


# Using Virtual Rehearsal in a Simulator to Impact the Performance of Science Teachers

Lisa A. Dieker, University of Central Florida, USA

Carrie Straub, Mursion Inc., USA

Michael Hynes, University of Central Florida, USA

Charles E Hughes, University of Central Florida, USA

 <https://orcid.org/0000-0002-2528-3380>

Caitlyn Bukathy, Hands on Educational Services, USA

Taylor Bousfield, University of Central Florida, USA

Samantha Mrstik, Georgia Gwinnett College, USA

## ABSTRACT

This study investigated the use of a virtual learning environment, TeachLivE, using pre-post group design to examine the effects of repeated virtual rehearsal sessions. Based upon past findings on the effectiveness of four 10-minute sessions, the research team used refined methods to examine the effects of these sessions on 102 secondary science teachers. Teachers who took part in the simulated activities significantly increased their targeted behaviors compared to colleagues who had not taken part in the simulation activities. These results of behavior changes that occurred in the simulation were found to transfer back to the real classroom settings for the experimental group (simulation use). Results from this study further validate the impact of simulation in teacher education, showing professional learning in virtual-reality simulated classrooms can positively impact targeted teaching practices in a concentrated amount of time.

## KEYWORDS

Biology, Professional Development, Science Teacher Education, Simulation Methods, Simulation, Teacher Education, Virtual Rehearsal

## INTRODUCTION

Though student scores on the National Assessment of Educational Progress (NAEP) are improving in eighth grade, only 34% of eighth grade students are reaching or exceeding proficiency in science (National Center for Education Statistics, 2017). The Next Generation Science standards aim to increase content knowledge of students with the expectation that students will take on greater challenges in their academic careers. While raising expectations for students, this shift in standards has altered the way teachers are expected to approach instruction, by moving to an increasingly student-centered,

DOI: 10.4018/IJGCMS.2019100101

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

inquiry-based approaches (Bartos & Lederman, 2014; Reiser, Berland, & Kenyon, 2012). These shifts in student expectations require a corresponding shift in teacher professional development. Teacher preparation programs need to prepare teachers for these shifting standards and practices. Moreover, in-practice teachers need immediate retooling to prepare students to meet these standards designed to result in better college and career options upon graduation. If teachers must undergo rigorous professional development (PD) to be ready to teach these standards, their training must be efficient and immediate with strong transference of knowledge (Windschitl, Thompson, Braaten, & Stroupe, 2012) to give students a quality understanding of these new standards in practice.

In inquiry-based science, students are encouraged to take a hands-on, investigative approaches to learning the functions of scientific concepts (Hobson, 2014). Yet, most science classroom discussions follow a pattern of initiation, response, and evaluation (I-R-E) (Tytler & Aranda, 2015). The I-R-E pattern is easy to recognize: the instructor asks a question, usually one to which he or she already knows the answer; the student responds, and subsequently receives feedback or constructive criticism. Instructors use this pattern for various purposes – to remind students of information or spark a discussion (Neal, 2008). The NGSS (2013) point to a departure from this structure to explore problems and ideas that, in the I-R-E pattern, might not see the light of day (National Science Teachers Association, 2014). The NRC Framework and NGSS focus not so much on memorization of facts in the scientific process but on a thorough understanding of core scientific ideas and the ability to discuss them with others (Reiser, Berland, & Kenyon, 2012). The transition from the typical I-R-E pattern to a more challenging and far-reaching exploration of science requires a conscious effort on the part of teachers directing student discourse, so that students can make sense of the material in a constructive way (Bacolor, Cook-Endres, Lee, & Allen, 2014).

## RETOOLING TEACHERS

To engage students at new levels of thinking related to science, teachers need to demonstrate an array of teaching practices in their classrooms, and teacher preparation and PD should target the practices teachers find most challenging (Windschitl et al., 2012). In the *Measures of Effective Teaching* study, Kane and Staiger (2012) report that teachers score lowest for complex teaching skills such as questioning, discussion techniques, and communicating with students about content. University of Michigan's *Teaching Works* (2014) analyzed core capabilities for teachers and developed a set of 19 high-leverage practices (HLPs) of which mastery will likely lead to increased advances in student learning. The HLPs are based on research linking particular practices to student achievement (Loewenberg Ball & Forzani, 2010). The Teaching Works' HLPs span across content, teacher style, and setting, and include eliciting and interpreting student thinking, and providing oral feedback on students' work (Loewenberg Ball, Sleep, Boerst, & Bass, 2009), both of which take place in inquiry-based discussions. Danielson (2011) provided indicators for similar teaching capabilities, including higher-level questioning. Higher-level questions are defined as open-ended questions that allow students to use past experiences, prior knowledge, and previously learned content in order to create a well-thought-out answer (i.e., question statements that begin with "How", "What", or "Why") that relates to new content. For science teachers in particular, questioning appears to be the weakest element of instruction, and researchers have proposed a core set of instructional practices for science teachers, including questioning to elicit student thinking (Windschitl et al., 2012).

## IMPROVING INSTRUCTION USING SIMULATION

This shift in teacher practices, which in turn should lead to shifts in student discourse and learning, is occurring at the same time as computer simulations are emerging in teacher education. Computer simulation is taking center stage as a next generation environment for teacher professional learning, allowing teachers to learn pedagogical skills within content areas. One such computer simulation

environment, TeachLivE, is an immersive, virtual reality classroom simulator that combines real and virtual worlds (a form of mixed reality) to give users a sense of immersion and presence. Teachers interact with student-avatars in real time, having authentic professional learning experiences within any content area. The software that drives TeachLivE allows digital puppeteering of the virtual characters by trained human actors, called interactors. The interactor employs gestures to control animation of the virtual students and allow for blending of human and artificial intelligence to create an authentic immersive experience. These features ensure that project objectives are met with consistency and teachers experience the core element of a simulator, the suspension of disbelief, having a sense the simulated environment is real. This human interaction also can be standardized to provide a rigorous research protocol for authentic but repeatable classroom experiences. Teacher behavior can then be coded while the user is in the system as annotations (tags) in association with key or targeted events. Tags allow for immediate data collection to allow teachers to reflect upon their performance in what is deemed after-action-review in the world of simulation. Simulators also can be used, where appropriate, for in-action feedback, e.g., letting a teacher know that his or her non-verbal messaging may be counter-productive (Dieker, Hughes, Hynes, & Straub., 2017).

Simulation can provide many educational experiences and opportunities not available in real-world settings (Dieker, Straub, Hughes, Hynes, & Hardin, 2014b; Dieker, Rodriguez, Lignugaris/Kraft, Hynes, & Hughes, 2014) while allowing for safe rehearsal of skills until mastery. A research base is emerging, focusing on the use of simulated environments (Mursion, TeachLivE, Sim School) with teachers and teacher candidates (see Andreasen & Haciomeroglu, 2009; Barmaki, 2016; Bukaty, 2016; Dawson & Lignugaris-Kraft, 2013; Elford, Carter, & Aronin, 2013; Elford, James, & Haynes-Smith, 2013; Gallegos, 2016; Straub, Dieker, Hynes, & Hughes, 2014; Vince Garland, Vasquez, & Pearl, 2012; Whitten, Enicks, Wallace, & Morgan, 2013).

Simulated environments provide safe places to practice teaching behaviors at an accelerated pace while receiving rapid corrective feedback (Dieker et al., 2014a; McPherson, Tyler-Wood, McEnturff, & Peak, 2011). Simulation that incorporates an after-action-review based on a theoretical model of performance mastery through feedback (e.g., Hattie & Timperley, 2007) has the potential to reduce discrepancies between current performance and a goal. This immediate shaping of behaviors cannot happen in a real classroom, as real students would be made to wait while their teacher received corrective feedback. Avatar-students can be “paused” and wait patiently without losing valuable instructional time. Most importantly, unlike in real classrooms, teachers can re-enter the environment to fix instructional errors, such as using a continuous I-R-E pattern, with student-avatars without affecting real students, known as virtual rehearsal (Straub et al., 2014). Immersive virtual environments have the potential to revolutionize teacher professional learning, but more research is needed to establish the efficacy and effectiveness of the use of simulation for teacher education.

The purposes of this research study is to evaluate the use of TeachLivE to affect the behaviors of teachers in science discourse during classroom instruction. In this study, science teachers were given an opportunity to practice their use of HLPs in TeachLivE and to evaluate the generalization of those practices to the traditional classroom setting.

## **THEORETICAL FRAMEWORK AND OVERARCHING HYPOTHESES**

Work in computer simulation is grounded in Brown, Collins, and Duguid’s (1989) theory of situated cognition, asserting, “what is learned cannot be separated from how it is learned and used” (p. 88). This theory builds the alignment that the research team sought between the similarities and differences in learning in a simulated versus a real classroom environment. The power of an effective simulator is the activity that occurs in the simulator should be situated in activity bound to social, cultural and physical contexts that align with the “real” environment. Hence, the research team was most interested in measuring the baseline skills of teachers in the situated context of the real environment, to then create a parallel environment in the simulator to practice those skills, and then to measure

the effective of the situated context of the virtual environment by conducting a post-observation of those skills again in the real environment.

The team built these assumptions on the belief that simulation experiences create a contextualized activity that provides learners with the opportunity to practice HLPs with student-avatars (Straub et al., 2014), with the theory of situated cognition being the transference of skills in the simulator to the real classroom setting. The theory of action arises from examining the critical features of professional learning for teachers related to increased student outcomes (e.g., active learning opportunities based on specific teaching practices, such as HLPs). Based on results from a national research study investigating simulation in middle school mathematics classrooms and coupled with findings from earlier studies related to using virtual environments for teacher preparation (Straub et al., 2014), the following overarching hypothesis was formed. “Teachers who engage in virtual environment simulations in high school science, specifically biology, will improve their practice in the simulator, and this higher level of performance will transfer back to their classroom. More specifically, four 10-minute sessions of virtual rehearsal (i.e., practicing the same lesson and HLPs in TeachLivE) will significantly increase teachers’ frequency of use of open-ended questions and content-related affirmation to students in both simulated and real classroom instruction, compared to their colleagues who did not engage in TeachLivE.”

## **RESEARCH QUESTIONS**

The following questions framed this research study. Research Question 1: To what extent does performance differ over four 10-minute sessions of TeachLivE (i.e., questions) when targeted performance feedback [after-action-review] is given? Research Question 2: To what extent do teachers’ performance differ over four 10-minute sessions of TeachLivE when no performance feedback is given? Research Question 3: To what extent does teacher practice of asking questions to initiate student-centered dialogue in a classroom differ after four 10-minute sessions of TeachLivE? Research Question 4: To what extent are the effects of teacher performance related to frequency of content-related affirmations in a classroom after four 10-minute sessions of TeachLivE?

## **METHOD**

### **Participant Characteristics**

Data analyzed in this study were collected at 11 separate research locations nationally comprised of university and school district partners. The 102 participants were high school science teachers. No restrictions were made based on education level of a teacher, number of years teaching, level of class taught, subject area within science taught, or any other demographic characteristics. Demographic data for all 102 participating teachers are presented in Table 1.

### **Sampling Procedures**

Participants were identified via convenience-sampling. All teachers agreed to teach a lesson plan based on specified science content created by the National Institute of Health as model science lessons and were structured to promote discourse. At each partnership site, teachers were nominated by their supervisors with the intent of receiving lesson resources for professional development. Of the 129 teachers contacted, 102 teachers completing all aspects of the study and their demographic data are included in Table 1. Participation by these teachers was entirely voluntary with minimal to no compensation provided.

Data were collected in two settings at all sites, the teachers’ classrooms and in the classroom simulator. A total of 102 teachers were observed in secondary classrooms located in 11 sites across the following states: Florida, Georgia, Idaho, Illinois, Louisiana, Maryland, New York, Texas, and Washington, D.C. Observations occurred for all teachers within a two-week window by trained

Table 1. Teacher demographic data

Variable	Comparison ( <i>n</i> = 51)		TeachLivE ( <i>n</i> = 53)	
	<i>n</i>	(%)	<i>n</i>	(%)
Professional licensure				
Yes	6	(12)	48	(91)
No	40	(78)	2	(4)
No response	5	(10)	3	(6)
If licensed, is license in science?				
Yes	35	(69)	37	(70)
No	1	(2)	4	(8)
No response	15	(29)	12	(23)
Area of certification				
Grades 6-12	26	(51)	26	(49)
Other	6	(12)	8	(15)
No response	19	(37)	19	(36)
Highest academic level				
Bachelor's	18	(35)	27	(51)
Master's	27	(53)	21	(40)
Doctorate	2	(4)	1	(2)
No response	4	(8)	4	(8)
Area of masters degree				
Biology	7	(14)	6	(11)
Other	19	(37)	18	(34)
Not applicable or No response	25	(49)	29	(55)
Years teaching science				
One year	3	(6)	7	(13)
Two years	10	(20)	5	(9)
Three years	1	(2)	3	(6)
Four years	3	(6)	4	(8)
5-10 years	13	(25)	12	(23)
More than 10 years	17	(33)	17	(32)
No response	4	(8)	5	(9)
Age				
18-29	14	(27)	19	(36)
30-39	12	(24)	14	(26)
40-49	13	(25)	8	(15)
50 or above	8	(16)	9	(17)
No response	4	(8)	3	(6)
Gender				
Male	10	(20)	9	(17)
Female	37	(73)	41	(77)
No response	4	(8)	3	(6)
Ethnicity				
Asian	5	(10)	3	(6)
Black	6	(12)	4	(8)
Hispanic	4	(8)	3	(6)
White	27	(53)	39	(74)
Other	5	(10)	1	(2)
No response	4	(8)	3	(6)
Teachers indicated having taught the following grade levels:				
Grade Level	Control Group		TeachLivE Group	
K - grade 5	2		2	
Grades 6 - 8	14		15	
Grades 9 - 12	46		43	

observers from across the country who made direct observation of the teachers' classrooms. These observers were required to complete an extensive online observation training delivered by the lead research site at a level of 95% fidelity prior to any pre or post-observations occurring. School settings ranged over urban, suburban, and rural with public or private enrollment. Classroom simulators were located near teachers' classrooms at university or school district partner sites. Institutional review boards at each site and within each school district examined and approved all research activities.

### **Sample Size, Power, and Attrition**

All 102 teachers delivered the same lesson plan for Observation 1 and another parallel lesson but with new content for Observation 2. These lessons were selected to ensure consistency in the content observed and were selected as model and parallel lessons on using data to make decisions. The two lessons used were model lessons created by the National Institute of Health and validated by content experts in the field to align with the NGSS and biology content. All pre and post observations occurred by the trained observers of the teachers using these required lessons within a two-week window and six weeks of time elapsed between observation 1 and 2. The power analysis reviewed the effect sizes to provide an estimate of the desirable effect size, yielding a range from small to large of  $\eta^2p = .025$  to  $.149$ . An a priori power analysis was conducted (Cohen, 1988) using a medium sized effect (0.25). Power analysis for an F-test Analysis of Variance (ANOVA) within-between interactions resulted in a required total sample size of 48 participants to have 80% power for detecting a medium sized effect (0.25) when employing a 0.10 criterion of statistical significance. Since this was new research in the field with a low risk to humans, a 0.10 criterion was selected. This level of effect size is common in teacher professional development studies as noted by Hattie (2009), as the ability to change a behavior with 40 minutes of intervention is efficient, therefore, a larger Type Two error was acceptable in considering the overall findings. The anticipated number of participants exceeded the suggested number of 48 participants for a medium sized power effect.

## **RESEARCH DESIGN**

The research design was a randomized controlled trial, consisting of two groups of teachers measured pre-post in the classroom, half of whom also were measured four times in the classroom simulator. The random assignment procedure took place at all 11 partnership sites, resulting in two experimental groups.

## **INTERVENTIONS**

Teachers were assigned to one of two groups, and both groups received the same lesson plan resources. The lesson plans selected were aligned with eliciting the HLP's of the project and were designed to enhance science literacy and aligned with disciplinary core ideas and cross-cutting concepts from the next generation science standards, as well as the Common Core Standards for literacy in science. The lesson plans were based on the 5E instructional model and were validated and field-tested in high school biology classrooms as part of a larger module from the NIH Curriculum Supplement Series "Using Technology to Study Cellular and Molecular Biology" (National Institutes of Health, 2005). Lesson 1, entitled "What is Technology?" was the basis for the first observation, while Lesson 2, entitled "Modeling Issues," was the basis for the second observation.

Both lessons had the same structure and parallel content. Each lesson began with a concept map, which served as a mini-task literacy activity (Literacy Design Collaborative, 2015). Teachers showed a model concept map and explained the components to students, then they gave their students five minutes to generate as many ideas as they could related to the prompt in the center of the map. The prompts corresponded directly to the lesson with the purpose of eliciting student thinking related

to the content just before and after the lesson. During the lesson, teachers facilitated a whole class discussion focused on interpreting, inferring and deducing from data, and integrating information to form conclusions.

All teachers received the lesson plans and accompanying video resources via email after orientation and prior to the first observation. Two observations were scheduled at least three weeks apart (one for each lesson) and treatment took place in between observations. Group 1 teachers served as a comparison group and received lesson plan resources while Group 2 teachers received the same lesson plan resources plus four, 10-minute sessions of TeachLivE.

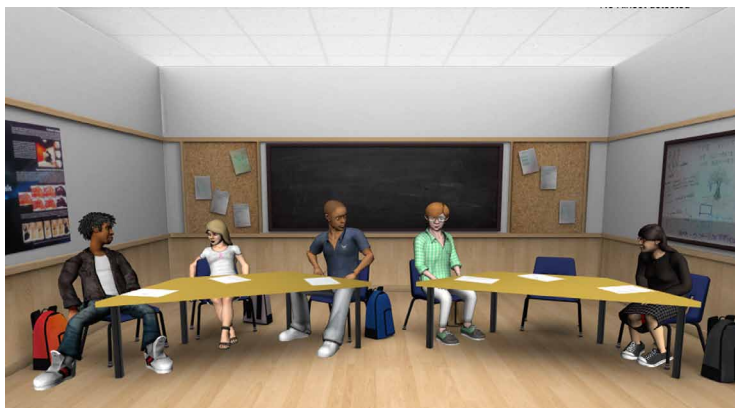
### Group 1: Comparison

Comparison teachers received Lessons 1 and 2 and the accompanying video for Lesson 1 via email. They were given no other intervention as a course of this study but did receive any PD provided by their district throughout the course of the school year. They taught Lesson 1 at pre-treatment observation and Lesson 2 at post-treatment observation.

### Group 2: Simulation

Simulation teachers received both lessons and resources (like the comparisons teachers), as well as four 10-minute virtual rehearsal sessions in the TeachLivE classroom simulator. In the simulator, teachers attended individually and interfaced with a computer-generated, animated student population of five high school avatars digitally controlled by a professional who enacted a highly interactive, authentic simulation of a high school classroom. The software is programmed to react to certain commands of the teacher and the interactor, with the purpose of increasing the teacher's aptitude in the classroom. Classroom simulators at 10 client sites across the country were connected via a secure server to the lead university site, which served as the central distribution point for TeachLivE and provided fidelity of treatment as all sessions were controlled in the same manner by the same team. The teachers experienced computer-simulated classroom activities with the student-avatars as they would with human students in a traditional classroom (see Figure 1: Image of classroom simulation). Visits to

Figure 1. Image of classroom simulation



the simulator occurred over the course of four weeks following the first classroom observation to determine teachers in both groups baseline performance aligned with the identified HLP's being measured in this study.

Simulation teachers participated in one 10-minute session to orient them to the simulation system. Data were not collected during the orientation session, as users were not teaching content but interacting with the student-avatars to learn about their virtual class. After orientation, teachers experienced four 10-minute sessions to virtually rehearse (i.e., targeted practicing of a skill in a virtual environment), with data on targeted behaviors gathered during each session. Teachers typically took part in two 10-minute sessions and returned within a month for two additional 10-minute sessions. Teachers were instructed to teach the first 10-minutes of Lesson 1 for each session, providing repeated rehearsal of the lesson they had already taught.

### *After-Action-Review*

At the close of each session, teachers took part in an after-action-review of their performance led by a facilitator using a digital chart displayed on a large video monitor. After-action-review consisted of four parts: (1) teachers were presented with frequency of observed behaviors (i.e., closed-ended questions (CE), open-ended questions (OE), and open-ended plus questions (OE+)), during the session verbally and on a large display; (2) teachers read examples of question types CE, OE, and OE+ on a large display; (3) teachers were asked to set a goal for their performance in the next session on OE questions; and (4) just prior to commencing the session, teachers stated their target goal of OE questions. Performance goals were not set for OE+ questions, because they are part of the larger category of OE questions. Upon completion of the after-action-review, teachers returned to the simulation for another rehearsal session.

### **Measures and Covariates**

To ensure that all 11 sites had high reliability in data collection, researchers employed methods to enhance the quality of measurements. All data collectors were trained online using a combination of asynchronous assessment and synchronous data collection training on the constructs (e.g., Danielson sub-constructs and HLPs) and methods (e.g., frequency counts during rotating intervals as described above) for data collection. Data collectors used the asynchronous online modules to demonstrate proficiency with the content of observations. Each practice was defined, and a case example was provided. Observers had to pass a multiple-choice content assessment with 90% accuracy for the asynchronous portion of the training. The synchronous online training consisted of a series of rigorous activities delivered via a video conferencing platform that exposed observers to operational definitions and required the collection of frequency counts in real time while watching a video online as a group to simulate classroom observations. Each observer was checked for reliability during the online training and required to complete a synchronous online activity with 95% accuracy.

### **Data Collection**

Quantitative and qualitative observations occurred using a validated practice tool from a previous study, the Teacher Practice Observation Tool (TPOT; Straub et al., 2014), to collect data on teachers in their classrooms pre- and post- treatment. The TPOT tool was used to measure the teacher's target behaviors in this study as well as to gather general data for future research on classroom performance. This instrument was validated by the research team in earlier studies for both face and content validity (Straub et al., 2014).

### *High-Leverage Practices*

Using research from the Measures of Effective Teaching project, descriptions of HLPs, and other empirically-based research in the field, operational definitions for observation were created. Data were collected on teachers' frequency and type of eliciting and interpreting individual students' thinking (HLP #3) both in the classroom simulator and pre-post in their respective classrooms. Data were classified as:



- **Closed-ended questions (CE):** Content questions that have restricted parameters, expecting one possible response as its only acceptable answer; constrains a student's response, such as test questions, yes–no questions and forces choice questions;
- **Open-ended questions (OE):** Content questions to which a number of different answers would be acceptable; content questions that have no parameters and do not constrain student's response.
- **Open-ended plus questions (OE+):** Content questions that ask a student to extend, produce, or combine ideas to generate new ideas (related to Bloom's highest cognitive domain –creating). OE+ questions were included within the open-ended questions category.

For the purposes of this study, the focus was on affirmation related to content (CRA) only and it was defined as: “teacher's positive verbal affirmation about what a student or group of students did or said related to content in a single episode within class” (multiple statements about the same episode counted as one occurrence of affirmation). This definition was used to avoid counting statements such as “good job” or “nice” as these were affirmations but may or may not have been aligned to the content. Hence, only statements such as “good job talking about a pencil being a type of technology” would be scored as a CRA.

### *Sub-Constructs from 2011 Danielson Framework for Teaching*

Eight sub-constructs that correlated with student achievement from the 2011 Danielson Framework for Teaching Evaluation Instrument (Measures of Effective Teaching Project, 2010) were identified. Key words from Danielson's indicators were chosen to create an abbreviated version for classroom observations and combined with the collection of frequency data in relation to describe/explain questions, specific feedback, and wait time. Danielson's levels of performance (i.e., unsatisfactory, basic, proficient, distinguished) were the basis for a four-point scale for each sub-construct: establishing a culture for learning, engaging students in learning, managing student behavior, managing classroom procedures, communicating with students, using questioning and discussion techniques, creating an environment of respect and rapport, and using assessment in instruction. Further, qualitative data were collected during the classroom observation on each sub-construct listed above using a field notes method. For a description of TPOT development, see Straub and colleagues (2014).

### *ReflectLivE: After-Action-Review System*

During each TeachLivE session, the teacher's virtual rehearsal was transmitted via secure Skype video and audio connection. The transmissions were recorded and coded for pedagogical strategy analysis using ReflectLivE software. ReflectLivE is a video tagging software integrated with the TeachLivE classroom simulator that records sessions, compresses the video to a smaller format, storing all data (video and tags) on the observer's workstation. These data can then be sent over a secure network to be stored at the originating research site computer containing the TeachLivE software. During each session, videos were tagged for frequency of questions and content-related affirmation.

## **RESULTS**

### **Treatment Fidelity in the Simulator**

Fidelity checks were in place throughout the study. All teachers received the lesson plan in digital format, as evidenced by a checklist of teacher contact information at each site. The professional who controlled the five computer-generated high school avatars was trained to follow five distinct patterns of behavior aligned to common student perceptions related to Lesson 1 content, and to maintain consistent, authentic responses through sessions that would reset with each interaction. During the TeachLivE sessions, the facilitator followed a detailed procedural checklist to turn on and operate

the software for the simulation, ensuring fidelity of implementation. Fidelity in all instances was reported to be at or above 90 per cent.

### Teacher Results

Teaching practices were defined on five distinct dimensions pre- and post-intervention: (a) close-ended questions (CE), (b) open-ended questions (OE), (c) open-ended plus questions (OE+), (d) content-related affirmation (CRA), and (e) summary score on the TPOT (TPOT Sum). Maxwell's (2001) recommendation of moderate correlation (0.3–0.7) was used as a threshold for all variables to determine if it was appropriate to conduct a multivariate analysis of variance. Content-related affirmation was excluded from the analysis, because the researchers predicted no significant findings. In the case of the variables under investigation, the majority did not meet correlation thresholds, so analysis of variance (ANOVA) tests were more appropriate. See Table 2 for correlations of dependent variables.

## CLASSROOM SIMULATOR RESULTS

### Research Question 1: Differences in Performance Over Time With Simulation and Performance Feedback

To examine performance of teachers over four 10-minute sessions, a within-subjects ANOVA was performed. Time was cast as a within-subjects factor with dependent variables of CE for question 1.1, OE for question 1.2, and OE+ for question 1.3. One observer collected data during all of the TeachLivE sessions. Due to the novel nature of the intervention (e.g., dearth of group design research identified on simulation in teacher education), an alpha level of .10 was established to judge statistical significance. Partial eta squared was used to interpret effect size rather than eta squared because a multifactor design was used (Pierce, Block, & Aguinis, 2004) to be able to compare effects across different factorial designs used in the study (Levine & Hullet, 2002).

#### Question 1.1: CE Questions in Simulator

After each session, teachers were presented with data verbally and on a large display on CE questions, but no performance goals were set for subsequent sessions. Analysis was conducted with a within-subjects design ANOVA. Mauchly's test of sphericity indicated that the assumption of sphericity had not been violated,  $X^2(5) = 6.772, p = .238$ . Results indicated a significant time effect ( $F(3,87) = 3.710, p = .015, \eta^2 p = .113$ ). Pallant (2007) recommends interpreting partial eta squared using Cohen's (1988) guidelines for eta squared effect size: small (.01), medium (.06), or large (.14). Mean scores decreased

Table 2. Correlations of dependent variables

		CE Pre	CE Post	OE Pre	OE Post	OE+ Pre	OE+ Post	TPOT Sum Pre	TPOT Sum Post
CE Pre	<i>r</i>	1	.244*	.117	.183	-.214*	.051	.111	.281
CE Post	<i>r</i>	.244*	1	.213*	.380**	.125	.183	.153	.183
OE Pre	<i>r</i>	.117	.213*	1	.238*	.275**	.157	.351**	.207
OE Post	<i>r</i>	.183	.380**	.238*	1	-.061	.094	.207*	.263
OE+ Pre	<i>r</i>	-.214*	.125	.275**	-.061	1	.141	.127	.031
OE+ Post	<i>r</i>	.051	.183	.157	.094	.141	1	.198*	.200
TPOT Pre	<i>r</i>	.111	.153	.351**	.207*	.127	.198*	1	.649**
TPOT Post	<i>r</i>	.281	.183	.207	.263	.031	.200	.649**	1

significantly at each session, which was expected because, although feedback on performance was given for CE questions, teachers were focusing on increasing a replacement behavior of CE to that of OE questioning. See Table 3 for mean CE questions across 10-minute TeachLivE sessions.

**Table 3. Mean CE questions across 10-minute TeachLivE sessions**

		TeachLivE Sessions			
		Session 1	Session 2	Session 3	Session 4
	<i>n</i>	M ( <i>SD</i> )	M ( <i>SD</i> )	M ( <i>SD</i> )	M ( <i>SD</i> )
Total	30	10.23 (6.1)	9.6 (6.2)	8.7 (3.9)	6.7 (5.4)

**Question 1.2: OE Questions in Simulator**

Teachers were primarily attempting to increase their frequency of OE questions and the same process was used for question 1.1. Analysis was conducted with a within-subjects design ANOVA. Mauchly’s test of sphericity indicated that the assumption of sphericity was violated,  $X^2(5) = 93.798, p = .000$ . Epsilon ( $\epsilon$ ) was 0.387, as calculated according to Greenhouse and Geisser (1959), and was used to correct the ANOVA. Results indicated no significant time effect ( $F(1.162, 33.694) = .320, p = .609, \eta^2p = .011$ ). Mean scores are displayed in Table 4.

**Table 4. Mean OE questions across 10-minute TeachLivE session**

		TeachLivE Sessions			
		Session 1	Session 2	Session 3	Session 4
	<i>n</i>	M ( <i>SD</i> )	M ( <i>SD</i> )	M ( <i>SD</i> )	M ( <i>SD</i> )
Total	30	16.65 (17.7)	15.37 (6.4)	17.5 (7.4)	17.5 (7.7)

**Question 1.3: OE+ Questions in Simulator**

As a specific subset of OE questions, OE + questions also were measured. After each session, teachers were presented with OE+ data verbally and on a large display, and a definition for OE+ questions was read aloud; however, performance goals were not set for subsequent sessions because OE+ questions are part of a larger category of OE questions. Analysis was conducted with a within-subjects design ANOVA. Mauchly’s test of sphericity indicated that the assumption of sphericity was violated,  $X^2(5) = 86.024, p = .000$ . Epsilon ( $\epsilon$ ) was 0.795, as calculated according to Greenhouse and Geisser (1959), and was used to correct the ANOVA. Results indicated a significant time effect ( $F(2.385, 69.178) = 4.789, p = .008, \eta^2p = .142$ ). Mean scores are displayed in Table 5.

**Table 5. Mean OE+ questions across 10-minute TeachLivE sessions**

		TeachLivE Sessions			
		Session 1	Session 2	Session 3	Session 4
	<i>n</i>	M ( <i>SD</i> )	M ( <i>SD</i> )	M ( <i>SD</i> )	M ( <i>SD</i> )
Total	30	.73 (1.7)	.77 (1.0)	1.27 (2.1)	2.1 (2.1)

## Research Question 2: Differences in Performance Over Time With Simulation and No Feedback On Performance

All of the above research questions were designed to investigate how providing performance feedback in after-action-review of simulation would impact teacher practice in a classroom simulator. Researchers for question 2 evaluated the effects of withholding feedback on a specific teacher practice (i.e., frequency of content-related affirmation) in a classroom simulator. To examine performance of teachers over four 10-minute sessions, a within-subjects ANOVA was performed. Time (four sessions) was cast as a within-subjects factor with a dependent variable of CRA. After each session, teachers were not presented with any data related to CRA. One observer collected data on frequency of OE questions asked by teachers per TeachLivE session. Analysis was conducted with a within-subjects design ANOVA. Mauchly’s test of sphericity indicated that the assumption of sphericity was violated,  $X^2(5) = 16.138, p = .006$ . Epsilon ( $\epsilon$ ) was 0.718, as calculated according to Greenhouse and Geisser (1959), and was used to correct the ANOVA. Results indicated no significant time effect ( $F(2.153, 62.43) = .455, p = .651, \eta^2p = .015$ ), which was expected, because no feedback had been provided. Mean scores are displayed in Table 6.

Table 6. Mean CRA across 10-minute TeachLivE Sessions

		TeachLivE Sessions			
		Session 1	Session 2	Session 3	Session 4
	<i>n</i>	M (SD)	M (SD)	M (SD)	M (SD)
Total	30	5.13 (4.0)	5.33 (3.2)	5.73 (5.7)	4.57 (3.6)

## Classroom Results

To examine impact of simulation of teachers in a real classroom with students present, the next research questions were designed to evaluate teacher performance on variables that had been part of the after-action-review in the simulator. Research question 3.1 evaluates teacher performance on a general measure of teacher practice (TPOT), while more specific practices were evaluated in questions 3.2 (CE), 3.3 (OE), and 3.4 (OE+).

## Research Question 3: Classroom Results of Simulation with Feedback Performance

### Question 3.1: TPOT Sum

An observer collected data on the TPOT Sum and two observers observed 30% of classes to establish inter-rater reliability. While performance feedback was not given using the TPOT Sum score as the measurement instrument, the score is considered to be a general measure of teacher performance. Reliability of scores between observers during both observations was calculated (pre-intervention,  $r = .932$ ; post-intervention,  $r = .882$ ). Results from an ANOVA indicated no statistically significant changes in TPOT Sum scores between Observation 1 and 2 based on treatment group ( $F(1,94) = .097, p = .757, \eta^2p = .001$ ). For main effects, no statistically significant difference was found between the first and second observation collapsed across treatment groups ( $F(1,94) = 1.460, p = .230, \eta^2p = .015$ ), nor between groups collapsed across observations ( $F(1,94) = .555, p = .458, \eta^2p = .006$ ). Mean TPOT scores are displayed in Table 7 and TPOT scores over time are displayed in Table 8.

### Question 3.2: CE Questions

An observer collected data on CE questions asked by the teacher, and two observers observed 30% of classes to establish inter-rater reliability. Reliability of scores between observers during both

Table 7. Mean TPOT scores over time

	<i>n</i>	Observations	
		1	2
		M ( <i>SD</i> )	M ( <i>SD</i> )
Comparison	46	20.43 (6.2)	19.89 (6.3)
TeachLivE	50	21.46 (6.1)	20.54 (6.3)
Total	96	20.97 (6.2)	20.23 (6.3)

observations was calculated (pre-intervention,  $r = .929$ ; post-intervention,  $r = .798$ ). Results from a mixed ANOVA indicated there were no statistically significant changes in frequency of CE questions between Observation 1 and 2 based on treatment group ( $F(1,100) = .796, p = .374, \eta^2p = .008$ ). See Table 8 for mean frequency of CE questions over time by group.

Table 8. Mean CE questions over time by group

	<i>n</i>	Observations	
		1	2
		M ( <i>SD</i> )	M ( <i>SD</i> )
Comparison	50	8.56 (6.0)	7.54 (4.3)
TeachLivE	52	7.87 (4.6)	7.98 (5.7)
Total	102	8.21 (5.3)	7.76 (5.1)

### Question 3.3: OE Questions

An observer collected data on the number of OE questions asked by the teacher, and two observers 30% of classes to establish inter-rater reliability. Reliability of scores between observers during both observations was calculated (pre-intervention,  $r = .864$ ; post-intervention,  $r = .954$ ). Results from a mixed ANOVA indicated there were not statistically significant changes in frequency of questions between Observation 1 and 2 based on treatment group ( $F(1,100) = 1.299, p = .257, \eta^2p = .013$ ). To determine the difference between groups at each level of time and vice versa, separate ANOVAs were run. There was no significant difference between treatment groups at Observation 1 ( $F(1,102) = 1.079, p = .301, \eta^2p = .010$ ) or Observation 2 ( $F(1,100) = .194, p = .661, \eta^2p = .002$ ). When comparing main effects over time by group, for the Comparison group, OE questions were not statistically significantly different between observations ( $F(1,49) = .282, p = .598, \eta^2p = .006$ ). However, for the TeachLivE group, OE questions were statistically significantly different between observations ( $F(1,51) = 4.403, p = .041, \eta^2p = .079$ ), with teachers increasing OE questions from Observation 1 ( $M = 9.81, SD = 5.83$ ) to Observation 2 ( $M = 12.35, SD = 7.62$ ). See Table 9 for mean frequency of OE questions over time by group.

### Question 3.4: OE+ Questions

An observer collected data on OE+ questions asked by the teacher, and two observers observed 30% of classes to establish inter-rater reliability. Reliability of scores between observers during both observations was calculated (pre-intervention,  $r = .586$ ; post-intervention,  $r = .792$ ). Results from a mixed ANOVA indicated there were statistically significant changes in frequency of OE questions

Table 9. Mean OE questions over time by group

	<i>n</i>	Observations	
		1	2
		M ( <i>SD</i> )	M ( <i>SD</i> )
Comparison	50	11.12 (7.3)	11.74 (6.2)
TeachLivE	52	9.81 (5.8)	12.35 (7.6)
Total	102	10.45 (6.6)	12.05 (6.9)

between Observation 1 and 2 based on treatment group ( $F(1,100) = 2.223, p = .030, \eta^2p = .046$ ). To determine the difference between groups at each level of time and vice versa, separate ANOVAs were run. There was no significant difference between treatment groups at Observation 1 ( $F(1,102) = 2.402, p = .124, \eta^2p = .023$ ) or Observation 2 ( $F(1,100) = 1.699, p = .195, \eta^2p = .017$ ). For the Comparison group, OE+ was statistically significantly different between observations ( $F(1,49) = 5.512, p = .023, \eta^2p = .101$ ), with teachers significantly decreasing their OE+ from Observation 1 ( $M = .96, SD = 1.93$ ) to Observation 2 ( $M = .36, SD = .69$ ). For the TeachLivE group, OE+ was not statistically significantly different between observations ( $F(1,51) = .323, p = .572, \eta^2p = .006$ ), although teachers increased from Observation 1 ( $M = .50, SD = .90$ ) to Observation 2 ( $M = .62, SD = 1.21$ ). See Table 10 for mean OE+ questions over time by group.

Table 10. Mean OE+ questions over time by group

	<i>n</i>	Observations	
		1	2
		M ( <i>SD</i> )	M ( <i>SD</i> )
Comparison	50	.96 (1.9)	.36 (.7)
TeachLivE	52	.50 (.9)	.62 (1.2)
Total	102	.73 (1.5)	.49 (1.0)

#### Research Question 4: Classroom Results of Simulation Without Feedback

An observer collected data on CRA asked by the teacher, and two observers observed 30% of classes to establish inter-rater reliability. Reliability of scores between observers during both observations was calculated (pre-intervention,  $r = .931$ ; post-intervention,  $r = .885$ ). Results from a mixed ANOVA indicated there were no statistically significant changes in frequency of CRA between Observation 1 and 2 based on treatment group ( $F(1,100) = .127, p = .722, \eta^2p = .001$ ). See Table 11 for mean CRA over time by group.

## DISCUSSION

All teachers, experimental and control, were observed in their classrooms teaching science content and those teachers who received TeachLivE also were observed in the classroom simulator. Results indicated that four 10-minute sessions in the TeachLivE simulator significantly improved the use of targeted teaching behaviors in the simulation scenarios, and those improvements transferred into the teachers' original classroom settings. Teachers were found to significantly increase the use of

Table 11. Mean CRA over time by group

	<i>n</i>	Observations	
		1	2
		M (SD)	M (SD)
Comparison	50	5.12 (5.4)	4.78 (3.9)
TeachLivE	52	5.42 (5.2)	4.65 (5.4)
Total	102	5.27 (5.3)	4.71 (4.7)

targeted teaching practices in the simulator (OE questions) and improvements transferred into the teachers' original classroom settings for OE questions. While there was a significant increase for OE+ questions in the simulator, those effects did not carry over to the real classroom. This finding was unanticipated. In the classroom, although teachers who received simulation increased their frequency of OE+ questions, they did not do so significantly in their actual practice with real students. It is interesting to note that their colleagues who did not receive simulation significantly decreased their frequency of OE+ questions from observation 1 to 2. Both groups were observed teaching the same lesson, so it is possible that the differences in performance can be attributed to practice in the simulator. Hattie and Timperley (2007) indicated that the impact of feedback was largest when given relative to performance on a specific task with low complexity. It is possible that the feedback model resulted in less impact on OE+ questions because of the level of complexity. It is also possible that teachers who received simulation focused on too many performance objectives and this resulted in a challenge to learning, reflected in the classroom when they only significantly changed practice related to one variable (OE questions). Consequently, the feedback model should be investigated to determine the best approach for impacting performance.

No significant difference in performance was observed in classrooms when participants did not set performance objectives on a variable (CE questions). Looking back to teacher performance in the simulator, CE questions decreased significantly, which was expected, because teachers were focusing on increasing an opposing behavior of OE questions. Finally, consistent with Phase I, when teachers were not provided with feedback on a variable (CRA), no significant differences in performance were observed. This finding also is not surprising, as many researchers indicate setting objectives and providing feedback are essential components to improving teacher performance (Hattie & Timperley, 2007). Our work underscores the importance of providing a structured after-action-review that takes into account best practices for providing feedback on performance. When teachers did not receive data on their performances, they did not change their practice.

As a whole, results add to emerging research in the field suggesting professional learning in virtual and mixed-reality simulated classrooms can be effective. We found simulation did increase teachers' frequency of OE questions and that this increase also was observed in their classrooms. Teachers who took part in a series of sessions (virtual rehearsal) significantly increased their instances of OE questions in the simulator, similar to studies conducted earlier (e.g., Dawson & Lignugaris/Kraft, 2013; Elford et al., 2013; Vince-Garland et al.; 2012), and their performance in OE questions increased significantly in comparison to colleagues who did not receive simulation. A limitation to these findings is the virtual rehearsal of teaching a lesson, but this treatment effect is part of using simulation and could not be eliminated. Also, in past research having a control group that just reviewed lessons but did not spend time in the simulator was found to negatively impact performance (Straub et al., 2014); hence this explains our reasoning for using only two groups, an experimental and a control group. Overall, findings provided support the hypothesis that teachers who engage in professional development in simulation can improve their pedagogical practices and these changes did transfer back into their practice.

## Limitations

Of the 11 research sites, in only one was the simulator sessions delivered in the school setting. At this site, a researcher brought the simulation equipment to the teachers' school using a mobile unit. This took a significant amount of coordination between technology staff on the research team and with the district, as the software requires specific network settings. At the other 10 sites, teachers traveled to simulation sites located at institutes of higher education. Teachers receiving simulation were required to visit the simulator three times, which required significant scheduling efforts in the cases of last-minute cancellations or delays resulting from technology issues. Cancellations due to travel were not an issue for the mobile lab; however, new barriers to scheduling arose, as teachers were more likely to run late to sessions as they tried to juggle on-site job duties. Future research should explore the idea of school district-level coordination for professional learning, so teachers do not have competing demands for their attention.

A second limitation is the type of PD occurring in each school site was not known or made available to the research team. Due to the national scope of this study additional skill development could be aligned to onsite PD. However, random assignment of teachers did occur to try and control for any onsite PD effects.

## Future Research and Implications

The next steps in simulation research are to evaluate the impact of varying session lengths, frequency of sessions, and total number of sessions to determine the optimal level of treatment needed to produce both immediate and sustained changes in teachers' behaviors. The question of dosage is critical to unlocking the benefits of simulation for busy teachers and school districts with limited financial resources; simulation has the potential to deliver professional learning in an accelerated format and in a compressed amount of time. Identifying the components of effective simulation could save valuable time and money. In light of findings, the team currently has three areas of unanswered questions related to time. First, if four 10-minute sessions impact practice, how long does this change in practice continue? Second, do teachers need to re-visit the simulator to rehearse once a month, a semester, or a year to regain the previous level of performance? Third, what is the optimum length of a session? Will shorter sessions still result in a significant increase in performance?

Our findings have important implications for researchers and educators designing simulation activities, as simulation with no feedback and no opportunity for informed reflection is likely a waste of resources. The researchers found that teachers did not change their performance when after-action-review was absent. However, future research should explore the aspects of after-action-review to determine the most effective model of applying after-action-review.

Just as models of feedback need further investigation, so do methods of grouping teacher participants in the simulator. For this project teachers attended sessions individually, resulting in the costliest use of the simulator. Would it be more effective for teachers to attend simulation activities in small groups (lesson studies or even Professional Learning Communities) with the twofold benefit of capitalizing on the social nature of learning and cost-savings? While questions surrounding group versus individual models of training were not explored in this project, there is much interest in the impact of these competing formats since understanding their relative benefits can further inform the field as more simulation technology is used for teacher learning.

Our future work will explore the use of simulation with adult avatars for other educational professionals such as administrators, guidance counselors, psychologists, and counselor educators. Use of simulation for parent-teacher conferences can provide invaluable experiences for professionals who likely did not engage often with parents during their teaching experiences or at any time during their preparation programs. Simulation can provide a safe practice ground, so individuals can learn from mistakes without harming relationships with parents. In the area of Counselor Education, counseling techniques can be provided with avatars to practice challenging interactions. Because teachers report that the mixed-reality simulation feels authentic to teaching, the same feelings of authentic presence



will likely be found with other avatar interactions. Beyond educational professionals' use of simulation, we are highly interested in how simulation might impact student learning, especially in critical area in the cross-cutting concepts outlined in the Next Generation Science Standards (2013). We also are just beginning to investigate social skill interactions for students in working in cooperative groups and looking at how the simulator might be used to help teachers and students with autism and intellectual disabilities in inclusive settings. As we explore the next generation of tools for teachers' professional and personalized learning, we keep in mind that the ultimate goal of the research team is that the simulator does not replace "real" teaching, but instead allows for safe rehearsal that is targeted and personalized. The vision is simulators can prepare and potentially retool the skills of teachers and other education professionals at all levels from pre-service to in-service where interactions are complex and nuanced like the act of teaching science.

## REFERENCES

- Andreasen, J., & Haciomeroglu, E. (2009). Teacher training in virtual environments. *Paper presented at the annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Academic Press.
- Bacolor, R., Cook-Endres, T., Lee, T., & Allen, A. (2014). *How can I get my students to learn science by productively talking with each other?* Stem Teaching Tools. Retrieved from <http://stemteachingtools.org/brief/6>
- Barmaki, R. (2016). *Gesture Assessment of Teachers in an Immersive Rehearsal Environment* [PhD]. University of Central Florida.
- Bartos, S. A., & Lederman, N. G. (2014). Teachers' knowledge structures for nature of science and scientific inquiry: Conceptions and classroom practice. *Journal of Research in Science Teaching*, *51*(9), 1150–1184. doi:10.1002/tea.21168
- Brown, J., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, *18*(1), 32–42. doi:10.3102/0013189X018001032
- Buckridge, H. (2016). *Mixed reality experiences in the M. Ed. educational leadership program: Student perceptions* [PhD]. University of Central Florida.
- Bukaty, C. (2016). *Innovative facilitation of requisite communication skills employment using mixed reality simulation to prepare young adults to problem solve in the workplace* [Dissertation]. University of Central Florida.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Common Core Standards Initiative. (2011). *Preparing America's students for college and career*. Retrieved from <http://www.corestandards.org>
- Danielson, C. (2011). The Framework for Teaching Evaluation Instrument. In T. D. Group (Ed.), *The Danielson Group*. Princeton, NJ.
- Dawson, M., & Lignugaris/Kraft, B. (2013). TLE TeachLivE™ vs. role-play: Comparative effects on special educators' acquisition of basic teaching skills.
- Dieker, L. A., Hughes, C. E., Hynes, M. C., & Straub, C. (2017). Using simulated virtual environments to improve teacher performance. *School University Partnership*, *10*(3), 62–81.
- Dieker, L. A., Rodriguez, J., Lignugaris-Kraft, B., Hynes, M., & Hughes, C. E. (2014a). The future of simulated environments in teacher education: Current potential and future possibilities. *Teacher Education and Special Education*, *37*(1), 21–33. doi:10.1177/0888406413512683
- Dieker, L. A., Straub, C., Hughes, C., Hynes, M. C., & Hardin, S. E. (2014b). Learning from virtual students. *Educational Leadership*, *71*(8), 54–58.
- Elford, M., Carter, R., & Aronin, S. (2013). Virtual reality check: Teachers use bug-in-ear coaching to practice feedback techniques with student avatars. *Journal of Staff Development*, *34*(1), 40–43.
- Elford, M., James, S., & Haynes-Smith, H. (2013). Literacy instruction for pre-service educators in virtual learning environments. In A. Hayes, S. Hardin, L. Dieker et al. (Eds.), *Conference Proceedings for First National TeachLivE Conference*. Academic Press.
- Fullan, M., & Langworthy, M. (2013). *Toward a New End: New Pedagogies for Deep Learning*. Seattle, WA: Collaborative Impact.
- Gallegos, B. (2016). *The Role of Virtual Avatars in Supporting Middle School Students with Learning Disabilities from Culturally and Linguistically Diverse Backgrounds in After School Programs on Science* [PhD]. University of Central Florida.
- Goodwin, M. (2007). Occasioned knowledge exploration in family interaction. *Discourse & Society*, *18*(1), 93–110. doi:10.1177/0957926507069459

- Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. doi:10.3102/003465430298487
- Hawkins, B. (2007). Open-endedness, the instructional conversation and the activity system: How might they come together? In R. Alanen & S. Poyhonen (Eds.), *Language in Action: Vygotsky and Leontievian Legacy Today* (pp. 245–279). Cambridge: Cambridge Scholars Publishing.
- Hodson, D. (2014). Learning science, learning about science, doing science: Different goals demand different learning methods. *International Journal of Science Education*, 36(15), 2534–2553. doi:10.1080/09500693.2014.899722
- Kane, T., & Staiger, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill & Melinda Gates Foundation.
- Levine, T., & Hullett, C. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research*, 28(4), 612–625. doi:10.1111/j.1468-2958.2002.tb00828.x
- Literacy Design Collaborative. (2015). *Mini-tasks*. Retrieved from <http://ldc.org/how-ldc-works/mini-tasks>
- Loewenberg Ball, D., & Forzani, F. (2010). Teaching skillful teaching. *Educational Leadership*, 68(4), 40–45.
- Loewenberg Ball, D., Sleep, L., Boerst, T. A., & Bass, H. (2009). Combining the development of practice and the practice of development in teacher education. *The Elementary School Journal*, 109(5), 458–474. doi:10.1086/596996
- Maxwell, S. (2001). When to use MANOVA and significant MANOVAs and insignificant ANOVAs or vice versa. *Journal of Consumer Psychology*, 10(1-2), 29–30.
- McPherson, R., Tyler-Wood, T., McEnturff, A., & Peak, P. (2011). Using a computerized classroom simulation to prepare re-service teachers. *Journal of Technology and Teacher Education*, 19(1), 93–110.
- National Center for Education Statistics. (2012). *The Nation's Report Card: Science 2011 (NCES 2012–465)*. U.S. Department of Education.
- National Center for Education Statistics. (2013). *NAEP questions tool*. Retrieved from <http://nces.ed.gov/nationsreportcard/itmrlsx/search.aspx?subject=mathematics>
- National Economic Council, Council of Economic Advisers, and Office of Science and Technology Policy. (2011). *A strategy for American innovation: Securing our economic growth and prosperity*. Washington, D.C.: The White House.
- National Institutes of Health. (2005). NIH Curriculum Supplement Series “Using Technology to Study Cellular and Molecular Biology.” Department of Bioethics, NIH Clinical Center.
- National Science Teachers Association. (2014). *Making Connections to the Common Core*. Retrieved from <http://ngss.nsta.org/making-connections-common-core.aspx>
- Neal, M. (2008). Look who’s talking: Discourse analysis, discussion, and initiation-response-evaluation patterns in the college classroom. *Teaching English in the Two-Year College*, 35(3), 272–281.
- NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. Washington, DC: The National Academies Press.
- Pallant, J. (2007). *SPSS Survival Manual*. New York: McGraw-Hill Education.
- Pierce, C., Block, R., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and Psychological Measurement*, 64(6), 916–924. doi:10.1177/0013164404264848
- Reiser, B., Berland, L., & Kenyon, L. (2012). Engaging students in the scientific practices of explanation and argumentation. *Science Scope*, 35(8), 6–11.

Straub, C., Dieker, L., Hynes, M., & Hughes, C. (2014). *Using virtual rehearsal in TLE TeachLivE™ mixed reality classroom simulator to determine the effects on the performance of mathematics teachers. 2014 TeachLive National Research Project: Year 1 Findings*. Orlando, FL: University of Central Florida.

Teaching Works. (2014). High Leverage Practices. Retrieved from <http://www.teachingworks.org/work-of-teaching/high-leverage-practices#sthash.I4xK7DG4.dpuf>

Tytler, R., & Aranda, G. (2015). Expert teachers' discursive moves in science classroom interactive talk. *International Journal of Science and Mathematics Education, 13*(2), 425–446. doi:10.1007/s10763-015-9617-6

Vince Garland, K., Vasquez, E., & Pearl, C. (2012). Efficacy of individualized coaching in a virtual classroom for increasing teachers' fidelity of implementation of discrete trial teaching. *Education and Training in Autism and Developmental Disabilities, 47*(4), 502–515.

Waring, H. (2009). Moving out of IRF (initiation-response-feedback): A single case analysis. *Language Learning, 59*(4), 796–824. doi:10.1111/j.1467-9922.2009.00526.x

Whitten, E., Enicks, A., Wallace, L., & Morgan, D. (2013). Study of a mixed reality virtual environment used to increase teacher effectiveness in a pre-service preparation program. In A. Hayes, S. Hardin, L. Dieker et al. (Eds.), *Conference Proceedings for First National TeachLivE Conference*. Academic Press.

Windschitl, M., Thompson, J., Braaten, M., & Stroupe, D. (2012). Proposing a core set of instructional practices and tools for teachers of science. *Science Education, 96*(5), 878–903. doi:10.1002/sce.21027

*Lisa Dieker (PhD) is a Pegasus Professor and Lockheed Martin Eminent Scholars in the College of Community Innovation and Education at the University of Central Florida (UCF). Her research focuses on harnessing the power of teachers working across disciplines in inclusive settings in teacher education, special education, and simulation.*

*Charles E. Hughes, co-director of the Synthetic Reality Laboratory, is a Pegasus Professor of computer science at UCF. His research interests are in virtual reality, and the applications of these technologies to interpersonal skills development, teacher preparation, physician training, de-escalation skills, and protective strategies for self and others.*

*Taylor Bousfield is an Exceptional Education Teacher for Orange County Public Schools and an Adjunct Professor in the College of Community Innovation and Education at the University of Central Florida. She holds a Ph.D. in special education and currently teaches secondary students with intellectual disabilities. Her research interests lie in the area of innovative teacher preparation for teaching students with low incidence disabilities. Specifically, preparing teachers in a community based model and using virtual reality.*

*Samantha Mrstik (PhD) is an assistant professor of curriculum and instruction at Georgia Gwinnett College. Prior to her work at Georgia Gwinnett, she obtained her Ph.D. at the University of Central Florida and has worked as a special education teacher in Florida for fifteen years.*