One alternative is to use computer-based programs that simulate the experience a student would have outdoors.

Our increasingly urbanized society and our technology-mediated lifestyles distance most of us from the biological and non-human physical world. Environmental problems present a challenge for education and outreach because they are inherently complex, interdependent, and interrelated. The need for environmentally literate citizens has never been greater. We need people who can work toward an ecologically and economically sustainable environment.

Students who use the computer game Urban Science learn to think like professional urban planners and, in the process, improve their ability to address environmental problems. The game creates opportunities for students to develop science, ethics, and practice to build a professional urban planning frame of reference, or epistemic frame.

In a recent study WCER researcher and former graduate student Elizabeth Bagley found that students who played the game used more scientific language and gave more specific recommendations for addressing environmental problems.

Bagley found that "virtual" environments made possible by computer games are well positioned to provide students with high-quality environmental education.

Students use Urban Science in a supervised setting where mentors play a central role. Mentors

- help students carry out complex tasks
- facilitate cycles of real-world learning through frequent and strategically-placed conversations with the students
- model a professional "epistemic frame" by asking players to reflect on what worked, what did not, and why scaffold a way of seeing and solving problems that the players can adopt.

Throughout the game, students and their mentors interact through reflection meetings where they discuss completed activities and plan next steps in the project.

(continued on next page...)
Our nation needs solid evidence about what programs will most benefit students, and we need more skilled researchers who can carry out that research. Research training programs can help.

For example, graduate students in WCER’s Interdisciplinary Training Program in the Education Sciences recently worked with Professor Geoffrey Borman to find that large-scale, data-driven reform efforts can lead to significant improvements in student mathematics achievement. You’ll read about their findings in this issue of Research Highlights. You’ll also read about the work of graduate student Elizabeth Bagley, who collaborated with Professor David Williamson Shaffer and colleagues in UW-Madison’s Epistemic Games laboratory. Bagley’s recent dissertation argues for the effectiveness of computer simulations in helping students improve their ability to think like professionals as they solve complex problems involving science, ethics, and professional practice.

Also in this issue, WCER researcher Jeffery Watson discusses the importance of accounting for team teaching. Based on his work at WCER’s Value-Added Research Center, Watson says districts and states should not implement high-stakes accountability policies based on student-teacher linkages without first analyzing the quality of those linkages, including team teaching.

And former WCER researcher Tony Milanowski discusses how to improve teacher evaluation. He urges evaluators to adopt recent advances in measuring teaching practice and to use them along with the developing technology for measuring outcomes.

Bagley’s research focused on whether and how these reflection meetings created opportunities for students to develop a combination of the skills, knowledge, values, and identity of environmentally literate urban planners. In particular she wanted to determine whether virtual (online) interactions between learners and mentors were as effective as in-person interactions. Specifically, how did mentoring communication change (a) the quantity of the discourse, (b) the quality of the discourse, or (c) the impact on players’ learning outcomes and engagement?

Bagley’s study found no significant differences in students’ level of engagement or learning outcomes between the in-person condition and the virtual condition. Players in the online condition were as engaged as those with face-to-face mentoring, and they derived similar benefits from playing the game.

Whether face-to-face or virtual, mentors led students to use similar professional discourse. The occurrence of elements of the “epistemic frame” within mentor-player discussions followed similar patterns, and students in both conditions produced professional-quality documents and learned professional ways of problem solving.

These results suggest that the key function of the mentors—to communicate professional ways of thinking—was not diminished in the online chat condition.

In other words, mentors should consider “stop talking and type” since Bagley’s results suggest that the mentoring condition didn’t affect the players’ reflection meeting discourse, learning outcomes, or level of engagement.

Moreover, the results of this study have the potential to influence the design, implementation, and assessment of virtual environments. This study suggests that learning in a virtual environment like Urban Science is viable and desirable because virtual environments can expand the range of what players can realistically do and thus also the problems they can address, the possible collaborations they can participate in, and the communities they can inhabit.

Learning in a virtual environment gives players a chance to see how the world—or at least some piece of it—works under the guidance of a mentor.
Information is power.
And good information empowers educators to improve teaching and student learning.

This is what Geoffrey Borman and graduate students Deven Carlson and Michelle Robinson discovered when studying whether implementing a data-driven reform effort could bring about district-wide improvements in students’ mathematics and reading achievement. They fielded a randomized experiment in over 500 schools within 59 districts and seven states in which approximately half of the participating districts were randomly offered quarterly benchmark student assessments and received extensive training on interpreting and using the data to guide reform.

They found that large-scale, data-driven reform efforts can lead to significant improvements in student achievement. Their findings will help guide the growing movement toward data-driven reform on achievement outcomes.

Deven Carlson and Michelle Robinson are both graduate research fellows in WCER’s Interdisciplinary Training Program in the Education Sciences. Carlson’s home department is political science; Robinson’s is sociology. The Training Program aims to prepare a new generation of education science scholars to provide solid evidence of “what works” in education.

The study examined an initiative fielded by the Center for Data-Driven Reform in Education (CDDRE) at Johns Hopkins University. The CDDRE program bridges two major approaches to low-achieving districts and schools: data-based district reforms and comprehensive school and classroom reforms. It aligns the efforts of state, district, and school-based educators with the goal of accelerating achievement in low-performing schools.

The research compared districts and schools in the CDDRE program to those that operated as usual, without benchmark assessments and associated services.

Results show that CDDRE helps school staff and district staff understand data on student performance, generate additional data to guide school improvement efforts, identify root causes of important problems, and select and implement evidence-based programs directed toward solving those problems.

Most of the districts and schools in the study were low performing, but in many other respects they were diverse, which lends authenticity to the findings:

Geography: The schools and districts are spread across seven states that represent nearly every region of the country.

District type: The sample contained rural districts and large, urban districts.

Socioeconomic: The proportion of students eligible for free or reduced-price lunch varied across districts.

The benchmark assessments monitored the progress of children in Grades 3 to 8 in mathematics and reading and guided data-driven reform efforts. The outcome measure was school-level performance on state-administered achievement tests.

Borman’s analyses tested the effects of the 1st-year components of the CDDRE treatment. The findings show significant improvements in math and some improvements in reading.

The data-driven CDDRE reform model was found to have a positive effect on student mathematics achievement: Assignment to the treatment group was estimated to increase average achievement by approximately 0.06 student-level standard deviations, which is statistically significant.

In reading, the results were positive but did not reach a conventional level of statistical significance.

Taken together, the results indicate that district-level assignment to implement a data-driven reform initiative can cause increased achievement, particularly in mathematics.

(continued on page 7...)
A recent study by Jeffery Watson and colleagues assessed a large urban school district’s student information system to determine whether and how student-teacher linkage data matched the reality of schools and classrooms. The VARC team assessed the degree to which teachers were correctly associated with courses, the degree to which student rosters were accurate, and the extent to which team teaching and additional instructional services occurred.

The study focused on nine schools in this urban district. The schools varied in terms of size, grades served, and location within the district. But all served at least two grade levels in which students took the state standardized test.

The study found three ways that errors in student-teacher linkage data are introduced:

- **Incorrect class rosters**: About 2.5% of all student records were associated with incorrect courses, although this varied by school.
- **Incorrect teacher-course associations**: Incorrect teacher-course assignments were observed to corrupt more student records (7.6%) than incorrect student rosters. That’s because a single misplaced teacher-course association resulted in incorrect student-teacher linkages for all students assigned to that course.
- **Failure to capture team-teaching instructional practices**: Team teaching was observed in all schools and it accounted for errors in 4.1% to 30.2% of student records validated by this study. Most teachers reported sharing between 50% and 75% of instruction with another teacher.

The validity of student-teacher linkage data and classroom-level value-added estimates depends on two things: the degree to which team teaching and innovative practices occur, and the degree to which a district is able to systematically record classroom-level practices.

Schools can be closed. Compensation for teachers and principals can be affected. Classroom Value-Added analysis metrics are now being applied in a variety of ways, ranging from simple school and classroom level reporting to high-stakes decisions. As school districts and states use these classroom value-added estimates in such high-stakes contexts it’s important to fully understand any limitations in the links between student achievement and teacher performance.

Linking performance data of students to their teachers might seem like something that district data systems automatically and routinely do, yet in most cases doing so is a real a challenge. That’s because most school student information systems were not designed to account for team-teaching approaches.

Jeffery Watson, a researcher with WCER’s Value Added Research Center, explains that most states and districts lack the capacity to measure the effects of team teaching. Their data systems don’t account for innovative models of instructional organization and can’t verify such student-teacher linkages.

Team Teaching Practices Affect Value-Added Measurements

The study found three ways that errors in student-teacher linkage data are introduced:

- **Incorrect class rosters**: About 2.5% of all student records were associated with incorrect courses, although this varied by school.
- **Incorrect teacher-course associations**: Incorrect teacher-course assignments were observed to corrupt more student records (7.6%) than incorrect student rosters. That’s because a single misplaced teacher-course association resulted in incorrect student-teacher linkages for all students assigned to that course.
- **Failure to capture team-teaching instructional practices**: Team teaching was observed in all schools and it accounted for errors in 4.1% to 30.2% of student records validated by this study. Most teachers reported sharing between 50% and 75% of instruction with another teacher.

If value-added methods do not account for these practices when attributing student growth to individual classrooms, educators who engage in team-teaching and multiple instructions will be less likely to buy in to, and trust, value-added estimates.
Team teaching occurs when more than one teacher plans and delivers instruction to a student or group of students. Team-teaching practices include these instructional methods:

**Pull-out instruction:**
a student is removed from a classroom to receive alternative instruction from another faculty member.

**Push-in instruction:**
instructional services are provided by a mainstream teacher for children who otherwise receive most of their instruction in a resource room, or self-contained special education room.

**Summary**

For the nine schools included in this study, much of the data (74.2%) in the district’s student information system validly reflected how teachers delivered instruction to students. But student-teacher linkage data quality was inconsistent among schools and was likely driven by characteristics of the schools themselves. For example, in schools where teachers use team teaching there’s more likelihood of error in student information system data. Overall, for the small sample included in this study, the accuracy of student teacher data varied between 54.7% and 92.5%.

Watson says that school districts implementing classroom-level Value-Added estimates should carefully consider how team teaching could affect the validity of their estimates. If team teaching introduces bias into value-added estimates then it will be important for districts to verify when team teaching occurs, and for economists to develop methods that can take into account information on team teaching.

Watson says districts and states should not consider implementing high-stakes policies based on student-teacher linkages without at least analyzing the quality of those linkages, including team teaching. If a data quality assessment is conducted, a validation process may be required before stakeholders will accept the validity of classroom level estimates of value-added.

Watson says the findings from this pilot study provide evidence that warrants larger scale analyses of student-teacher linkage data. Next steps for his research include replicating the current study using an online verification process for an entire district so that more information about these factors can be gathered.
Assessing Teaching Practice

Done well, evaluation of teacher performance has the potential to improve student achievement. Done indifferently, it has the potential to stagnate student achievement.

Too often, teachers are evaluated by a single administrator, with minimal training, who rates the teacher satisfactory or unsatisfactory, based on a single classroom observation. Former WCER researcher Tony Milanowski, now with Westat, urges evaluators to adopt recent advances in measuring teaching practice and to use them along with the developing technology for measuring outcomes.

Policymakers are disappointed by the tendency of current performance evaluation practices to rate almost every teacher alike. Many have turned to performance measures based on students’ standardized test scores. But even when used correctly, these measures don’t provide enough information to improve teacher performance. Outcome measures, such as those based on value-added, and measures of instructional practice need to be used together.

Milanowski recommends that states and districts develop practice measures by translating their visions of effective instruction into explicit models of competency. These would include what teachers need to know and do to carry out state or district priorities. The model can serve as the foundation for a set of practice measures that include observational rubrics for:

- performance evaluation and management,
- performance assessments that would be part of tenure and pay systems, and
- walkthrough tools for day-to-day performance management and for evaluating the implementation of instructional strategies.

Developing a teacher competency model doesn’t require reinventing the wheel, Milanowski says. Designers can start with state teaching standards, for example, those developed by the National Board for Professional Teaching Standards. It’s good to begin with an already existing model, he says, because that captures the aspects of teaching that are similar across states and districts. Evaluation designers can then add competencies that reflect the needs of a particular state or district, and those needed to support local instructional initiatives and strategies.

Milanowski says three measurement systems are needed to assess teaching practice: 1) observations of classroom practice, for use in periodic formal teacher evaluation, (2) teaching “work samples,” or performance assessments, for decisions including granting tenure or movement up a career ladder, and 3) classroom walkthroughs that provide information for everyday performance management.

*Productive classroom observations* use a measurement system that is reliable and valid. Evaluations should help teachers learn from the results. Feedback should be specific and should refer to the rating scale. It should help teachers understand why they received the scores they did. A trained person should be available to assist teachers who want to improve. Other kinds of professional development should be available and linked to the competencies.

*Work samples, or performance assessments,* complement classroom observations. They better portray the teacher’s content knowledge, instructional planning skills, use of formative assessment data, and differentiation of instruction. Teachers describe examples of practice related to specific competencies in response to prompts or questions. They share artifacts, including unit or lesson plans, assignments, completed student work, and assessments.

Formal observations, even frequent ones, can’t provide a clear picture of typical instructional practice, especially about how key instructional strategies are routinely
implemented. *Classroom walkthroughs* (brief, focused visits) are more efficient for this purpose. Walkthroughs get school leaders, instructional coaches, and mentors into classrooms frequently enough to see whether key instructional strategies are being implemented.

Together, the three measurement systems provide a rounded picture of instructional practice. These tools can then be combined with measures of teaching productivity, such as value-added.

**Value-added measures**

As an emerging technology for measuring student performance, and thereby a tool for evaluating teachers, value-added measures have raised some questions, but Milanowski says they are the best productivity measure available. Value added measures compare average student academic growth in a school to the average growth of similar students across a district, and they can account for such factors as a student’s prior performance level or socioeconomic status. Many states are adopting value-added methods, and recent federal education policy has promoted them (e.g., the Race to the Top competition).

The best foundation for making management decisions about teachers is using value-added estimates of classroom productivity together with assessments of teaching practice. But Milanowski cautions that averaging them into one overall measure of teacher performance is not as simple as it seems. They represent two different constructs and have different measurement properties. The two scores can be used together, so just adding them together would be like adding a person’s weight and height.

Setting cutoff points for the value-added measure will require substantial thought, because there is no natural cutoff point that represents acceptable performance.

Schools should use multiple years of value-added data for such decisions as tenure, pay raises, or termination, and the cut-off for tenure or termination decisions should take into account the variability of the value-added measures. Generally, the cut off should be set well below the value-added average.

Value-added productivity measures should also be used to evaluate the quality of teaching practice measures and the effects of human capital management programs.

(...continued from page 3)

**Assessing the magnitude of these effects**

Cluster-randomized designs, like this one, are becoming more common for evaluating the effects of educational interventions. Yet Borman’s is the first known evaluation in which school districts served as the unit of randomization.

Several educational interventions have been found to be effective in small-scale efficacy trials, but they were later found to produce no positive impacts when evaluated on a larger scale. By randomizing nearly 60 school districts, Borman’s results are largely insulated from such concerns. The external validity of these results is further enhanced by the fact that the study design incorporates districts from seven states.

Previous research has provided suggestive evidence that data-driven reform can produce improved student achievement. But these earlier studies were either somewhat underpowered or focused on the evaluation of a pilot program, Borman says. This study provides the best evidence to date that data-driven reform efforts, when implemented at scale, can result in substantive and statistically significant improvements in achievement outcomes.

For more see “A Multistate District-Level Cluster Randomized Trial of the Impact of Data-Driven Reform on Reading and Mathematics Achievement,” Educational Evaluation and Policy Analysis, September 2011, vol. 33 no. 3, 378-398

*Available online at http://epa.sagepub.com content/33/3/378*

---

*Adapted from the article Strategic Measures of Teacher Performance. Kappan, v92 N7, April 2011, pp 19-25.*