Research Article

Investigating the Effectiveness of Computer Simulations for Chemistry Learning

Jan L. Plass,1 Catherine Milne,1 Bruce D. Homer,2 Ruth N. Schwartz,1 Elizabeth O. Hayward,1 Trace Jordan,1 Jay Verkuilen,2 Florrie Ng,1 Yan Wang,1 and Juan Barrientos1

1New York University, 82 Washington Square East, New York, New York 10003
2Graduate Center, City University of New York, New York 10016

Received 1 June 2011; Accepted 31 December 2011

Abstract: Are well-designed computer simulations an effective tool to support student understanding of complex concepts in chemistry when integrated into high school science classrooms? We investigated scaling up the use of a sequence of simulations of kinetic molecular theory and associated topics of diffusion, gas laws, and phase change, which we designed and experimentally tested. In the two effectiveness studies reported, one in a rural and the other in an urban context, chemistry teachers implemented two alternate versions of a curricular unit—an experimental version, incorporating simulations, and a control version, using text-based materials covering the same content. Participants were 718 high school students (357 rural and 361 urban), in a total of 25 classrooms. The implementation of the simulations was explored using criteria associated with fidelity of implementation (FOI). Each context provided insights into the role of FOI in affecting the effectiveness of the interventions when working with groups of teachers. Results supported the effectiveness of this sequence of simulations as a teaching tool in a classroom context, and confirmed the importance of FOI factors such as adherence and exposure in determining the specific environments in which these materials were most effective. © 2012 Wiley Periodicals, Inc. J Res Sci Teach 49: 394–419, 2012

Keywords: multimedia simulations; scaling up; fidelity of implementation; chemistry; cluster analysis

In the pursuit of a society in which individuals are prepared to make informed decisions about their health and welfare and the environment in which they live, science education is considered to be of critical importance (Glenn, 2000). At the high school level in particular, success in science classes is seen as opening doors to science careers, as well as promoting scientific literacy—a prerequisite for being an informed citizen. Yet science is often poorly taught, and students lose interest, electing only the minimum number of required science courses.

For some years now, advances in computer-based educational environments have been accompanied by great hopes for their ability to foster interest and improve learning in
The use of multimedia materials seems to be a natural fit with good science teaching practice, offering opportunities for active learning, contextualized instruction, and the use of visualizations to clarify difficult concepts. In order to explore the usefulness of interactive computer-based simulations in teaching high school science concepts, our multi-disciplinary team—science educators, developmental psychologists, and experts in educational technology and learning sciences—designed, developed, and tested a series of simulations to teach concepts of kinetic molecular theory and associated topics to high school chemistry students. In particular, our goal was to make these simulations accessible to and useful for those learners whose experience of chemistry is very limited. The studies reported here investigated the implementation and effectiveness of our simulations when scaled up for use in broader school contexts. Two research questions framed our effectiveness studies:

1. How effective were the simulations in supporting student learning outcomes when scaled up and implemented in rural and urban contexts?
2. What variations in fidelity of implementation (FOI) were observed within these effectiveness studies?

The effectiveness studies that are the focus of this paper represent the final phase of a four-year research project. As detailed below, the project included the design and development of a sequence of simulations on chemistry topics; a series of efficacy studies of the simulations; and finally the effectiveness studies, reported here, which investigated scaling up the implementation of the simulations. Efficacy studies are used to investigate the success of an intervention design under optimal conditions (Flay, 1986); in our case this involved the use of the simulations in our lab or during a single class period, with the research team in a prominent support role. In contrast, the subsequent effectiveness studies, like other investigations of effectiveness (Flay, 1986), were designed to evaluate the use of the simulations in the real-world messiness of everyday classroom practice. The number of classrooms and teachers involved was increased, teachers were supplied with a standardized lesson plan packet, and research staff stepped back to allow teachers to lead the implementation. The studies reported in this paper complied with the current guidelines for Institute of Education Sciences scale up evaluation projects: The delivery of the intervention was under “routine” conditions, the intervention of using simulations in classrooms had already been tested under “ideal” conditions, the participant sample was diverse, and cluster analysis allowed us explore and interrogate diversity (IES, 2011). The exception was that the studies did not employ an external evaluator.

In the following sections we discuss the use of multimedia simulations as a curricular innovation for science learning; the process of scaling up the use of an innovation; and the question of fidelity of implementation. We then describe the initial phases of our four-year research project as they relate to our effectiveness studies. Finally, we report the results of our effectiveness studies and discuss the implications of these results.

Multimedia Simulations for Chemistry Learning

Today, science educators have access to sophisticated multimedia simulations that allow learners to view and interact with models of phenomena and processes. Such simulations can provide learners with visual representations of dynamic theoretical entities that are difficult to represent in the static environment of the science textbook but are critical for understanding why matter behaves as observed (Ardac & Akaygun, 2004; Honey & Hilton, 2011). These simulations may also encourage active learning by giving students opportunities to manipulate
complex systems and discern patterns through their own investigations (Lindgren & Schwartz, 2009; National Research Council, 1996). While interacting with simulations, learners are engaged in processes of scientific reasoning, such as problem definition, hypothesis generation, experimentation, observation, and data interpretation (de Jong & van Joolingen, 1998; Kim & Hannafin, 2011). As a result, simulations have the potential not only to help learners understand scientific phenomena, but also to foster inquiry and problem solving skills.

While substantial resources have been invested to develop and disseminate computer-based animations and simulations of chemical processes at the molecular level, there has been less investment in evaluating what features of simulations best support chemistry learning for a diverse range of learners (Honey & Hilton, 2011; Plass, Homer, & Hayward, 2009a). Typically, the focus in simulation design has been on whether representations are scientifically accurate rather than whether specific design features support student learning (Plass et al., 2009b). Our goal was to develop and test an innovative sequence of simulations designed to support student learning, and to study whether that sequence would be effective when scaled up for use in multiple classrooms.

Scaling Up: Introducing an Innovation into the Real World

Innovations in education are often developed in protected environments, such as laboratories or test classrooms, but then face challenges when implemented in the field. For example, how a teacher decides to integrate new materials has a direct effect on a student’s experience with those materials (Fogleman, McNeill, & Krajcik, 2011). In this paper we explore whether the learning gains we observed for students using our multimedia chemistry simulations in relatively controlled classroom contexts (Plass et al., 2009a) could be replicated in two quasi-experimental effectiveness studies that were conducted in the complexity of everyday classroom practice (O’Donnell, 2008). Such studies of effectiveness are designed to address whether an intervention “does more good than harm when delivered under real-world conditions” (Flay, 1986, p. 451). Means and Penuel (2005) noted that investigations of scaling up with technology-based innovations should address the question, “What works, when and how?” (p. 176). Although scaling up is described in various ways (see IES, 2011), we found Dede’s (2006) approach helpful for locating our study. He identified four models of scaling up:

1. Systemic change (e.g., literacy in Union City (Carrigg, Honey, & Thorpe, 2005)).
2. Scaling up sets of exemplary curricula and instructional practices that involve collaboration between education contexts (e.g., LeTUS (Blumenfeld, Fishman, Krajcik, Marx, & Soloway, 2000), instructional congruence and English language learners (Lee & Luykx, 2005), Kids as Global Scientists (Songer, Lee, & McDonald, 2003)).
3. Scaling up “in settings that are not only unwilling to undertake full-scale systemic reform, but largely uninterested in even isolated innovations” (p. 559) (including innovations designed to function without a local partner, such as graphical multiuser virtual environments (Dede, Nelson, Ketelhut, Clarke, & Bowman, 2004)).
4. Professional development independent of context (e.g., Keeping Learning on Track (Thompson & William, 2008)).

Our studies represent a variant of these models, which incorporates elements of Models 2 and 3. Over the course of several years, we partnered with high school teachers and students to design and develop a sequence of simulations for use in high school chemistry classrooms. As in Dede’s second model, our design process was based on a theoretical framework of how
people learn, principles of multimedia design, and best practices in science education; we used an iterative design model that allowed us to test and redesign the simulations. Our aim was not to supplant the teacher’s role but to develop an accessible tool to support student visualization of phenomena that cannot be observed directly. Similar to Dede’s third model, our interventions were designed without a local partner. However, we readily found schools and teachers that were interested in innovation, even if they were not in a position to undertake full-scale systemic reform. Principals and teachers were willing, even eager, to try the tool and accept the support we were offering. The process of simulation design and development, efficacy studies, and effectiveness studies is described in more detail later.

Assessing Effectiveness: Fidelity of Implementation

In his discussion of scaling up, Dede stressed the importance of acknowledging specific “contextual factors,” which may influence effectiveness (2006, p.563). Similarly, we understood that the effectiveness of our program could not be evaluated without an understanding of how it was actually used. Fidelity of implementation (FOI), which originated in medical research, is conceptualized in education research as the implementation of curriculum plans as intended by the original design (Lynch & O’Donnell, 2005; O’Donnell, 2008). The need for an examination of FOI in K-12 educational research emerged along with the recognition that teachers and students are active agents with their own ideas of how to use educational interventions—ideas that are not always in line with the intended use. Lee, Penfield, and Maerten-Rivera (2009) argue that FOI is key in scaling up because it helps researchers evaluate whether non-significant results could be an outcome of lack of fidelity in implementation or poor theorizing of the conceptual basis of the intervention.

Following Lynch and O’Donnell (2005) and Songer and Gotwals (2005), criteria for fidelity relevant to our research include:

- **Structure:**
  Adherence between the intended delivery and actual delivery of the curricular unit,
  Exposure to the unit in terms of time associated with the unit implementation and the skills that are emphasized, and
  Program differentiation, which addresses the issue of how the intervention differs from the more traditional version of the unit.

- **Process:**
  Quality of delivery of the intervention compared with traditional approaches.

- **Participant responsiveness:**
  Evaluating student engagement as a measure of student agency associated with their learning, and
  Identifying forms of teacher-student verbal interactions as a measure of participant responsiveness.

In scaling up implementation of our simulations, we recognized the importance of including a plan for assessing FOI criteria. We took a qualitative approach to FOI that could help us identify the quantitative data needed for our two effectiveness studies and serve as a tool to interpret some of the outcomes from these studies. There remains a lack of consensus about whether all elements of FOI must be measured to determine fidelity (Lee et al., 2009; Songer & Gotwals, 2005). We chose to emphasize the FOI criteria that were most relevant to our purpose. In light of the diversity of youth, teachers, and schools participating in the New York City study, we also considered Lee et al. (2009), who noted, “[t]he importance of examining FOI in an educational intervention is greater with increasing student diversity,
since the impact of the intervention may vary among diverse student groups and the overall outcomes may mask differential outcomes among the groups” (p. 836). We therefore planned to conduct classroom observations of teachers and to videotape the implementation of both experimental and control curricular units as permitted by the schools involved.

Molecules and Minds: Designing and Implementing Chemistry Simulations

The effectiveness studies reported in this paper were the culmination of *Molecules and Minds*, a four-year research project. The project included three phases:

<table>
<thead>
<tr>
<th>Phase 1 (Year 1–2)</th>
<th>Simulation design and development; efficacy studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 2 (Year 3)</td>
<td>Preparation of the curriculum unit for use in effectiveness studies</td>
</tr>
<tr>
<td>Phase 3 (Year 3–4)</td>
<td>Effectiveness studies</td>
</tr>
</tbody>
</table>

**Phase 1: Simulation Development and Efficacy Studies**

Phase 1 began with the design and development of the chemistry simulations, based on theories of learning, research in cognition and multimedia, and best practices in science education. We co-developed the simulations with teachers and students in New York City, employing an iterative design process to test and refine each simulation in usability studies. Outcomes from our design experiments informed our simulation design, for example using culturally familiar icons rather than symbols, supporting exploration rather than using “worked out” examples, and using a consistent interface throughout all simulations. This is a departure from the typically accepted strategy, which approaches simulation design from an expert chemistry perspective.

In designing and developing these simulations, we were influenced by the work of Gabel (1999) and Johnstone (1982), both of whom explored the idea of *levels of representation* and its role in how chemists understand the world. Gabel argued that chemical phenomena are initially described at the *observable or macroscopic level*, which is how people experience chemistry. However, these phenomena are usually explained using the properties and behavior of atoms and molecules at the *submicroscopic level*. To complicate matters further, chemists represent the macroscopic and submicroscopic levels *symbolically*, using chemical symbols, formulas, and equations (Scerri, 2006). One of the central goals of chemistry education is to assist learners in navigating between these different types of representations. As Kozma notes, “the use and understanding of a range of representations is not only a significant part of what chemists do—in a profound sense it is chemistry” (2000, p. 15). Kozma and Russell (2005) talk of students achieving *representational competence* when they develop the capacity to move between levels of representation. Accordingly, our simulations were designed by representing the observable using narrative, the explanatory using a representation of the particle model, and the symbolic to communicate connections between explanations and phenomena (see Figure 1).

**Observable Phenomena: Using Narrative to Introduce Everyday Experiences.** One of the cornerstones of science is that observable phenomena need to be explained through the use of theories and models. We have found that narrative is a powerful tool for introducing students to an everyday phenomenon that needs explaining (Milne et al., 2010). Each simulation uses a narrative to present a familiar phenomenon, which students can then investigate by interacting with a dynamic explanatory model of particles in motion. Structurally, our narratives are typical of the genre: They have a beginning, middle and an end; involve agents with intention (our heroine, Gabriella); and result in a final product that is the outcome of interaction
between the author and the reader (Bruner, 1991). Research suggests that humans organize events based on narratively structured thinking, and tend to unconsciously impose temporal and causal relationships on the logico-semantic structure of events (Bruner, 1991; CTGV, 1992; Herman, 2003). We saw narrative as the genre most familiar to students and across cultures (Cortazzi, 1993); our efficacy studies subsequently supported our hypothesis that using narrative would promote student achievement with the simulations (Milne et al., 2010).

**Presenting an Explanatory Model.** Our choice of kinetic molecular theory as the explanatory context for the design of multimedia simulations was based on its significance as a major theory in high school chemistry curricula and a key idea in the Chemistry Core Curriculum of New York State and the Texas Education Agency, consistent with the newly released Framework for K-12 Science Education (Board on Science Education, 2011). Our experience with high school youth also suggested that many students were not introduced to important chemical principles in their middle school science classes, and that providing a basic curricular unit on kinetic theory could serve students well.

Kinetic molecular theory requires an understanding that matter is composed of particles in constant motion, and multimedia can communicate this idea in a visually dynamic way that is not available in static resources such as textbooks. The topics of the four simulations we developed (Kinetic Molecular Theory, Ideal Gas Laws, Diffusion, and Phase Change) often occur sequentially in the high school curriculum, providing the opportunity to develop a curricular unit in which our simulations could be integrated in order to study the process of student learning over a period of time. The design of each simulation was informed by a small set of ideas that are central to understanding kinetic molecular theory and associated topics (Krajcik, 1991; Taylor & Coll, 2002). Chemistry teachers who provided feedback to our proposed key understandings supported these claims from the literature (see Table 1 for a summary).

Presenting the explanatory models through interactive computer simulations provided students with an active learning experience, allowing them to engage in ways that might otherwise be impractical or impossible (Hennessy, Deaney, & Ruthven, 2006). In this case, our simulations presented visualizations of the behavior of particles, otherwise not readily observable. In depicting the particulate model, we chose abstract representations (e.g., a generic

*Figure 1.* Three interconnected levels of representation of chemical phenomena.
container shape rather than a pictorial image of a boiling pot of water) in order to support students in moving beyond the specific everyday examples introduced in the narratives.

*Symbolic Representations.* Both texts and symbols are important elements of our simulation design. Each explanatory model has an associated graph of the relationships between variable that is populated by data points based on students’ decisions and actions in real time (Figure 2). While acknowledging the complexity of these graphs, we also submit that they are powerful symbolic representations that capture relationships among categorical and continuous variables. The ability to create and interpret graphs is a key skill in learning science (Roth, Pozzer-Ardenghi, & Han, 2005): Graphs assist scientists in recognizing and interpreting patterns that are drawn from raw data (Lemke, 1990) and visualizing large amounts of data in systematic ways (Latour, 1987). In the context of our simulations, the embedded graphs introduce students to quantitative aspects of chemistry, an area of focus in the national standards (NRC, 1996).

The four simulations we developed integrate all three levels of representation: observable, explanatory, and symbolic. Consider the example of our simulation of phase change

---

**Table 1**

*Key Understandings for simulations and curricular unit*

<table>
<thead>
<tr>
<th>Overarching Understandings for Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observable (Macroscopic/Phenomenological/Everyday Experience)</td>
</tr>
<tr>
<td>Observed phenomena can be explained by the behavior of particles</td>
</tr>
<tr>
<td>The observed pressure of a gas arises from the combined effect of many particles</td>
</tr>
<tr>
<td>Gases can be compressed, liquids and solids cannot</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Explanatory (Models, Theories)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All matter is made up of tiny particles called atoms</td>
</tr>
<tr>
<td>The particles in solids, liquids and gases are always moving (never still)</td>
</tr>
<tr>
<td>The properties of matter you see are a result of how those atoms behave</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Symbolic (Graphs, Formulas)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A variable is something that changes</td>
</tr>
<tr>
<td>Relationships between variables can be plotted on a graph</td>
</tr>
<tr>
<td>Relationships between variables can be interpreted from a graph</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Content Understandings for Simulations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Kinetic molecular theory</strong></td>
</tr>
<tr>
<td>Gases are composed of particles that are constantly in motion</td>
</tr>
<tr>
<td>There is empty space between the particles of a gas</td>
</tr>
<tr>
<td>The average speed of particles in motion is related to the temperature of the gas (higher speeds correspond to higher temperatures)</td>
</tr>
<tr>
<td>Internal pressure is directly proportional to temperature</td>
</tr>
<tr>
<td>Internal pressure is directly proportional to number of particles</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Gas laws</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal pressure is directly proportional to temperature</td>
</tr>
<tr>
<td>Internal pressure is inversely proportional to volume</td>
</tr>
<tr>
<td>Volume is directly proportional to temperature</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Diffusion</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mass of the particles affects their rate of diffusion</td>
</tr>
<tr>
<td>Temperature affects the rate of diffusion</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Phase change</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Adding heat energy changes state</td>
</tr>
<tr>
<td>Interactions between particles affects the amount of heat energy needed for phase change</td>
</tr>
<tr>
<td>Phase change requires energy change but no change in temperature</td>
</tr>
</tbody>
</table>

---

*Journal of Research in Science Teaching*
At the observable level, phase change is introduced in a narrative involving our heroine, Gabriella, her brother, Tac, and their attempt to heat water to make hot chocolate. This narrative is used to contextualize phase change in an everyday case familiar to students. At the explanatory level, we provide a visualization of the particulate model of the dynamic behavior of molecules undergoing phase change from liquid to gas. In our design, pairs of green, red, or blue dots represent various gas molecules inside a container; burners represent heat (changed by moving a dial); and a thermometer connected to the interior of the container represents temperature. At the symbolic level, a graph is generated with data points based on each user’s manipulation of the particulate model. Each of the four simulations was developed with a similar structure to introduce students to these three interconnected levels of representation.

Simulation design questions (e.g., the role of icons and symbols in design; whether worked-out or exploratory conditions better supported learning; the role of narrative for...
introducing everyday experiences; the value of including the graph and model together) were first explored in a series of studies in a laboratory setting. Efficacy studies were then conducted, first under laboratory conditions and then in intact classrooms where researchers controlled all structural and process elements of FOI. The classroom efficacy studies were conducted in several public schools in the New York City area and rural Texas. This sequence of studies resulted in a simulation design that was theoretically derived, empirically validated, and highly usable (for more information on the efficacy studies conducted see Plass et al., 2009a, 2009b).

Phase 2: Preparation of the Curriculum Unit for Implementation in Effectiveness Studies

Once the simulations had been experimentally tested, we began to develop the curricular units that would be used in our effectiveness studies. Our approach to intervention development references a seminal paper on fidelity criteria by Mowbray, Holter, Teague, and Bybee (2003): We used a set of materials that had been proven efficacious in earlier studies, and consulted experts as we planned for implementation. We solicited input from teachers and science educators—our pedagogical and curriculum experts—and conducted observations of teachers implementing the simulations into their chemistry curricula.

In order to develop the curricular units, we first secured the cooperation of three teachers interested in integrating the simulation sequence into their classrooms. During the Fall 2007 and Spring 2008 semesters, we observed and recorded all class sessions in which these teachers and their students used the simulations, discussed the simulations, or performed demonstrations or experiments pertinent to the content of the simulations. We found that each teacher used the material in very different ways. One teacher prepared worksheets for his students to use as they worked on the simulations. He utilized simulations every other day, with classroom demonstrations and reflective discussion in the intervening lessons. By the end of two weeks, his students were better able to manipulate the simulations systematically in order to conduct experiments and graph data points. Based on our observations and student learning outcomes, we adopted his approach for our effectiveness study.

Consistent with findings summarized in the meta-analysis by Smetana and Bell (2011), we took the approach that simulations are most effective when used together with, rather than in lieu of, other materials and experiences. With input from a curriculum design expert, we adapted and refined the teacher’s worksheets as well as the labs and demonstrations he had used, adding supplementary information and demonstrations/labs. Using this material, we then developed two parallel versions of a two-week curricular unit on kinetic molecular theory and its applications for use in the effectiveness study. The experimental version of the curricular unit incorporated a series of lesson plans, worksheets, and demonstrations/labs, along with the simulations. The control version used nearly identical plans, worksheets, and demonstrations/labs, but replaced use of the simulations with additional text-based instruction (Table 2). Each teacher taught both control and experimental classes.

Key to the organization of these curricular units was the sequencing of the simulations and text-based materials, beginning with simple chemical systems in which intermolecular interactions are minimal (e.g., kinetic molecular theory) and the number of variables that must be conceptualized is restricted (e.g., diffusion). Next followed the gas laws, in which the number of variables under consideration is greater, and finally phase change, in which intermolecular forces are important. Content goals of the curriculum materials were aligned with key understandings associated with kinetic molecular theory, which we had identified during the design process (Table 1) and which were the basis of our learning outcome assessment. Skill development goals included students engaging in scientific reasoning by examining
variables to make causal claims, using some of the tools and language of science to represent these understandings, and improving graphical understanding. Behavioral goals were to engage students in the study of chemistry by sparking their interest and helping them to make connections to everyday phenomena and chemistry.

**Phase 3: Effectiveness Studies**

Once the two versions of the curricular unit were prepared, we ran effectiveness studies, one in rural southeast Texas (late Spring of Year 3, 2007–2008) and one in New York City (Fall of Year 4, 2008–2009). Our experience working with teachers, students, and schools
from these two contexts added to our appreciation that even though they face similar deep challenges associated with funding, infrastructure, and resources, there can be significant differences as well (Theobald, 2005). Some of our research procedures and analysis differed in response to the differences in the two research settings, as detailed below.

In both settings, we supported teachers in their implementation of the materials. A senior member of the research team ran a workshop with teachers prior to the beginning of the study to distribute a written packet describing the materials, discuss the course of study, and provide an introduction to the simulations; teachers were invited to get in touch with that researcher at any time. Just before the beginning of the study, a senior researcher visited each site to get acquainted with the teacher and the site and to check on technical issues. During the study, a senior researcher was present in both experimental and control classrooms for the first few days in order to address any questions or concerns teachers may have had.

Study 1, conducted in a rural setting in Texas, addressed our first research question, which asked about the effectiveness of simulations to support student learning when scaled up in rural and urban settings. Our student sample was ethnically relatively homogeneous, with 76% of the students identifying themselves as Hispanic; the participating schools were comparable to one another with regard to demographic make-up, neighborhood, and attendance rates. With a focus on recording information on the adherence and exposure elements of the structure criteria of FOI, we embedded trained observers into each classroom to collect field data (Table S1).

During the period of our study, Texas chemistry education was undergoing some change. For the first time, aspects of kinetic molecular theory were to be included in the essential knowledge and skills for Integrated Physics and Chemistry course, typically a minimum science requirement for high school graduation in the schools we worked with. One implication of this change was that even experienced chemistry teachers had relatively little experience with this area of chemistry. This meant that our simulation sequence was of particular relevance, and we were able to interest the superintendents of several school districts in our work. In Texas, once the superintendent of a school district had agreed that the district would participate, there was an expectation that schools and teachers would also agree.

Study 2, conducted in the urban setting of New York City (NYC), is more representative of the complexity of implementation research. The NYC study involved public schools that differed significantly from one another with respect to demographic make-up, neighborhood, and attendance. Unlike the centralized approach in Texas, all approval for participation in our NYC study had to come from individual school principals, who would never agree to participate in a study without the approval of the subject teacher. One result of this was that participation of chemistry teachers in our effectiveness study was completely voluntary, which likely resulted in a highly motivated pool of teachers. We expanded our effort to capture FOI elements of structure and participant interaction by videotaping classes (not available to us in Texas), as well as conducting detailed classroom observations. In Study 2, in addition to comparing groups on the basis of learning outcomes as we did in Study 1, we conducted further analysis of data in an effort to capture some of the complexity in performance of this diverse population of students by conducting an exploratory cluster analysis. Procedures for, and results from, the two studies are described below.

Simulation Effectiveness: Two Studies on Implementation

**Study 1: Initial Comparison Study**

*Participants.* A total of 357 students in 20 classrooms from four public high schools in rural Texas (School 1, \( n = 90 \); School 2, \( n = 60 \); School 3, \( n = 46 \), and School 4, \( n = 161 \))
participated. 129 students (7 classes) were randomly assigned to the control condition, 228 (13 classes) to the experimental simulation condition. Each of the 5 participating teachers taught at least one class in the control condition and one in the experimental condition.

**Instruments.** Student learning outcomes measured were chemistry comprehension, chemistry transfer, and graphing skills. Chemistry knowledge was assessed using an 18-item measure ($\alpha = .79$) adapted from the New York State Chemistry Regents Examination; this exam was used as a basis for the measure in order to insure content validity of our chemistry knowledge assessment. This test employed 10 multiple-choice comprehension items (Figure S1) and 8 more complex open-ended transfer items (Figure S2). Comprehension items are designed to test whether a student has understood the information directly addressed in instructional materials; transfer items demonstrate a student’s ability to apply concepts learned in one situation, such as that represented by the narrative element of the simulation, to a different situation or problem. One point was given for each correct response to multiple-choice comprehension items. Two independent graders scored the answers to the open-ended transfer test questions. The initial inter-rater reliability, measured by percentage agreement, was .88. Further discussion between the graders resolved all differences. Learners’ graphing skills were assessed using a 15-item, multiple-choice test (TOGS; $KR = .81$; McKenzie & Padilla, 1986). Construct and criterion validity of this measure has been previously established (McKenzie & Padilla, 1986). Both the chemistry knowledge and graphing tests were administered as paper-based tests prior to the instruction and again after the two-week instructional sequence (Additional Methods are available as Supporting Information accompanying the online article).

**Study 1 Results**

In conducting our analyses, we realized that our findings on FOI had implications for our group comparison analysis. Consequently, we first report our findings on FOI.

**FOI.** Data provided by observers placed in all classrooms allowed us to address our second research question about FOI. As detailed below, two of the teachers (Schools 3 and 4) decided not to implement the curricular units as intended, replicating findings from other studies in which teachers agree to adopt an intervention but do not implement it in a way that is consistent with the proposed structure (Mills & Ragan, 2000) or decide to change the implementation protocol in some way (e.g., Mowbray et al., 2003).

**Adherence.** Teachers in Schools 1 and 2 achieved a high level of adherence. They implemented the curriculum package as intended, fulfilling the requirements of scope and sequence for both the simulation-based and text-based units, and completed all pretests and posttests. The teacher at School 3 retained the scope of the curricular units but not the sequence. The teacher at School 4 did not administer the posttests.

**Exposure.** In School 3, presentation of the curriculum over three days instead of the recommended two weeks meant that the exposure criterion of FOI was not met, which is associated with missing data as well. Other researchers have also identified the issue of missing data in evaluating the effectiveness of an intervention (Fishman & Pinkard, 2001). Student absences from specific classes because they are involved in other school organized activities or absences for the day can be addressed in the quantitative analysis through the use of techniques such as multiple imputation, but the issue remains that whatever the
Program Differentiation. Because all teachers in this study used the packets developed for the intervention, the control and experimental units were comparable with regard to the topics covered, lesson plans, and student worksheets for the classroom demonstrations/labs. They differed only in whether or not simulations were integrated into the curriculum (Table 2). Each teacher taught some classes in the control and some in the experimental condition, in order to minimize teacher effects in our study.

As a result of our FOI findings, Schools 3 and 4, which did not meet the FOI criteria, were excluded from the group comparison of learning outcomes. While we respect that these teachers made informed educational decisions that led them to adapt the unit plan, the lack of adherence with missing post-test data in one case, and the lack of adherence and exposure in the other case, led us to exclude them from the quantitative analysis.

Group Comparison of Learning Outcomes. The following results report the analysis of data from teachers who used the simulations as recommended (Schools 1 and 2). This data set includes 207 students from 10th and 11th grade (54% female; 76% Hispanic). We used students rather than classes as the unit of analysis, with teacher included as a random factor. A total of 144 students were in the experimental simulation condition and 63 in the control condition.

The statistical model employed to analyze the results of this study is a mixed regression (Raudenbush & Bryk, 2002), an analytic approach capable of addressing dependencies due to students being clustered within classrooms. Separate analyses were conducted for each dependent variable (posttest scores on comprehension, transfer, graphing skills). Condition (simulation group vs. control group) was dummy-coded and included as a fixed factor; classroom was included as a random factor. For each equation, condition, classroom, and pretest score on each dependent variable were regressed on the posttest score of that variable.

Another reality of school-based implementation research is that of missing data. In this analysis, missing data were handled by using the method of multiple imputation (MI; Graham, 2009). MI applies a simulation methodology to optimally average out missing data by simulating values from the predictive distributions of the missing data, given the observed data in several datasets, which are then analyzed in the usual way. MI is substantially more efficient than older procedures such as listwise deletion/complete case analysis. Amelia II software (Honaker, King, & Blackwell, 2007) was used to generate 20 imputations, following standard recommendations in the literature. A very mild prior was necessary to obtain good convergence, and all recommended diagnostics were satisfactory. Subsequent data analysis was done in Stata 10.1 (StataCorp, 2007) using the xtregr program with maximum likelihood. No problems were observed with convergence in any of the MI replications.

After accounting for effects of pretest comprehension scores and the random effects of classroom, no effect of treatment was observed for posttest comprehension scores. In contrast, after accounting for pretest transfer scores and the random effects of classroom, a significant treatment effect was found for posttest transfer scores, with students from the simulation-based curricular unit (experimental) having a significantly greater increase in transfer scores than students in the textbook-based curricular unit (control) (Table 3), $\beta = .68$, $p < .05$. This difference corresponds to a small effect size, Cohen’s $d = .25$, with an observed power of .43. Finally, after accounting for effects of TOGS pretest scores and the random effects of classroom, a significant treatment effect was found for TOGS posttest scores, with the experimental group having significantly greater increases in posttest TOGS scores than the control
group, $\beta = 2.07, p < .01$. This difference corresponds to a large effect, Cohen’s $d = .56$, with an observed power of .98.

The results from Study 1 indicate that for students from rural communities in Texas, the simulation-based curriculum promoted the transfer of chemistry knowledge and enhanced graphing skills. The improvement in graphing skills suggests that the design of the simulation, in which the explanatory model is associated with a graph based on the variable values chosen by students, supported understanding of graphing principles and conventions, and that such a focus may not be a common education experience for a significant number of the student participants. The results also indicate that simulation use supported student visualization of particulate nature of matter and that they were able to apply these understandings to other contexts. Students’ ability to better complete transfer questions provided evidence that simulation-based curricular units do provide a structure for supporting students to move between levels of representation.

**Study 2: New York City**

**Participants.** A total of 361 students in 15 classrooms from six New York City public high schools—School 1 ($n = 61$), School 2 ($n = 45$), School 3 ($n = 53$), School 4 ($n = 52$), School 5 ($n = 50$), and School 6 ($n = 100$)—participated in Study 2. Once again, classes were randomly assigned to either condition. After randomization, 148 participants were in the control group and 213 in the experimental group. Of these 361 students, only 194 had complete pretest and posttest data. We mentioned earlier the diversity in NYC that existed at the level of gaining individual principal approval for the research to be conducted in the school, but schools also varied in their structure and character: One was a small, progressive, university-affiliated public school; another was an alternative transfer high school for students.
who had been long-term truants or had been suspended from other schools; another was a day and night school with a significant ELL population; and the other two were comprehensive high schools, which have a mandate to accept all local students who apply for admission.

Instruments and Procedure. All measures, classroom materials, and procedures used in Study 2 were equivalent to those applied in Study 1. Additionally, class periods in which our materials were used or discussed were observed and videotaped.

Classroom Observational Data. Observers sat in on each class session, using paper forms (Figure S3) to note basic information such as daily attendance and which topics were covered. Additionally, observers recorded information at regular intervals, noting what the classroom activity was at that moment (e.g., simulation, lecture, discussion), percentage of students on-task with respect to the assigned activity (e.g., working on the simulation, listening to the teacher lecture, attending to class discussion), and percentage of students engaged in off-task behavior (e.g., chatting, surfing the internet, disruptive behavior).

In addition to classroom observations, we recorded video data in order to conduct a more nuanced analysis of the FOI criteria. Video provides information that is not available from other commonly used methods to verify FOI, such as self-report, which has the limitation of only accessing teachers’ awareness (see also Lee et al., 2009); classroom observations, which we used in Study 1; or completed student work as a de facto measure of FOI (see Songer & Gotwals, 2005). A combination of these strategies provides a richness of data not available from one strategy (Mowbray et al., 2003). Due to guidelines set by the New York City Department of Education, video data focused on the teacher. Each teacher wore a wireless microphone, and the camera was mounted on a tripod with the teacher as the central subject.

Video data was coded by a team of seven trained coders, who transcribed videos using InqScribe (Inquirium, 2005–2008), and then exported the codes to Excel, which converted the codes to a numerical format. Finally, the numerical codes were exported to PASW Statistics 18, Release Version 18.0.0, for data analyses.

The development of the coding system entailed an iterative process. Senior researchers examined a number of videos and proposed an initial set of codes. These codes were tested by multiple coders and discrepancies in coding resolved, with more specific or more general codes added as needed. This cycle was repeated until the team decided that the coding system was feasible and that the available codes adequately captured the constructs of interest in the videos. Ultimately, the codes implemented included structural FOI criteria, such as class activity (e.g., Was the teacher lecturing? Was a simulation in use as intended?), and participant responsiveness criteria, such as teacher-student verbal interactions (e.g., Did a verbal interaction occur? Did it include factual information, evidence of comprehension, or evidence of higher-level thinking?). Each video was coded minute-by-minute. For every minute, coders noted which if any of the coded activities had been observed. Once the video had been coded for a specific class session, the coder retrieved the paper observation form associated with that period. He or she then used the coded video information to verify notes on the paper observation form, transferring the information on percent of students engaged and off-task from the paper form to the Excel sheet containing the coded video data.

Each coder’s work was randomly spot-checked by a senior researcher and reliability statistics assessed. Overall, reliability was satisfactory and overall agreement between the coders was adequate ($\kappa = .70$). Codes with a reliability of less than .65 were not included in the analysis.
Study 2 Results

**FOI.** The richer quality of FOI data obtained from access to video data as well as observers’ notes allowed us to carry out a more detailed analysis of FOI in Study 2, addressing issues of student responsiveness as well as the adherence, exposure, and differentiation criteria.

**Adherence.** Two of the participating teachers did not follow the recommended two-week schedule (see Table 2) for the sequence: The teacher at School 1 used the materials for 29 days in both control and simulation classes, and the teacher at School 2 used the text-based control intervention over 5 days and the simulation-based intervention over 10 days. Teachers in the other four schools implemented the curricular units as intended.

**Exposure.** Exposure, identified as student attendance, was a significant issue for FOI in New York. For the six schools and participating grades (10th, 11th, and 12th), attendance ranged from 75.70% to 93.50% (NYC Department of Education, 2011). In the NYC study, only School 4 was ultimately excluded based on lack of exposure (over half of student data was missing) but student attendance and timeliness were issues for all the schools participating in the study. Attendance and timeliness are especially important aspects of exposure for interventions like ours that are based on a specific sequence and scope. Inadequate exposure accounts for much of the missing data in both Texas and New York City.

**Program Differentiation.** In Study 2, as in Study 1, each teacher taught at least one class using the experimental unit, and at least one using the control unit.

**Participant Responsiveness—Student Engagement and Classroom Discourse.** An analysis was conducted to explore student engagement associated with the intervention because teachers often associate engagement with learning, which has implications for how teachers might continue to implement the intended intervention (Milne et al., 2010).

For purposes of quantification by our observers, we defined *engagement* as students being on task with regard to the assigned classroom activity. Because video data focused on the teacher, with audio obtained from a microphone worn by the teacher, we defined *classroom discourse* as exchanges between the teacher and a student or students. Coding of classroom discourse specified whether an exchange was on-topic (about class topic of study) or off-topic (e.g., about a school assembly or discipline problem). To examine whether integrating simulations into the unit promoted students’ engagement and affected classroom discourse, we compared three types of class periods, with simulation, lecture, or worksheet as the main activity.

Regarding student engagement, analysis of variance (ANOVA) indicated a significant difference among the three types of class periods, $F(2, 37) = 5.27, p < .01$. Bonferroni-adjusted post-hoc comparisons indicated that students were significantly more engaged during class periods with a simulation as the main activity than those with a worksheet as the main activity ($p < .05$). There was no significant difference in student engagement between simulation and lecture periods. This outcome was unexpected; we had expected higher observed student engagement with the simulations. However, this may reflect our operationalization of engagement, which can be defined in many different ways (Domagk, Schwartz, & Plass, 2010; Olitsky & Milne, 2012). Our definition of engagement as students simply being on task, that is, essentially, doing what they were supposed to be doing, may be problematic. In a lecture situation, being on task means paying attention to the teacher: Activity emanates from the teacher, and it is the teacher who dominates discourse (Cazden, 2001). In contrast,
when students are working with a simulation, being on task means attending to that simulation; it is the student who drives the activity and is the agent of his or her own learning. We speculate that the superior learning outcomes of students in the experimental condition could be attributed to the fact that they had more control of their learning as they actively explored chemistry simulations.

Regarding classroom discourse, ANOVA indicated a significant difference in how much teachers and students talked about the course content across the three types of class periods, $F(2, 38 = 14.89, p < .001)$. Post-hoc comparisons indicated that teachers and students engaged in less discussion about the course content during simulation class periods than during lecture or worksheet class periods ($p < .001$). From an FOI perspective this outcome was not unexpected because during these lessons students would be expected to be interacting with the simulation rather than with the teacher. However, the experimental group scored higher on measures of both comprehension and transfer. Did this indicate that the less teachers talked, the more students learned? We believe there is another interpretation of this observation. Rather than discovering the absence of constructive teacher-student dialogue, perhaps we were observing the presence of something else of value—space to think, or cognitive engagement. Alerby and Alerby (2003) note, “it is in the silent reflection that our thoughts take shape…mak[ing] the experience into learning” (p. 46). Perhaps, then, using simulations in the classroom can allow students that space to think; perhaps they may learn things they might never have learned if the teacher had “leapt in with another question” (Seidman, 2005, p. 77). Further, more fine-grained analysis is necessary to explore differences in the content and structure of student-teacher interactions during simulation use as opposed to at other times. We have begun to analyze computer log files of students interacting with the explanatory model and graph to look for metacognitive processes associated with student interactions with the simulations (Brady et al., 2011).

Group Comparison of Learning Outcomes. Based on our findings on FOI in this study, several schools had to be excluded from this analysis. As noted above, neither the teacher at School 1 nor the teacher at School 2 followed the recommended two-week schedule for the sequence. School 4 was excluded based on lack of exposure, as detailed above. We also had to exclude School 3, not because of obvious FOI issues, but because the quantitative analysis tool used for the group comparison, ANOVA, assumes homogeneity of variance. We were concerned that the students at this alternative high school, in which all those enrolled have either been truant for an extended period or previously suspended, represented a unique population that could not be compared with other groups for the purposes of this analysis. Of course, such exclusion makes it even more imperative that other methods, including FOI and cluster analysis, be considered for effectiveness studies like ours (see Exploratory Cluster Analysis, below).

Ultimately, 126 participants from two schools (Schools 5 and 6) were included in the group comparison analysis, all from 10th and 11th grade (52% female; 36% Hispanic, 22% Black, 12% Asian, 9% White). The simulation-based curricular unit (experimental) group scored higher on the transfer test than the textbook-based curricular unit (control) group, controlling for pretest scores, $\beta = 1.59, p = .058$. This $p$ value approaches significance and corresponds to a large effect size, Cohen’s $d = .49$, with an observed power of .80. The experimental group also outperformed the control group on the comprehension test, $\beta = 1.24, p = .05$. This difference, not found in Study 1, corresponds to a large effect, Cohen’s
d = .54, with an observed power of .85. Unlike Study 1, there was no difference in graphing skills between the two groups.

Given that an improvement in comprehension was found in NYC but not in Texas, we explored whether this might be due to different learner characteristics in the two samples, specifically students’ initial chemistry knowledge. We used an ANOVA to compare the two samples on their comprehension and transfer pretest scores and found that students in NYC had significantly lower pretest scores for comprehension $F(1, 270) = 5.94, p < .05$, and transfer $F(1, 270) = 12.29, p < .001$ than their counterparts in Texas.

The results from Study 2 indicate that among the NYC students the simulation-based curriculum promoted gains in both comprehension and transfer of chemistry knowledge. These gains were significant because the participating NYC students began with a lower knowledge base in kinetic molecular theory and associated topics than did their counterparts in Texas. Overall, the simulation-based unit enhanced students’ learning about kinetic molecular theory and associated topics; this was especially so for students with low prior knowledge. We were heartened by this result because our intervention was designed with the goal of supporting the learning of chemistry learners with low prior knowledge.

**Exploratory Cluster Analysis.** All participants with complete pretest and posttest data from the six schools were included in the exploratory cluster analysis ($N = 194$). Cluster analysis provides a compelling means of capturing groupings within complex data sets, making it particularly well suited to implementation research in complex and diverse school environments, such as the ones in this study. Nagin clustering techniques were used to establish the number and shape of groups apparent in this data set (Nagin, 1999). Nagin techniques allow for the clustering of data over time, allowing us to develop information on learning trajectories for participating students. Furthermore, this technique can be used to explore the effect of fixed-factor covariates, such as gender, race, and school, in cluster formation. The Nagin cluster analysis was first run with no covariates, assuming a linear trajectory, to determine the number of clusters present. We then explored fixed-factor covariates such as race, gender, experimental group, and school.

**Establishing Number of Clusters.** Chemistry knowledge pretest and posttest scores were entered as variable of interest, with Time 1 and Time 2 entered as independents. Two-, 3-, 4-, and 5-cluster solutions were investigated, all with linear trajectories. The Bayesian information criterion (BIC; Table 4) of each of these four solutions was compared. The BIC can be defined as the likelihood of the data grouping in this way minus the number of parameters in the model. Comparing BICs, it was apparent that the 3-cluster solution is the largest of the four, and is therefore preferable. The 2-cluster solution has a BIC very close to that of the 3-cluster solution (Figure S4).

Table 4

<table>
<thead>
<tr>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-Cluster solution</td>
</tr>
<tr>
<td>3-Cluster solution</td>
</tr>
<tr>
<td>4-Cluster solution</td>
</tr>
<tr>
<td>5-Cluster solution</td>
</tr>
</tbody>
</table>

$^a$BIC closest to 0.
Both the 2- and 3-cluster solutions give two clusters of roughly equivalent size. In the three-cluster solution, Cluster 1 accounts for 58.5% of cases and Cluster 2 accounts for 36.6% of cases. One of these clusters, Cluster 2, demonstrates a much steeper slope than Cluster 1 in both solutions. However, the slopes of both Clusters 1 and 2 are significant, suggesting that across both these two clusters there was improvement from pretest to posttest. The three-cluster solution adds a third cluster, which accounts for roughly 5% of cases and has a much higher intercept point and a non-significant slope. In interpreting the three-cluster solution, it appears that there is a cluster of 113 students who scored low on the chemistry pretest and whose scores increased somewhat on the chemistry posttest (Cluster 1), a cluster of 74 students who scored low on the pretest and showed a large increase on the posttest (Cluster 2), and a small cluster of 10 students who started out scoring high on the pretest and remained high on the posttest (Cluster 3).

**Investigating Fixed “Risk” Factors.** Nagin cluster analysis allows for investigating fixed-factors such as language, gender, race/ethnicity, school, and treatment group, that may have an impact on groupings of the three clusters. With the addition of each fixed covariate as risk factors, the model is adjusted. Language and gender did not significantly factor in cluster membership; when added into the model, the cluster model output remained essentially the same, with three clearly defined clusters. However, the reference group, Caucasian/Other, was found to be a significant factor in determining the clusters. Students who identified themselves as Caucasian or Other in their ethnicity were found to be 3.2 times more likely to be in Cluster 1 than in Cluster 2, suggesting that these students began the intervention with higher level of prior knowledge than students in other groups. Furthermore, 9 out of 10 Asian students were members of Cluster 2. The three-cluster solution with experimental condition included as a risk factor is shown in Figure 3. Findings from this analysis indicate that experimental condition as a factor was significant, and that students in the simulation classes were 2.5 times more likely to be in Cluster 2 than in Cluster 1 or Cluster 3. This is of interest in particular because Cluster 2 shows the most learning gains, suggesting that students who received the simulations were more likely to demonstrate significant increases in chemistry knowledge.

The 3-cluster solution with School included as a risk factor is represented in Figure 4. The size of the sample prohibited all six schools in the data set being entered into the model.
with a three-cluster solution. Therefore, two schools that were considered to be the most disparate in terms of demographic features, neighborhood, and quality of teachers were investigated as risk factors in order to get a general sense of the extent to which school differences impact student performance. Attending School 6 (one of the schools where all FOI were high) was found to be a highly significant factor; School 6 students were found to be 15.33 times more likely to be members of Cluster 2 rather than Cluster 1. By comparison, students attending School 1 (where FOI was low) were found to fall almost uniformly into in Cluster 1. These results suggest that FOI was a significant factor in relation to our observed learning gains.

The findings from the cluster analysis allow us to acknowledge the diversity of participants—students, teachers, and schools—and to explore the impact of this diversity on student learning. The analysis also provides support for the role of FOI because divergence of implementation among schools impacted the extent to which students’ scores increased from pretest to posttest. Finally, this analysis supports the claim that those students who received the simulation treatment were more likely to increase their chemistry knowledge, indicating that use of our simulations can significantly affect student learning.

Conclusions and Discussion

When we state in this paper that our innovation has been scaled up, we mean that we have moved to broader use by teachers beyond early adopters with whom we worked initially (Fishman, 2005). Our goal in the present research was to see if our simulations of kinetic molecular theory and associated topics could make a positive difference to student learning when integrated into high school chemistry classrooms. In taking our innovation to rural and urban schools, we experienced some of the challenges associated with diverse populations.

Learning Outcomes

The results of our effectiveness studies showed that in both urban and rural settings, use of the simulations in a sequence based on conceptual development of ideas associated with kinetic molecular theory led to better performance in chemistry. Our simulation sequence was effective in supporting both students’ comprehension of ideas and their ability to transfer their understanding to new contexts. This was especially true for students with low prior knowledge: Students in New York City, who scored lower on the pretest of chemistry knowledge than students in the Texas study, made better gains on questions of basic comprehension as
well as on the more difficult transfer questions. Since an important initial goal of our work was to support learners with low prior knowledge, this was a significant finding.

The Nagin cluster analysis (1999) performed on our New York City sample further highlighted the capacity of our simulation intervention to make an impact on learning for lower-performing students. Not only did the analysis indicate that, in general, students who used the simulations were more likely to demonstrate gains in chemistry knowledge, but it also allowed us to discern that there was one group of students who began with low pretest scores and showed particularly dramatic improvement from pre- to posttest. These students were more likely to be in experimental (simulation) rather than control classes. We found Nagin cluster analysis to be a powerful tool for our analysis and see its potential for other effectiveness studies that are conducted in diverse settings because it allows for the results to be both descriptive and explanatory.

Fidelity of Implementation

In our studies, use of FOI was informed by the work of other scholars, such as Lee et al. (2009) and Songer and Gotwals (2005), which encouraged us to examine how fidelity “enhances or constrains the effect of interventions on learning outcomes” (O’Donnell, 2008, p.34). A high level of FOI is expected in efficacy studies, where researchers have more control over implementation, but is more difficult to achieve in effectiveness studies, where the real world complexity of classroom and school contexts intrudes. FOI helps researchers identify whether the absence of an effect is the result of limitations of theory guiding the intervention or the quality of the implementation. However, as O’Donnell noted, rarely is a measure of FOI used to adjust or interpret outcome measures in effectiveness studies. Our approach therefore was to use observation and video recording to obtain rich information about FOI, which helped us identify the school data we could use for quantitative analyses.

Both in Texas and in New York, half of our participating teachers modified the intervention in some way, leading to low fidelity in coherence, exposure, or both. Other researchers have also identified this tension between implementation and adaptation (Mowbray et al., 2003; O’Donnell, 2008). Even though we had organized the sequence of the units with the expectation that they would be implemented this way, we found that the sequence was frequently disrupted, for various reasons. Often, teachers had to respond to the messiness of school life: Students were absent or late on a regular basis; the lab assistant was out; lessons were cancelled for events such as Career Day or a field trip. Such interruptions were observed more frequently in New York City than in Texas, leading us to speculate whether the observed lower prior knowledge of students in New York City could be attributed, in part, to this level of discontinuity. One wonders how, under these conditions, teachers achieve any level of pedagogical continuity, and students any level of learning continuity.

A number of teachers also made independent pedagogical decisions during the implementation of the curricular units that were inconsistent with our preliminary agreements and presented a challenge to the structure element of FOI, which, as results indicated, had negative implications for student learning. This was especially the case in Study 2, in which one chemistry teacher spread the use of the unit out almost four times longer for all classes than we had originally agreed, while another teacher used the materials for different periods of time with the control and the experimental groups. Such actions are consistent with the notion of pedagogical design capacity, which describes how teachers respond to curriculum materials by improvising, offloading, or adapting, depending on their own experiences and

Journal of Research in Science Teaching
interpretations of available resources (Brown & Edelson, 2003). While these might have been informed decisions, abandoning the structural integrity of the curricular sequence impeded our ability to make quantitative judgments about the potential of the simulations to be educationally effective under such different conditions. This situation highlighted the tension between how some teachers see themselves—as experts with respect to knowing the needs and capacities of their students—and the desire of researchers working with teachers to implement specific interventions. Our findings suggest the need for further study in this area of research where teacher aspirations intersect with innovation interventions.

Our investigation of students’ engagement and classroom discourse showed no significant differences between how students responded to the simulations as compared to their response to teacher lecture. Though the level of interaction does differ, our current instruments did not allow us to explore such interactions in sufficient detail. However, this experience has highlighted the value of developing instruments that will allow a finer-grained analysis. In general, our results from both the cluster analysis and group comparisons showed that FOI is a key element that must be considered in effectiveness studies.

Scaling Up

Earlier, we suggested that our model for scaling up included elements of two models outlined by Dede (2006). His second model describes collaboration with schools and districts, while in the third model, conditions for success are often missing, including administrative support, qualified and supportive teachers, well-maintained technology infrastructure, and a population of students that attend consistently. Consistent with Dede’s second model, our scaling up had partial local support. In New York City, our collaboration with the school district was limited to obtaining Institutional Review Board approval for the research, so we lacked broad support for the study. In Texas, we had more initial administrative support, which might help to explain greater FOI. In both contexts we worked with qualified and supportive teachers and a reasonably well maintained technology infrastructure, although technology also introduced challenges. For example, many of the schools had computer carts, which the teacher had to reserve well in advance; though these carts were normally designed to hold enough computers for a full class, our researchers would often find that several computers were not working properly, were missing power cords, or would not connect to the Internet. The activities of distributing computers at the beginning of class, getting all of them up and running, and putting them away at the end of class—while ensuring that none of the machines “walked away”—cut into instructional time and distracted the teacher’s focus from the lesson at hand.

Fogleman et al. (2011), in their examination of teacher adaptation of an innovative curriculum for middle school chemistry colloquially called Stuff, found that attendance of students in urban schools was a significant issue for data collection and, as we also discovered, has implications for how student learning can be assessed. Consistent with these observations, attendance was an issue for students in our study, particularly for those in our New York City partner schools. Poor attendance contributed to the amount of missing data and was evidence of a systemic problem that had implications for all our analyses. Statistical data from the schools documented the attendance problem, and it was confirmed as well by anecdotal data from teachers and observational data recorded by our researchers. One observer, for example, noted that 10 minutes into a first-period class with 28 students registered, only 5 were present. Observations of this nature raise the question of the instructional effectiveness of any innovation if students are not in class on a regular basis. Some schools in New York City have begun to address this issue by making structural changes to the scheduling of the school.
day, using spiral scheduling or responding to research on the biological rhythms of teenagers and the school day (Carskadon, 1999).

In summary, as a body of work, our studies demonstrate practical principles for the effective design and use of interactive multimedia simulations in science classrooms, and contribute to the theoretical understanding of multimedia learning of science. Across both sites of implementation, Texas and New York, there was consistent evidence that the integration of simulations enhanced students’ learning of chemistry in the complex environment of an authentic classroom. In response to the challenge by Means and Penuel (2005) that scaling up technology interventions should address the question of what works, when and how, we can say, based on the data collected in the present studies, that efficacious simulation-based learning environments can be cognitively efficient and pedagogically effective when implemented with fidelity. In addition, the outcomes of our project indicate specific empirically based design principles that should be considered in the design of simulations for low prior knowledge learners, a category which encompasses most high school students beginning the study of chemistry or other science disciplines.

Directions for Future Research

Our findings on integrating our simulation sequence on kinetic molecular theory and associated topics suggest that we need to explore further how concepts should be organized to support student learning. Log files generated by student use offer the potential to provide a richer understanding not only of how students actually use a simulation sequence, but also of their capacity for and processes of metacognition (Chang, Plass, & Homer, 2011). Other directions for future research include a focus on the value of being more explicit about FOI for curriculum interventions and developing models of integration for simulation use in curriculum planning that are effective in a range of different contexts. The tension between adoption and adaptation, and how best to support teachers in helping their students achieve, remains an important focus for future studies of the integration of innovative technology interventions: We need to continue to explore the ongoing challenge in scaling up of finding balance between FOI criteria that are core elements of simulation-based instruction and allowing space for teacher adaptation of the intervention. Finally, in the effectiveness studies reported here, we focused on the content goals associated with our curricular unit, with some assessment of skills development with respect to graphing. Future research should expand this investigation to a broader examination of achievement of the skills development and behavioral goals. Overall, the evidence that our simulation sequence provided effective support for student learning of content goals suggests that the next step is a full scale-up to see if these positive results can be replicated.

The research presented in this paper was supported in part by the Institute of Education Sciences (IES), U.S. Department of Education (DoEd) R305K05014 awarded to Jan L. Plass, Catherine Milne, Bruce Homer, and Trace Jordan.

References


Journal of Research in Science Teaching


Brown, M., & Edelson, D. C. (2003). Teaching as design: Can we better understand the ways in which teachers use materials so we can better design materials to support their changes in practice? Evanston, IL: LeTUS Report Series.


Journal of Research in Science Teaching


StataCorp. (2007). Stata Statistical Software: Release 10. College Station, TX: StataCorp LP.

