, Persky, Weiss, and Jenkins (2010) described both progress in applying probabilistic models and the very difficult challenges involved. Behrens, Frezzo, Mislevy, Kroopnick, and Wise (2007) described ways of embedding assessments into structured simulations; and Shute, Ventura, Bauer, and Zapata-Rivera (2009) delineated a framework for incorporating stealth assessments into games.

Not surprisingly given these analytic difficulties and the open-ended nature of immersive learning environments compared to structured simulations and

games, we encountered many challenges in attempting to understand students' progress during the course of the River City curriculum (Ketelhut et al., 2007). River City was deliberately designed as a very open-ended learning environment in which many paths to success were available. Further, resolving the town's issues with illness required students first to infer that three different diseases were simultaneously present and then to shift their activities to studying one of these diseases. While scientifically valid and important as a learning experience in complex inquiry, this complicated situation necessitated relatively unfocused data gathering by students until the realization of superimposed illnesses was reached.

As a result of all these factors in River City, interpreting students' detailed actions to understand their intent—and what level of ongoing performance they actually achieved—was very difficult, although Clarke (2009)developed analytic methods to accomplish this. For each student, by combining records of movements, interactions with the world, chat-logs with team members, and artifacts produced, we were able to formulate case studies documenting individual learning trajectories. However, these case studies required intensive human effort and expertise, and the methods used are impractical and unscalable for real-time diagnosis formative for instruction. SAVE Science (http://www.savescience.net/) is a research project currently studying the issues of data-mining records of student actions in an immersive virtual environment for learning science inquiry (Ketelhut et al., 2012).

The virtual performance assessment project

To better understand the assessment challenges involved with the MUVE interface, in 2008 my colleagues and I began our Virtual Performance Assessment (VPA) project, funded by the Institute for Education Sciences of the U.S. Department of Education, as well as the Gates Foundation. We are developing and studying the feasibility of immersive virtual performance assessments to assess the scientific inquiry skills of middle grades students as a standardized component of an accountability program (http://vpa.gse.harvard.edu). The goal is to provide states with reliable and valid technology-based performance assessments linked to state and national science education standards for inquiry processes (Clarke-Midura, Dede, & Norton, 2011).

Applying the evidence-centered design (ECD) approach (Mislevy & Haertel, 2006; Mislevy & Rahman, 2009) allowed us to articulate every aspect of the VPAs, from the knowledge, skills, and abilities (KSAs) measured to the types of evidence that allow making claims about what students know and do not know. Using the principled assessment designs for inquiry (PADI) system has enabled

us to create multiple forms of the same assessment, for a generalizability study, and to reframe science inquiry constructs (theorizing, questioning and hypothesizing, investigating, analyzing and synthesizing) into specific KSAs aligned with current national standards (Clarke-Midura, Mayrath, & Dede, in press). Clarke-Midura is currently studying the extent to which the attribute hierarchy method (Wang & Gierl, 2011) and Bayesian network models (Mislevy, Steinberg, & Almond, 2003) provide measurement models suitable for the types of data generated from virtual performance assessments.

The VPA work has clarified how to design ECD-based summative assessments of sophisticated cognitive performances in virtual worlds. We have established that this type of assessment is practical and affordable at scale, as well as more valid for sophisticated performances like scientific inquiry than paper-and-pencil, item-based assessments (Clarke & Dede, 2010). While we are still determining the psychometric properties of our virtual performance assessments, we believe this medium has great potential to complement more standard forms of testing. However, our research has also revealed that the design of a virtual world as a summative assessment results in an environment necessarily too narrow in its structure to allow many of the powerful forms of open-ended learning described above.

All these issues lead to the conclusion that the difficulties of embedding diagnostic assessments in immersive simulations, as well as in other open-ended learning environments with complex tasks, cannot be fully resolved—at least in the near-term—through more sophisticated analytic techniques. New types of design strategies are also needed to create immersive authentic simulations that have interwoven aspects amenable to diagnostic measurement and real-time formative intervention.

Design Strategies for Ameliorating Analytic Difficulties in Interwoven Assessments

This section describes design strategies that can preserve the open-ended learning that immersive interfaces empower, while at the same time enabling tractable real-time analysis of data diagnostic of students' understandings. An important constraint in these designs is that, as much as possible, the assessment activities and data collection must be unobtrusive. Otherwise, the assessment dimension of the experience disrupts immersion and engagement through undercutting flow and authenticity, which in turn can undermine learning (National Research Council, 2011). These design strategies are illustrated with examples from River City and with potential applications to EcoMUVE and EcoMOBILE.

Paths and heat maps

The paths that a student takes in exploring a virtual world to determine the contextual situation, identify anomalies, and collect data related to a hypothesis for the causes of an anomaly are an important predictor of the student's understanding of scientific inquiry. In the River City curriculum, participants inside the multi-user virtual environment explore various places in the town (Figure 1) and collect data on changes over time, acting in gradually more purposeful ways as they develop and test hypotheses (Dede, 2009). Students help each other and also find experts and archives to guide them (Figure 2). Further, learners use virtual scientific instruments, such as microscopes to test water for bacteria (Figure 3 and Figure 4).

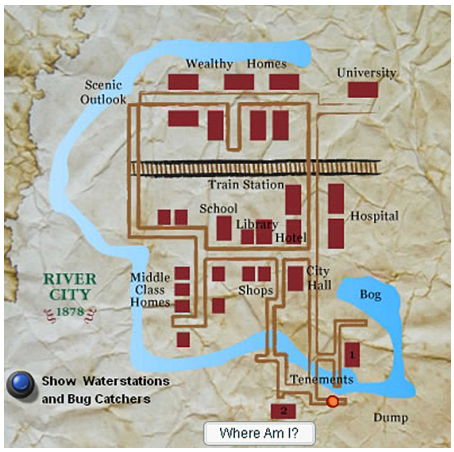


Figure 1: Map of River City

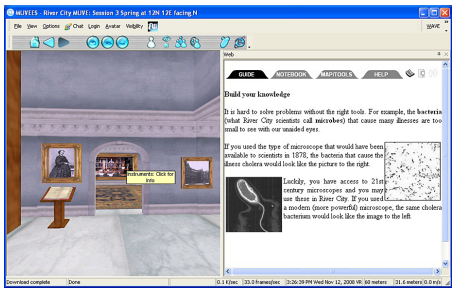


Figure 2: View of 3-D environment and web-based content on right side of screen

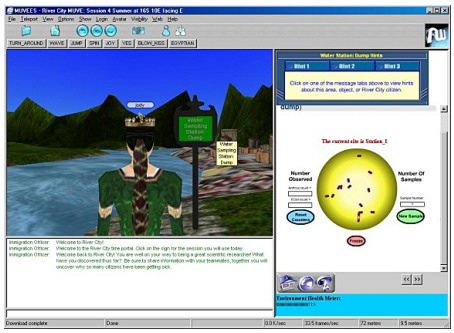


Figure 3: Taking a water sample with the virtual microscope.

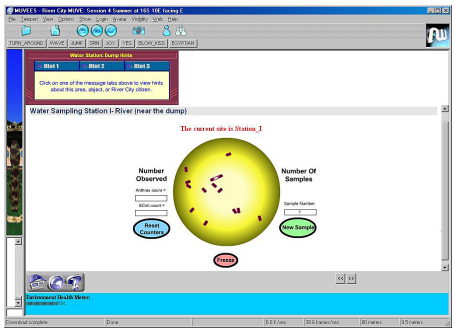


Figure 4: Close up of Microscope. Students click "Freeze" and count the number of EColi and Anthrax in the water

This immersive simulation allowed them to conduct an experiment by changing an independent variable they select, then collecting data in the city to test their hypothesis. Students not only hypothesize what would happen if, for example, a sanitation system were built—they can actually visit the simulated city with a sanitation system added and see how this change affects the patterns of illness.

In analyzing learning in River City, we used log file data to generate event paths (Figure 5) for both individual students and their three person teams. Students and teachers found this a useful source of diagnostic feedback on the relative exploratory skills—and degree of team collaboration—that these performances exhibited. Dukas (2009) extended this research by developing an avatar log visualizer (ALV), which generates a series of slides depicting the relative frequency events of one or more subpopulations of students, aggregated by user-specified location and time bins. Figure 6 displays an ALV visualization that contrasts the search strategies of the high-performing and low-performing students in a class, displaying the top 10 scores on the content post-test (in green) and the lowest 10 scores (in pink).

Figure 5: Event paths in River City Session 3 for a three-person team

Figure 6: An avatar log visualizater (AVL) heat map showing high-performing and low-performing students in River City Session 4

The high performing students' preferred locations provide an expert model usable in diagnostic feedback, formative about their search strategies, to students in subsequent classes. The low performing students' locations may offer insights into what types of understanding they lack. Overall, path analysis is a potentially powerful form of unobtrusive assessment, although choosing the best way to display student paths through a learning environment is a complex type of visualization not well understood at present (Dukas, 2009). The utility of this diagnostic approach also depends on the degree to which exploration in the virtual world is an important component of learning.

Accessing an individualized guidance system

As another example of the rich analytic power possible through the use of log files, Nelson (2007) developed a version of River City that contained an interwoven individualized guidance system (IGS). The guidance system utilized personalized interaction histories collected on each student's activities to generate real-time, customized support. The IGS offered 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 (Figure 7).

Figure 7: Individualized guidance system (IGS) interface

As an example, if a student were to click on the admissions chart in the River City hospital, 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 or she had previously visited the tenement district, and asking the student how many patients listed on the chart came from that part of town.

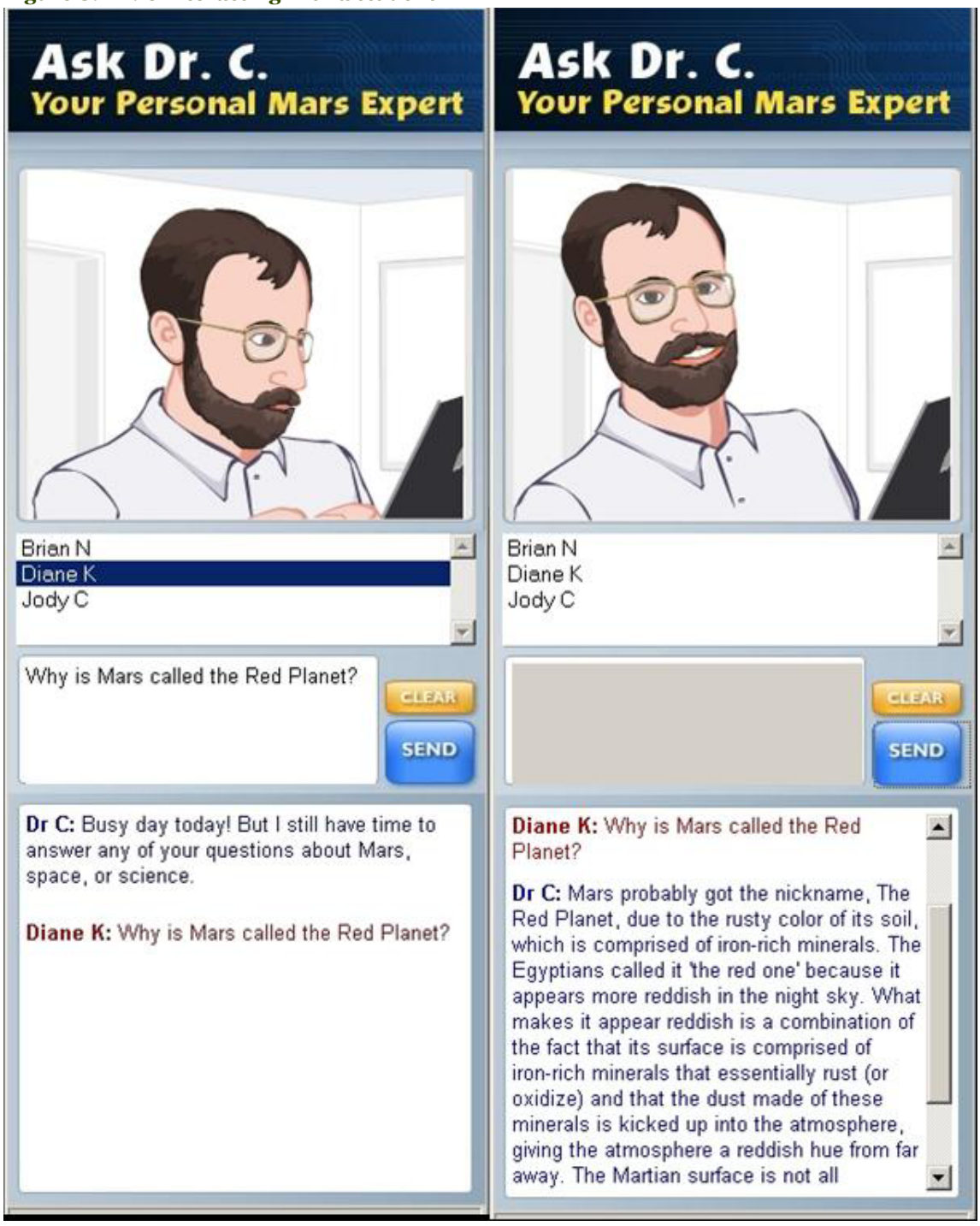
Multilevel multiple regression analysis findings showed that use of this guidance system with our MUVE-based curriculum had a statistically significant, positive impact ($p < .05$) on student learning (Nelson, 2007). In addition to using the log files to personalize the guidance provided to each student, we conducted analyses of guidance use. We knew when and if students first chose to use the guidance system, which messages they viewed, where they were in the virtual world when they viewed them, and what actions they took subsequent to viewing a given guidance message. This potentially provides diagnostic information that could guide instruction.

Asking questions of an "expert" agent

Animated pedagogical agents (APAs) are "lifelike autonomous characters that co-habit learning environments with students to create rich, face-to-face learning interactions" (Johnson, Rickel, & Lester, 2000, p. 47). Bowman (2011) created Dr. C (Figure 8), an APA that served as a mentor for middle school students studying space science. Interwoven in a web browser, Dr. C was a computerized simulation of the primary scientist involved with NASA's Mars Student Imaging Project. One version of Dr. C provided both career (content-

focused) and psychosocial (interpersonal-focused) mentoring; the latter meant Dr. C would answer some questions about the scientist's personal and professional life. The back end of the system was a large set of short responses to frequently asked questions, with a relatively simplistic pattern recognition system scanning the text of the student's question for words it recognized. A study revealed that the Dr. C template for expert mentoring was flexible, reliable, and engaging, distributing substantial amounts of content knowledge on an as-needed basis (Bowman, 2011).

Figure 8: Dr. C interacting with a student



Research suggests that APAs can fill various roles of mentorship, including expert, motivator, collaborator, and learning companion (Chou, Chan, & Lin, 2003). For example, Baylor and Kim (2005) created three versions of an APA: the Expert, designed as older than the participants, formal in appearance and language, and providing domain-specific information; the Motivator, casual in appearance and language, providing encouragement; and the Mentor, less formal than the Expert yet older than the Motivator, providing a mix of information and encouragement. The results from their study confirmed that the agent (APA) roles were not only perceived by the students to reflect their intended purposes, but also led to significant changes in learning and motivation as intended by their design. Specifically, the Expert agent (APA) led to increased information acquisition, the Motivator led to increased self-efficacy, and the Mentor led to overall improved learning and motivation (Baylor & Kim, 2005).

One can imagine tailoring a wide range of APAs to various student needs and embedding these in immersive learning environments (Dede, 2012). Beyond engaging students and providing a limited form of mentoring, APAs have advantages for interwoven diagnostic assessment in immersive authentic simulations in two respects: First, the questions students ask of an APA are themselves diagnostic—typically learners will ask for information they do not know, but see as having value. Sometimes a single question asked by a student of an APA may reveal as much about what that learner does and does not know than a series of answers the student provides to a teacher's diagnostic questions. Both EcoMUVE and EcoMOBILE could embed APAs of various types for eliciting a trajectory over time of questions that reveal aspects of students' understanding and motivation, as well as aiding learning and engagement by the APA's responses.

Second, APAs scattered through an immersive authentic simulation can draw out student performances in various ways. In a virtual world, a student could meet an APA who requests the student's name and role. Even a simple pattern recognition system could determine if the student made a response indicating self-efficacy and motivation ("scientist" or some variant) versus a response indicating lack of confidence or engagement ("sixth grader" or some other out-of-character reply). As another example, an APA could request a student to summarize what the student has found so far, and some form of latent semantic analysis could scan the response for key phrases indicating understanding of terminology and relevant concepts. The important design considerations of this method for evoking performances are that (a) the interaction is consistent with the overall narrative, so not too disruptive of flow, (b) the measurement is relatively unobtrusive, and (c) the interactions themselves deepen immersion.

Structured benchmark assessments that measure learning progress and scaffold transfer

Periodically, designers could structure brief benchmarking episodes into the immersive learning experience. In a virtual world, these might be the equivalent of the VPA assessments: shifting a student to a semantically identical but syntactically dissimilar environment in which the student must identify a problem (earlier in the curriculum) or resolve a problem (later in the curriculum) by performances based on current understandings. As an example, one of our VPA assessments (Clarke-Midura, Dede, & Norton, 2011) involved a student determining which among several environmental factors was causing the depletion of a kelp forest in an Alaskan bay (Figure 9, Figure 10).

The ECD of this structured assessment would enable benchmarking to document a student's progress, without undercutting the open-ended nature of the overall learning experience. As another benefit, these benchmarking assessments could help to scaffold transfer. Further, by centering on learners' common misconceptions about that domain (e.g., in the case of ecosystems, action across distance and time, invisible causes, the flows of matter and of energy), immediately conveying the results of these benchmarks to students could prompt "aha" moments that help to synthesize new levels of understanding.

Figure 9: Avatar exploring simulated Alaskan bay

Figure 10: Documenting data and hypotheses

Reporting mechanisms consistent with the immersive simulation's narrative and flow.

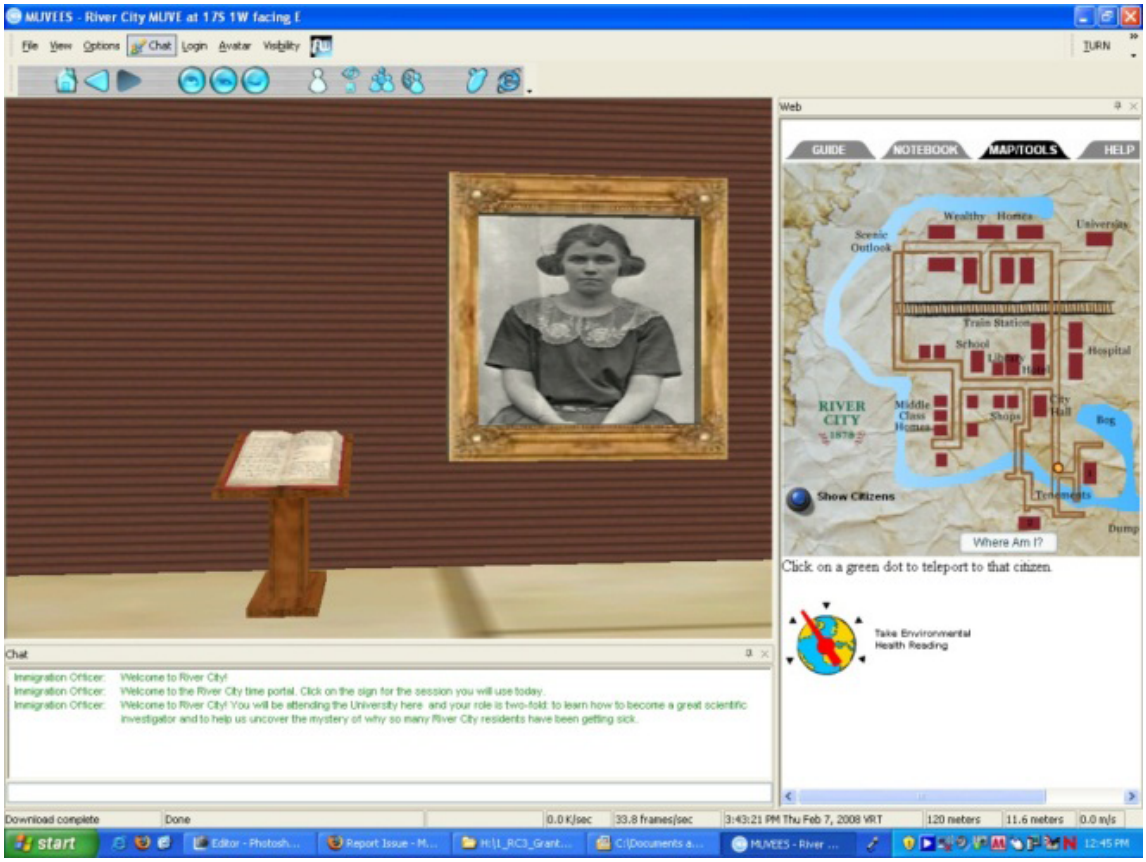
Game designers have developed a variety of reporting mechanisms that provide diagnostic feedback to participants, formative to improved player strategies, without disrupting flow or narrative (Gee, 2008). These include *dashboards* of various types that provide ongoing progress levels for key variables (e.g., health and stamina in an adventure game). Beyond providing ongoing feedback, this design strategy increases immersion and engagement—and learning, if the variables tracked correspond closely to instructional goals.

In River City, we went beyond dashboards to give students an opportunity parallel to leveling up in games: On each visit to River City, teams of students could attain new powers through reaching a threshold of experiences and accomplishments (Nelson et al., 2007). We centered our powers narrative on a specific location inside the world, a spooky building initially closed to all students (Figure 11). When a team of students achieved powers for a given session, they gained access to a new room in the house where they found a number of magical tools (Figure 12), such as a special interactive map that allowed students to check on the health of all residents of the city.

The powers design approach provides a way of motivating a team's collective success on a variety of interwoven assessments, then uses that achievement to promote further learning with added opportunities for diagnostic inferences.

Figure 11: "Powers" Mansion

Figure 12: "Magical" Health Checker



Next Steps in Attaining Immersive Assessments: Recommendations for the Field

The U.S. National Research Council (2011) report on learning science through computer games and simulations made several recommendations on a research agenda (pp. 126–127):

- Applications of the ECD approach to the development of assessments of learning through simulations and games. Developers and testing experts should collaborate to clearly identify desired learning goals and the kinds of evidence needed to show learner progress toward these goals; they should use these specifications to design tasks and test items in ways that will provide the needed evidence. Modeling of the motivation and thinking of the learner will need to evolve simultaneously with the "physical" modeling of the game or simulation.

  In addition to ECD, I would include design for ameliorating the challenges of interwoven assessment in open-ended learning environments.

- The development and use of flexible statistical models and machine learning to make meaning from the large amounts of data provided by simulations and games. These measurement methods are well suited to application in simulations and games because they can handle uncertainty about the current state of the learner, provide immediate feedback during tasks, and model complex patterns of student behavior and multiple forms of evidence. Continued research on these methods will help to improve assessment in simulations and games.

  In addition to new types of mathematical models, I would include embedded measurement models analogous to those utilized by mentors in internship or apprenticeship settings.

- Researchers should continue to advance the design and use of techniques that (a) rapidly measure and adapt to students' progress in a specific learning progression, (b) dynamically respond to an individual student's performance, and (c) allow for the summative evaluation of how well students are learning.

A culminating achievement of this type of research would be to alter, in real time, the context and activities of the immersive simulation to make salient what the student needs to understand next in their learning trajectory.

Collectively, these research achievements could help to scaffold two Grand Challenges in the 2010 National Educational Technology Plan (U.S. Department of Education, 2010, p. 78):

1.0: Design and validate an integrated system that provides real-time access to learning experiences tuned to the levels of difficulty and assistance that optimize learning for all learners and that incorporates self-improving features that enable it to become increasingly effective through interaction with learners.

2.0: Design and validate an integrated system for designing and implementing valid, reliable, and cost-effective assessments of complex aspects of 21st-century expertise and competencies across academic disciplines.

These in turn could empower powerful digital teaching platforms that enable customizing classroom learning for each student (Dede & Richards, 2012).

What does all this mean for designers? Working closely with researchers to articulate and understand the interconnections between design decisions and consequences for understanding engagement and learning is very important. Developing evidence- and theory-based design heuristics for combining emerging types of media (e.g., immersive environments, social media) is also crucial. By accomplishing these two goals as a field, we can empower better assessment and research proving the effectiveness of our designs.

Acknowledgments

I have greatly benefited from the insights of my colleagues Dr. Tina Grotzer, Dr. Amy Kamarainen, Dr. Shari Metcalf, and Shane Tutwiler from the EcoMUVE/EcoMOBILE projects, supported by the U.S. Department of Education's Institute for Education Sciences (IES), Qualcomm, Texas Instruments, and the U.S. National Science Foundation (NSF), as well as from the insights of Dr. Jody Clarke-Midura, Director of the Virtual Performance Assessment Project funded by IES and the Gates Foundation. In addition, I appreciate the feedback of Drs. Cassie Bowman, Ed Dieterle, Brian Nelson, and Diane Jass Ketelhut, all members of the River City project team, which was supported by NSF. The viewpoints expressed in this article are mine and are not necessarily those of my collaborators or the funders.

References

Barab, S. A., Gresalfi, M., & Ingram-Goble, A. (2010). Transformational play: Using games to position person, content, and context. *Educational Researcher,* 39(7), 525–536.

Baylor, A. L., & Kim, Y. (2005). Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education,* 15, 95–115.

Beal, C. R., & Stevens, R. H. (2007). Student motivation and performance in scientific problem solving. In R. Luckin, K. R. Koedinger, & J. Greer (Eds.), *Artificial intelligence in education: Building technology rich learning contexts that work* (pp. 539–541). Amsterdam, Netherlands: IOS Press.

Behrens, J. T., Frezzo, D., Mislevy, R., Kroopnick, M., & Wise, D. (2007). Structural, functional, and semiotic symmetries in simulation-based games and assessments. In E. Baker, J. Dickieson, W. Wulfeck, & H. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 59–80). Mahwah, NJ: Lawrence Erlbaum Associates.

Behrens, J. T., DiCerbo, K. E., & Ferrara, S. (2012). *Intended and unintended deceptions in the use of simulations*. Princeton, NJ: ETS.

Bennett, R. E., Persky, H., Weiss, A., & Jenkins, F. (2010). Measuring problem solving with technology: A demonstration study for NAEP. *Journal of Technology, Learning, and Assessment,* 8(8), 1–45.

Bowman, C. D. D. (2011). Student use of animated pedagogical agents in a middle school science inquiry program. *British Journal of Educational Technology,* 43(3). Retrieved from http://onlinelibrary.wiley.com /doi/10.1111/j.1467-8535.2011.01198.x/pdf

Brown, J., Hinze, S., & Pellegrino, J.W. (2008). Technology and formative assessment. In T. Good (Ed.), *21st century education*. Thousand Oaks, CA: Sage.

Chou, C., Chan, T., & Lin, C. (2003). Redefining the learning companion: The past, present, and future of educational agents. *Computers and Education,* 40, 255–269.

Clarke, J. (2009). *Exploring the complexity of inquiry learning in an open-ended problem space* (Unpublished doctoral dissertation). Harvard Graduate School of Education, Cambridge, MA.

Clarke, J., & Dede, C. (2010). Assessment, technology, and change. *Journal of Research on Technology in Education,* 42(3), 309–328.

Clarke-Midura, J., Dede, C., & Norton, J. (2011). Next generation assessments for measuring complex learning in science. In *The Road Ahead for State Assessments* (pp. 27–40). Cambridge MA: Rennie Center for Education and Public Policy. Retrieved from http://renniecenter.issuelab.org /research

Clarke-Midura, J, Mayrath, M., & Dede, C. (in press). Thinking outside the bubble: Virtual performance assessments for measuring complex learning. In M. Mayrath, J. Clarke-Midura, & D. H. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*. Charlotte, NC: Information Age Publishing.

Dawley, L., & Dede, C. (in press). Situated learning in virtual worlds and immersive simulations. In M. J. Bishop & J. Elen (Eds.), *Handbook of research on educational communications and technology* (Vol. 2., 4th ed.). New York, NY: Macmillan.

Dede, C. (2009). *Immersive interfaces for engagement and learning*. *Science, 323(5910)* , 66–69.

Dede, C. (2012). Customization in immersive learning environments: Implications for digital teaching platforms. In C. Dede & J. Richards (Eds.). *Digital teaching platforms: Customizing classroom learning for each student* (pp. 119–133). New York, NY: Teacher's College Press.

Dede, C., Grotzer, T., Kamarainen, A., Metcalf, S., & Tutwiler, M.S. (in press). EcoMobile: Blending virtual and augmented realities for learning ecosystems science and complex causality. *Journal of Immersive Education*

Dede, C., & Richards, J. (Eds.). (2012). *Digital teaching platforms: Customizing classroom learning for each student*. New York, NY: Teacher's College Press.

Dede, C. (2012b). *Interweaving assessments into immersive authentic simulations: Design strategies for diagnostic and instructional insights* (Commissioned White Paper for the ETS Invitational Research Symposium on Technology Enhanced Assessments). Princeton, NJ: Educational Testing Service. http://www.k12center.org/rsc/pdf/session4-dede-paper-tea2012.pdf

Dukas, G. (2009) *Characterizing student navigation in educational multiuser virtual environments: A case study using data from the River City project* (Unpublished doctoral dissertation). Harvard Graduate School of Education, Cambridge, MA.

Gee, J. P. (2008). Learning and games. In K. Salen (Ed.), *The ecology of games: Connecting youth, games, and learning* (pp. 21–40). Cambridge, MA: MIT Press.

Heffernan, N. T., Heffernan, C. L., Decoteau, M., & Militello, M. (2012). Effective and meaningful use of educational technology: Three cases from the classroom. In C. Dede & J. Richards (Eds.), *Digital teaching platforms: Customizing classroom learning for each student* (pp. 88–102). New York, NY: Teachers College Press.

Heffernan, N. T., & Koedinger, K. R. (2012). *Integrating assessment within instruction: A look forward*. Princeton, NJ: ETS.

Johnson, W. L., Rickel, J. W., & Lester, J. C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education,* 11, 47–78.

Ketelhut, D. J., Avirup, S., Yates, A., Shelton, A., Natarajan, U., Nelson, B., Schifter, C., & Karakus, M. (2012). *Insights into science learning using immersive environments as assessments: Data mining SAVE science*. Manuscript submitted for publication.

Ketelhut, D., Dede, C., Clarke, J., Nelson, B., & Bowman, C. (2007).Studying situated learning in a multi-user virtual environment. In E. Baker, J. Dickieson, W. Wulfeck, & H. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 37–58). Mahwah, NJ: Erlbaum.

Ketelhut, D. J., Nelson, B. C., Clarke, J. E., & Dede, C. (2010). A multi-user virtual environment for building and assessing higher order inquiry skills in science. *British Journal of Educational Technology,* 41, 56–68.

Klopfer, E. (2008). *Augmented learning: Research and design of mobile educational games*. Cambridge, MA: MIT Press.

Koedinger, K. R., & Corbett, A. T. (2006). Cognitive tutors: Technology bringing learning science to the classroom. In K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 61–78). New York, NY: Cambridge University Press.

Levy, R. (2012). *Psychometric advances, opportunities, and challenges for simulation-based assessment*. Princeton, NJ: ETS.

Mislevy, R., & Haertel, G. (2006). *Implications of evidence-centered design for educational testing* (Draft PADI Technical Report 17). Menlo Park, CA: SRI International.

Mislevy, R., & Rahman, T. (2009). *Design pattern for assessing cause and effect reasoning in reading comprehension* (PADI Technical Report 20). Menlo Park, CA: SRI International.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). *On the structure of educational assessments*. *Measurement: Interdisciplinary Research and Perspectives,* 1, 3–62.

National Research Council. (2011). *Learning science through computer games and simulations*. Washington, DC: National Academies Press.

Nelson, B. (2007). *Exploring the use of individualized, reflective guidance in an educational multi-user virtual environment*. *Journal of Science Education and Technology* 16(1), 83–97.

Nelson, B., Ketelhut, D. J., Clarke, J., Dieterle, E., Dede, C., & Erlandson, B. (2007). Robust design strategies for scaling educational innovations: The River City MUVE case study. In B. E. Shelton & D. A. Wiley, *The design and use of simulation computer games in education* (pp. 219–242). Rotterdam, Netherlands: Sense Press.

Quellmalz, E. S., & Haertel, G. (2004). *Technology supports for state science assessment systems.* Paper commissioned by the National Research Council Committee on Test Design for K-12 Science Achievement. Washington, DC: National Research Council.

Quellmalz, E. S., Timms, M. J., & Schneider, S. A. (2009). *Assessment of student learning in science simulations and games*. Paper prepared for the National Research Council Workshop on Gaming and Simulations. Washington, DC: National Research Council.

Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. J. Cody, & P. Vorderer (Eds.), *The social science of serious games: Theories and applications* (pp. 295–321). Philadelphia, PA: Routledge.

Slator, M. (2009). Place illusion and plausibility can lead to realistic behavior in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 364, 3549–3557.

U.S. Department of Education. (2010). *Transforming American education: Learning powered by technology* (National Educational Technology Plan 2010). Washington, DC: Office of Educational Technology, U.S. Department of Education. Retrieved from http://www.ed.gov/technology/netp-2010

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education* 16(3), 227–265.

Wang, C., & Gierl, M. J. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement,* 48(2), 165–187.

Williamson, D. M., Bejar, I. I., & Mislevy, R. J. (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Erlbaum.

About the Author

Chris Dede is the Timothy E. Wirth Professor in Learning Technologies at Harvard's Graduate School of Education. His fields of scholarship include emerging technologies, policy, and leadership. His funded research includes five grants from NSF and the Gates Foundation to design and study immersive simulations, transformed social interactions, and online professional development. In 2007, he was honored by Harvard University as an outstanding teacher, and in 2011 he was named a Fellow of the American Educational Research Association.

Chris has served as a member of the National Academy of Sciences Committee on Foundations of Educational and Psychological Assessment and a member of the 2010 National Educational Technology Plan Technical Working Group. His co-edited book, Scaling Up Success: Lessons Learned from Technology-based Educational Improvement, was published by Jossey-Bass in 2005. A second volume he edited, Online Professional Development for Teachers: Emerging Models and Methods, was published by the Harvard Education Press in 2006. His latest book, Digital Teaching Platforms, will be published by Teachers College Press in 2012.

Dede, C. (2013) Opportunities and Challenges in Embedding Diagnostic Assessments into Immersive Interfaces. *Educational Designer,* 2(6). Retrieved from: http://www.educationaldesigner.org/ed/volume2/issue6/article21/