Learning Disabilities Research & Practice, *00*(0), 1–21

DOI: 10.1111/ldrp.12232

Effects of Multimedia versus Live Professional Development on Teachers' and Students' Performance Related to the Question Exploration Routine

Jean B. Schumaker
University of Kansas
Joseph B. Fisher
Grand Valley State University
Lisa D. Walsh
Autism Behavior and Psychological Services
Paula E. Lancaster
Central Michigan University

In each of two studies, teachers were randomly assigned to either a Virtual Workshop (VW) group that used a computerized professional development program or an Actual Workshop (AW) group that participated in face-to-face professional development, including discussion, feedback, and collaboration. In both studies, teachers' posttest scores related to their knowledge of the Question Exploration Routine and their plans for using it were significantly higher than their pretest scores. In Study 2, both groups' posttest scores with regard to implementation and planning of the routine were significantly greater than their pretest scores. There were no significant differences between the groups at posttesting on any measure. The posttest knowledge scores of the whole groups of students and the subgroups of students with LD being taught by both groups of teachers were significantly higher than their pretest scores. All teachers indicated that they were satisfied with the training and the routine. VW teachers in both studies indicated that they were satisfied with the software program.

Over the past 20 years, researchers have been evaluating alternatives to live, face-to-face professional development sessions for teachers (e.g., Griffin & Brownell, 2018). One reason for these efforts is that face-to-face sessions are costly. Few school districts have the funds to bring in highly skilled professional development experts for the amounts of time required to create lasting change in a school (e.g., Blanchard et al., 2016; Elges et al., 2006). Additionally, such sessions are typically delivered to every teacher in a school, resulting in a "one-size-fits-all" type of effort. In addition, few professional development experts are available, and one may be more skilled and knowledgeable than another. Additionally, many professional development efforts may be futile because teachers who have been trained in new practices may leave the district, while additional funds are not available to train the new personnel on a catch-up basis. Finally, new empirically validated instructional practices are being developed regularly, and live professional-development efforts face constant challenges related to keeping pace with these new developments. As a result, the quality of live, face-to-face professional development efforts can vary widely, and these efforts often fall short of their intended outcomes.

Reprints can be requested from Jean B. Schumaker at Edge Enterprises, Inc., P.O. Box 1304, Lawrence, KS 66044. Electronic inquiries should be sent to jschumak@edgeenterprisesinc.com.

For these reasons, technology that does not involve a live instructor has been recommended for replacing, or at least becoming a component of, face-to-face professional development sessions (Bates et al., 2016; Collins & Liang, 2015; Edinger, 2017; Geiman, 2011; Wei et al., 2010). Technology has the advantage of being available constantly; its quality is generally stable; it can be used for years by new teachers joining a school, and by teachers in remote locations; it can be used by different types of teachers as appropriate, and according to individual needs; and it is affordable (Dede et al., 2009). Nevertheless, technology can be impersonal and can rely on specific operation system requirements. Additionally, some of the benefits of face-to-face professional development (e.g., discussions, a sense of learning community, a presenter who can address confusion and questions) can potentially be lost.

RESEARCH ON USING TECHNOLOGY FOR PROFESSIONAL DEVELOPMENT

Some research has been conducted on the use of technology without a live instructor. Kennedy et al. (2017), for example, employed a package of components, including *podcasts*, to teach teachers about evidence-based practices related to science vocabulary instruction. Specifically, their instructional

package included (a) a PowerPoint slide show with narration, (b) video clips of an instructor modeling instructional practices with no students present, (c) sample instructional materials, and (d) written feedback from an instructor. The multiple-baseline across-teachers design showed that the percentage of time spent in vocabulary instruction and the number of practices used with fidelity increased after instruction for all three teachers. Unfortunately, interobserver reliability was not reported for the practices measure. A social validity questionnaire, however, showed that the teachers were positive about the intervention.

More recently, Peeples et al. (2018) compared the same methods created for the Kennedy et al. (2017) study to a lecture condition and a condition where teachers simply read an article. All 200 teachers received written feedback after observations, but the podcast group received performance feedback based on the Kennedy observation system. All three groups of teachers improved their performance regarding teaching vocabulary, but the podcast group spent significantly more time using the practices than the other groups. Unfortunately, student learning was not measured, so whether this extra time resulted in more learning is not known. Furthermore, the different feedback methods (and not the podcasts) might have been a confounding factor.

An alternative to the podcast approach is the use of a *multimedia software program* that includes interactive screens, which present a variety of learning experiences to the learner. This software format requires the learner to interact with the software program constantly and provides feedback to the learner regarding comprehension and application of the content. Narrated and animated text is displayed on the screen, and some screens also include a video clip in an actual classroom, showing a teacher and students interacting to demonstrate a component of the practice. Some screens include quiz questions the learner has to answer correctly before moving on. Still other sections provide many examples of the whole intervention being used in elementary, middle-, and high-school classrooms in different subject areas (e.g., science, social studies). Some sections of the program give the learner an assignment to complete by filling out a digital planning diagram for a lesson and then printing it out. In other words, the learner progresses through this interactive program by answering questions, reviewing video clips and lesson plans that align with the grade level and subject area the teacher is teaching, and completing planning activities digitally.

The team of researchers that developed the type of software program described above has conducted several studies in which two groups of teachers taking a college course were randomly selected to either (a) use the software program (the "Virtual Workshop" [VW] group) or (b) receive live, face-to-face instruction (the "Actual Workshop" [AW] group). Pretest/posttest control-group designs and multiple-baseline across-teachers designs were utilized. Each study focused on instructing the participants to implement a single Content Enhancement Routine which had previously been validated as being effective in improving student learning in inclusive general education subject-area courses (see Bulgren & Schumaker, 2006; Schumaker & Deshler, 2010; and Schumaker et al., 2002 for reviews). The routines chosen for these studies were specifically designed to enhance the learning of *all* students in inclusive general education classes. Each routine was designed to be used by teachers to co-construct with students an understanding of an important abstract concept (e.g., "*democracy,"* "*biological weapon*," "*tragedy"*), or to compare and contrast concepts (e.g., "*biological weapon*" with "*chemical weapon*").

As an example of one of the projects focusing on multimedia PD, Schumaker et al. (2010) focused on the Concept Comparison Routine (Bulgren et al., 1995, 2002). This routine is used while teaching students how to compare and contrast two concepts. Interestingly, in Study 1, the VW *preservice* teachers earned significantly higher knowledge test scores than those in the AW *preservice* group. In Study 2, the VW *inservice* teachers earned significantly higher scores on their performance of the routine in the classroom than did the AW *inservice* teachers. Otherwise, the scores of teachers in the VW and AW groups in both studies were equivalent. When Schumaker et al. (2010) disaggregated the results for 73 students with disabilities from the results of the whole group of 292 students, they showed significant gains for the student LD groups associated with both teacher groups.

Thus, as a group, these studies on podcasts and multimedia instruction provide support for the effectiveness of technology as a tool for training teachers. They have shown that teachers' knowledge, planning, and practice can change significantly after taking part in some type of computerized instructional experience. In a few instances, the computerized training produced teacher knowledge and performance scores that were significantly higher than the face-to-face training. Because Schumaker et al. (2010) disaggregated the learning of students with disabilities from that of the whole group of students, the results have shown that students who have the greatest learning deficits (Warner et al., 1980), as well as other students, can benefit from the instruction when their teachers are taught through technology.

NEED FOR ADDITIONAL RESEARCH

Nevertheless, some issues have been raised about technology-driven PD (Hill et al., 2013). For example, some authors (e.g., Darling-Hammond et al., 2017; Desimone, 2009) have held that effective PD must include several features (e.g., content focus, active learning, modeling). Unfortunately, some of the recommended features (e.g., sustained duration, live feedback, collaboration, coaching) cannot be incorporated into podcasts and multimedia programs. It stands to reason that technology-driven programs might be discounted because of this very fact. Some researchers, however, have questioned these assumptions based on recent studies wherein the effects of instruction of teachers trained through PD programs containing the recommended core features were compared to the effects of instruction of control-group teachers. Garet et al. (2016), for example, employed the recommended PD features of professional learning communities and coaching, along with other features, in a study with 221 fourth-grade math teachers. Although teacher knowledge and some behavior

in the classroom changed, student achievement did not. Garet et al. (2011) also provided PD with sustained duration, a content focus, active learning, and feedback. Some teachers received coaching, and others did not. Although teacher knowledge and some instructional practices improved in both teacher groups, no differences were found between student pretests and posttests for either group of teachers.

Thus, because of the lack of clarity about the types of PD features that are required to produce both teacher and student learning and the possible issues that might be raised about computerized programs, the current project was designed to take a closer look at a couple of recommended PD features: (a) live interaction with and feedback from an instructor and (b) collaboration. These components were selected because they could be easily incorporated into the short PD sessions. A decision was made not to include coaching as a component of the PD since Garet et al. (2011) found no added value when coaching was included. Additionally, funding was not available to pay coaches as well as to conduct a component analysis with a group of teachers who received an IM program plus coaching, another group who learned through an IM program but did not receive coaching, as well as two control groups of teachers who received live instruction with and without coaching.

An empirically validated Content Enhancement Routine, called the Question Exploration Routine (Bulgren et al., 2001), was chosen as the focus of PD in this project for several reasons. First, this routine has been shown to enable substantially larger numbers of students with LD, low achievers, and average achievers to earn passing grades on tests when it is used as opposed to when it is not used (Bulgren et al., 2011; Bulgren et al., 2009). Second, the routine has not been the focus of PD studies in the past. Third, the routine involves the complex process of answering a major course question, which requires teachers and students to analyze the question, create and sequence a series of subquestions, create a main idea answer to the course question, and apply learned knowledge to new situations. Whether teachers could learn about such a complex cognitive process via technology is unknown. Third, previous studies on the Question Exploration Routine involved either the researcher instructing the students on information organized by herself into a graphic device (Bulgren et al., 2009; Bulgren et al., 2011) or two 9th-grade ELA teachers instructing students in two lessons on *Romeo and Juliet* (Shakespeare, 1992), which had been created for them by other teachers (Bulgren et al., 2013). Whether teachers teaching a variety of grade levels in a variety of subject areas can be taught to organize their own information into such a graphic device and implement the routine in a way that enhances student learning for a variety of students is unknown. Additionally, because teachers have not been asked to provide ratings of the software and the routine in previous studies, whether they will be satisfied with the software program and the routine, once they have used it in the classroom, is not clear. Finally, because students have not been asked for ratings, whether they will react differently, given that their teachers were trained differently, is unknown. Thus, the research questions for the current studies are as follows.

1. Do teachers who learn about the Question Exploration Routine in a live workshop involving collaboration and feedback earn significantly higher scores on measures of knowledge, preparation, and implementation than teachers who learn about it through a software program? (Studies 1 and 2)
2. Do students of teachers who learn about the routine in a live workshop earn significantly higher knowledge scores and rate their teachers' instruction more highly than students of teachers who learn about it through a software program? (Study 2)
3. Do teachers who learn about the routine in a live workshop rate their training and the routine more highly than teachers who learn about it through a software program? (Study 2)

STUDY 1: METHODS

Participants

A total of 20 certified teachers, who were enrolled in a college course as part of a retraining program for individuals seeking certification as special education teachers, participated; they volunteered and provided written informed consent.[1] Ten teachers were randomly assigned to the experimental group (hereafter referred to as the "Virtual Workshop [VW] group"). Eight were females, and two were males. Ages ranged from 22 to 50 years ($M = 34$ years); they had an average of 5 years of teaching experience. Ten teachers served in the alternate treatment group (hereafter referred to as the "Actual Workshop [AW] group"). Nine were females, and one was male. Ages ranged from 25 to 44 years ($M = 31$ years); they had an average of 6 years of teaching experience.

Settings

The Virtual Workshop (VW) took place in a university computer lab, outfitted with 25 desktop computers arranged in rows. Also present were an instructor's computer, a data projector, a screen, and a whiteboard. The Actual Workshop (AW) took place in a university classroom furnished with tables and chairs. Also present were an instructor's computer, a digital versatile disc (DVD) player, a data projector, a document camera, a screen, and a whiteboard.

The Instructional Practice

The instructional practice for which all the teachers received professional development was the Question Exploration Routine (QER) (Bulgren et al., 2001), an empirically validated inclusive practice (Bulgren et al., 2009; Bulgren et al., 2011; Bulgren et al., 2013) designed to improve the educational outcomes of diverse groups of students enrolled in inclusive general education, subject-area instruction. The routine involves a systematic step-by-step procedure for facilitating class discussion and using strategies to answer a critical course question (e.g., "*How does the*

deforestation of the rainforest in South America contribute to the greenhouse effect?"). The procedure is designed to make the information related to the critical question more explicit and accessible to students, especially those with special needs (Bulgren et al., 2006). The teacher follows a set of steps built into the routine[2] to introduce the critical question and model cognitive strategies while thinking aloud. Over time, as the routine is used many times, students are engaged in a cognitive apprenticeship related to answering difficult course questions. For example, through the teacher modeling aloud, they learn how to "unpack" a larger critical question into smaller questions. In other words, they are guided to brainstorm some focused, supporting questions (e.g., "*What's currently happening to rainforests*?" "*What results from the burning of rainforests*?," "*What is the effect of increased CO2 in the Earth's atmosphere*?"), sequence them logically, and develop corresponding answers. The answers are then summarized to build an accurate and concise main-idea answer collaboratively (e.g., "*When rainforests are burned, the resulting increase in CO2 contributes to the greenhouse effect*"). Information derived during the routine is recorded by the teacher and the students on their own personal copies of a *graphic organizer* called the Question Exploration Guide (QEG) (see Figure 1[3]). The teacher's copy is displayed for all the students to see, and the teacher models how to fill it in during the discussion. As a result, students walk away from the lesson with a permanent record of the discussion, a record that they can use as a study guide for tests and as the basis for essays and other written products. Although the teacher creates a draft of the guide before class, the final version of the QEG is created by the teacher and students in partnership throughout the discussion.

The Professional Development Programs

The Virtual Workshop

The Virtual Workshop (VW) was composed of a multimedia software program (Schumaker et al., 2007), which had been downloaded onto computers. Participants worked through the program individually. If they had a technical problem, they could request help, but they could not discuss the content of the program with anyone. The program has six lessons: (a) an introduction to the QER; (b) information about the parts of the QEG, the graphic organizer; (c) information and video clips about each part of the routine, with associated quizzes to check understanding; (d) video clips of teachers at the elementary-, middle-, and high-school levels using the whole routine in their classrooms in relation to different subject areas; (e) example QEGs for science, social studies, language arts, and elective courses; and (f) digital activities where the user can practice creating QEGs for assigned topics and receive feedback.

With regard to the screen layout, each section title is listed in a table of contents displayed along the bottom of the computer screen. Using a mouse, the user can select any section title at any time and as often as wanted by "clicking" upon it. After clicking, information about a given section is displayed in the top three quarters of the computer screen. Every screen contains buttons which a teacher can use to progress to the next screen, return to the previous screen, have the narrator repeat the same message, increase or decrease the volume of the narrator's voice, or exit the program. When a teacher exits the program, the program provides a pass code that the teacher can use to return to the same screen in the future. By following the narrator's instructions and using the on-screen buttons, a teacher can easily progress in a linear fashion through the whole program with no outside support.

Each lesson contains specific information about the QER in the form of audio, video, animated graphics, text, or combinations of these media. For example, during the explanation of the QEG, one section of the QEG is displayed on the screen at a time, and the types of information to be displayed in that section are described. Examples are shown. During the explanation of each step of the routine, text on the left-hand side of the screen describes, in bulleted form, the behaviors in which a teacher should engage. A narrator covers these bulleted items. A rectangle on the right side of the same screen contains a video clip, which a teacher can play using a play button (and other buttons that rewind, replay, and stop the video clip) to see a teacher in a classroom using the step in partnership with students. Audio controls are also available. On the quiz screens, a question and related multiple-choice answers are displayed, and the narrator states the question and the answers. The teacher then can choose the correct answer by clicking on it. Immediately, the teacher sees and hears positive praise for correct answers, or a statement that the answer was not correct. The teacher is then given an instruction to answer the question again, and the question screen is reshown. During the digital activities related to the creation of QEGs, three QEGs are created: one on the Civil War, one on Content Enhancement Routines, and one on a topic chosen by the teacher. The teacher is given a handout containing information for the first two QEGs. The teacher is instructed to read the handout. Then the software program provides the teacher with optional pieces of information that can be chosen (by clicking) to be placed in each section of the QEG. For the final QEG(s), the teacher can choose a topic and begin creating QEGs for lessons in the classroom. A certificate of completion can be printed as each lesson is completed.

The Actual Workshop

In contrast, the Actual Workshop (AW) was a live workshop that was designed to teach participants about the QER. It was comprised of the same content as the VW, corresponding to the table of contents for the VW. A PowerPoint slide was created for the text from each screen of the VW, and the content was presented orally by the session leader.[4] Additionally, all of the video segments in the software program were stored on a DVD and played by the instructor in coordination with the PowerPoint slides. The same multiple-choice questions, sample QEGs, and practice activities used in the VW were included in a paper packet distributed to AW participants. The instructor led discussions and guided participants to examine and discuss items and complete practice activities

Question Exploration Guide

1. What is the Critical Question?

How did states and framers of the U.S. Constitution compromise when several states would not ratify the Constitution?

2. What are the Key Terms and explanations?

Constitution	A document that explains how a government is set-up/organized
Compromise	An agreement between groups in which each group gives up something to gain something
Ratify	Approve

3. What are the Supporting Questions and answers?

Why did some states object to the Constitution?	People felt it did not protect the rights of individuals.
Which rights needed to be protected?	1. Freedom of religion, speech, press, assembly, and petition. 2. Protection from unlawful searches, arbitrary arrest, and punishment.
What did the states suggest?	A Bill of Rights be added to the Constitution.
What did the framers agree to do?	To add the Bill of Rights after the Constitution was ratified.

4. What is the Main Idea Answer?

The framers agreed to add the Bill of Rights to protect individuals' rights *after* the states ratified the Constitution.

5. How can we use the Main Idea?

Assignment: Read the Bill of Rights, and find a newspaper article that shows how the rights of a person have been protected by it.

6. Is there an Overall Idea? Is there a real-world Use?

Citizen rights can be protected through use of the Bill of Rights.

FIGURE 1 Sample Question Exploration Guide.

collaboratively. Corrective feedback was provided to participants as needed. Additionally, the participants freely asked questions and received answers.

In addition to containing the same content, both workshops integrated the same known principles of effective professional development, such as video analysis of the key aspects of an instructional practice performed by an expert teacher, review of multiple video-recorded exemplars, examination of multiple teacher-prepared lesson plans, opportunities for interactive learning, multiple practice activities to check understanding and rehearse key aspects of the instructional practice, and feedback on performance (e.g., Desimone, 2009; Knight, 2004, 2007; McDonald et al., 2013; Snow-Renner & Lauer, 2005). Both groups of teachers received paper copies of the same sample QEGs and a copy of the QER instructor's manual (Bulgren et al., 2001). Both groups spent one three-hour class period in their workshop. The only difference was that the AW involved a live instructor, discussions, collaboration, immediate feedback, and answers to questions. The VW involved no live interaction.

Measures

Teacher Knowledge Test (A Measure of Teacher Learning)

This test was composed of 13 short-answer questions in an open-ended format to measure a teacher's recall and understanding of the QER's components and procedures. Teachers were allowed 15 minutes to answer the questions on this paper-and-pencil test. To score teachers' answers, written guidelines specifying acceptable responses were used. Different point values were awarded according to the number of answers required for each question. For example, if a question asked for the six steps in the routine, six blanks followed the question, and six points were available, one for each step. Teachers earned a maximum score of 24 points on the test. The percentage of points earned was calculated.

Question Exploration Guide Test (A Measure of Teacher Learning)

For this test, teachers completed a blank Question Exploration Guide (QEG) (Figure 1) (Bulgren et al., 2001) to measure their knowledge of the type of information (e.g., the critical question, key terms, definitions, supporting questions, etc.) that belongs in each section. They were given a one-page single-spaced description of information to read (pretest: *U.S. Civil War;* posttest: *Three Branches of the U.S. Federal Government*) and then were asked to fill in a blank QEG based on that information. They had an unlimited amount of time to read the document and fill in the eight sections of the guide. Each teacher had to choose a question and supporting information associated with the document and could earn a maximum score of 8 points, one point for each section of the QEG. Written guidelines specifying acceptable responses for each section were used by scorers. The percentage of points earned was calculated.

Training Satisfaction Questionnaire (A Measure of Teacher Reaction)

This questionnaire was developed to assess teachers' opinions about their workshop. Each of the 10 questionnaire items included a 7-point Likert-type scale ranging from "strongly disagree" (1) to "strongly agree" (7). Example items related to the teacher's understanding of the information and its applicability to their teaching. A mean rating was determined for each item, as well as a mean overall rating for each group.

Software Satisfaction Questionnaire (A Measure of Teacher Reaction)

This questionnaire contained 11 Likert-type scale items focused on the software program. Each scale ranged from "completely dissatisfied" (1) to "completely satisfied" (7). Items related to such topics as the narration, the video clips, the quizzes, and the time required to watch the whole program. It was administered only to the VW teacher group after completing the VW. A mean rating was calculated for each item as well as a mean overall rating.

Interscorer Reliability

Two trained scorers independently scored at least 20% of the Teacher Knowledge Tests and the Question Exploration Guide Tests administered both before and after the workshops to determine interscorer reliability. All of the tests and guides had been given ID numbers instead of names. The reliability scorer was blind to the teachers' assignment to workshops. The points awarded by the two scorers were compared item-by-item. The observers had to award the same score on a given item for an agreement to be tallied. The percentage of agreement was calculated by dividing the total number of agreements by the total number of agreements plus disagreements and multiplying by 100. For the Teacher Knowledge Test, the percentage of agreement was 91.0% (440 agreements out of 480 opportunities to agree). For the Question Exploration Guide Test, the percentage of agreement was 91.3% (146 agreements out of 160 opportunities).

Procedures

The same session leader[5] was present at both workshops. In a separate session prior to each workshop, the leader administered both the Teacher Knowledge Test and the Question Exploration Guide Test. Once each teacher's workshop was complete, he/she was given the Teacher Knowledge Test, the Question Exploration Guide Test, and the Training Satisfaction Questionnaire. The VW teachers also completed the Software Satisfaction Questionnaire.

During the VW, the session leader used the instructor-station computer, data projector, and screen to briefly demonstrate how to navigate through the software program.

Afterward, teachers turned on their computers, launched the program, and navigated the program independently for a maximum of 3 hours. An observer recorded responses to quiz questions and completion of the various parts of the program, using a checklist that listed all the parts of the software program. The observer also recorded any issues or questions raised by the participants. The session leader provided technical support with computer hardware or software. No one provided support with workshop content.

During the AW, the session leader presented the specified information, video clips, and activities, using the available equipment. The session leader presented the PowerPoint slides and played all the video segments used in the VW at appropriate times in the workshop. The teachers completed the same practice activities as teachers in the VW in paper form, but they worked together to complete the activities. In addition, the session leader provided corrective feedback. The AW lasted 3 hours. Any questions asked by participants about the QER or QEG were answered by the session leader immediately. An observer recorded the information covered by the session leader on a checklist listing all the required slides, video clips, and activities.

Experimental Designs and Data Analysis

Two experimental designs were used. A pretest/posttest control-group design (Campbell & Stanley, 1963) was used to compare the Teacher Knowledge Test scores and the Question Exploration Guide scores of participants in the AW as opposed to the VW. To compare the differences between the pretest and posttest scores within each treatment group, repeated-measures analyses of variance (ANOVAs) were performed. To determine whether the in-person AW workshop was superior to the VW workshop, analysis of covariance (ANCOVA) was employed with the teachers' posttest scores serving as the dependent variable and their pretest scores serving as the covariate. Group was the main effect of interest. A posttest-only control-group design was also used to compare the Training Satisfaction Ratings of AW and VW teachers. To determine whether the AW participants reported greater training satisfaction than the VW participants, an ANOVA was performed for each item, and for the mean overall rating.

STUDY 1: RESULTS

Workshop Fidelity Results

The observers' records showed that all of the VW teachers completed 100% of the segments and activities in the software program. Similarly, the observer's records showed that the AW leader presented all of the slides, videotapes, and activities in the AW.

Teacher Knowledge Test Results

Mean scores, standard deviations, and statistical results are displayed in the first section of Table 1 for the Knowledge Test. A repeated-measures ANOVA indicated that the posttest scores of VW teachers were significantly different from their pretest scores, as were those of the AW teachers. The effect sizes were very large. The ANCOVA revealed no difference between the posttest scores of AW and VW teachers.

Question Exploration Guide Test Results

The repeated measures ANOVAs indicated that the posttest scores of the VW teachers were significantly greater than their pretest scores, as were the posttest scores of the AW teachers. (See the second section of Table 1.) The ANCOVA showed no difference between the posttest scores of AW and VW participants.

Training Satisfaction Questionnaire Results

The overall mean ratings across all items on the Training Satisfaction Questionnaire provided by AW teachers ranged from 3.5 to 7.0 ($M = 5.30$, $SD = 0.95$). Similarly, mean ratings by VW teachers ranged from 3.9 to 6.3 ($M = 5.10$, $SD = 0.81$).[6] The ANOVAs did not reveal differences between the mean ratings for individual items, or between the mean overall ratings.

Software Satisfaction Questionnaire Results

The mean ratings of the teachers in the VW group on individual items ranged from 5.1 to 6.7 ($M = 6.27$, $SD = 0.69$). With the exception of one item, all of their mean ratings on individual items were in the satisfied to completely satisfied range. The one mean rating below 6.0 (5.10) related to the length of the software program. The participants had spent the allotted 3 hours working through the program.[7]

STUDY 2: METHODS

Purpose

Once the results of Study 1 were reviewed, the software program was revised according to the notes taken and the minor issues that had surfaced. The purpose of Study 2 was to answer the research questions with inservice teachers and their students as participants.

TABLE 1
Means Standard Deviations, and Statistical Comparisons for Studies 1 and 2

		Means		*Within-Group Effects*			*Between-Group Effects*	
Measures	*Group*	*Pre (SD)*	*Post (SD)*	*Statistic*	p-*Value*	*Effect size*	*Statistic*	p-*Value*
Study 1								
Teacher Knowledge Test	AW	1.20% (2.70)	66.70% (13.38)	$F(1,9) = 234.80$	$p < .001$*	$d = 9.69$[c]	$F(1,17) = 0.856$	$p = .368$
	VW	0.80% (1.69)	73.00% (19.39)	$F(1,9) = 149.24$	$p < .001$*	$d = 7.73$[c]		
QEG Test	AW	22.5% (19.9)	92.7% (8.6)	$F(1,18) = 104.887$	$p < .001$*	$d = 4.58$[c]	$F(1,17) = 2.369$	$p = .142$
	VW	25.1% (19.5)	93.9% (8.6)	$F(1,18) = 103.961$	$p < .001$*	$d = 4.56$[c]		
Study 2								
Teacher Knowledge Test	AW	8.33% (7.63)	80.90% (9.84)	$F(1,9) = –484.21$	$p < .0001$*	$d = 13.92$[c]	$F(1,16) = 2.17$	$p = .16$
	VW	7.90% (8.50)	86.80% (8.47)	$F(1,9) = –520.26$	$p < .0001$*	$d = 14.43$[c]		
QEG Test	AW	22.50% (19.47)	92.70% (8.72)	$F(1,9) = 180.0$	$p < .001$*	$d = 8.49$[c]	$F(1,17) = 0.03$	$p = .864$
	VW	25.10% (19.47)	93.90% (8.62)	$F(1,9) = 164.0$	$p < .001$*	$d = 8.10$[c]		
QER Implementation Checklist	AW	25.98% (6.30)	84.57% (8.27)	$F(1,9) = 216.78$	$p < .001$*	$d = 9.31$[c]	$F(1,17) = 0.245$	$p = .627$
	VW	25.24% (7.60)	86.20% (4.39)	$F(1,10) = 517.21$	$p < .001$*	$d = 13.71$[c]		
Knowledge Test: All Students	AW	10.52% (14.15)	47.14% (31.25)	$t(32.3) = 4.43$	$p < .0001$*	$d = 1.539$[c]	$F(1,20.8) = 3.28$	$p = .085$
	VW	8.94% (13.09)	63.25% (27.60)	$t(57) = 5.26$	$p < .0001$*	$d = 2.514$[c]		
Knowledge Test: Students w/ LD	AW	8.20% (14.13)	30.79% (27.80)	$t(37) = 4.678$	$p < .001$*	$d = 0.76$[b]	$F(1,31) = .528$	$p = .47$
	VW	3.94% (10.16)	48.52% (33.99)	$t(29) = 7.546$	$p < .001$*	$d = 1.0$[c]		

*Statistically significant. LD = Learning disabilities; AW = Actual Workshop; VW = Virtual Workshop.
[a] Reflects small effect size.
[b] Reflects medium effect size.
[c] Reflects large effect size.

TABLE 2
Demographic Data on All Students and Students with LD in Study 2

	All Students		*Students with LD*	
Category	*VW*	*AW*	*VW*	*AW*
Total number of students	127	116	32	38
Gender				
Male	62	64	25	24
Female	65	52	7	14
Average age	14.84 Years	14.09 Years	14.64 Years	14.54 Years
Average grade	8.8	7.7	8.6	7.5
Ethnicity				
Caucasian	80	82	20	26
African American	16	6	3	2
Hispanic	10	6	3	3
Mixed	13	17	3	5
Other	8	5	3	2
Standardized test scores				
Reading	62%ile	44%ile	41%ile	22%ile
Math	55%ile	46%ile	35%ile	20%ile

VW = Virtual Workshop; AW = Actual Workshop.

Participants

Teachers

A total of 20 teachers who were currently teaching grades 4 through 12, in a variety of subject areas, volunteered and provided written informed consent to participate. Ten teachers were randomly selected to serve in the VW group: one 4th-grade teacher, one 7th-grade teacher, three 8th-grade teachers, three 9th-grade teachers, one 9th/10th-grade teacher, and one 10th-grade teacher. Eight teachers were females, and two were males. Their ages ranged from 23 to 62 years ($M = 44$ years), and their average number of years of teaching experience was 12 years (range = 0–26 years). The 10 remaining teachers served in the AW group. They included: one 4th-grade teacher, three 8th-grade teachers, four 9th-grade teachers, one 9th/10th-grade teacher, and one 10th-grade teacher. Six teachers were females, and four were males. Ages ranged from 25 to 57 years ($M = 37$ years), and their average number of years of teaching experience was 9 years (range = 4–26 years). All Study 2 teachers were paid $250 for their participation.

Students

In addition, a total of 243 students with written permission from their parents or guardians participated (see Table 2 for demographic information). They also signed consent forms. The students were regularly enrolled in one of the 20 participating teachers' inclusive general education classes at nine schools in a metropolitan area surrounding a large Midwestern city. Twenty-nine percent of the students ($n = 70$) had learning disabilities (LD) and active Individualized Education Programs. Students were identified as having LD through Kansas or Missouri state guidelines.[8] The age of the 127 students whose teachers were in the VW group ranged from 10 to 18 years, with 37% of the students representing minority populations. Thirty-two of these students (25%) had LD. The ages of the 116 students of teachers in the AW group ranged from 9 to 17 years, with 29% representing minority populations and 33% diagnosed with LD.

Measures

Study 2 employed the same measures as Study 1, with a few additions. In Study 2, the teachers filled in a QEG each time they identified a critical question that they would be presenting during observed classes. These QEGs provided a repeated measure across time, showing whether the teachers could apply their skills to a variety of topics in a stable way and maintain their skills over time. Furthermore, four new measures were employed.

QER Implementation Checklist

Just before beginning the study, participating teachers were asked to identify critical course questions that students should be able to answer at the end of upcoming units. Then they were asked to identify class periods when they would be presenting content related to these identified questions. Trained observers used the QER Implementation Checklist while observing the teachers during those specified class periods. A total of 15 teacher behaviors (see Table 3) were assessed. If a behavior was performed, the teacher could earn 1 point for that behavior, for a total of 15 points. If a behavior was not performed, the teacher received zero points. A percentage score was calculated.

Student Knowledge Test

This test assessed students' understanding of content related to the critical question specified for the class period that was observed. This generic 13-point, open-ended short-answer

TABLE 3
Question Exploration Routine Implementation Checklist

CUE

- The teacher named the routine or the Question Exploration Guide.
- The teacher explained how the routine aids learning.
- The teacher handed out blank Question Exploration Guides.
- The teacher explained what students are to do during the routine.

DO

Ask a Critical Question

- The teacher asked a critical question and modeled writing it on the Guide.

Note and Explain Key Terms

- The teacher noted and explained key words in the question.
- The teacher elicited definitions for the key words.

Search for Supporting Questions and Answers

- The teacher prompted students to create Supporting Questions and Answers.
- The teacher elicited questions and answers from students in the process.

Work out the Main Idea Answer

- The teacher prompted students to create a Main Idea Answer.
- The teacher elicited ideas from the students.

Explore the Main Idea within a Related Area

- The teacher prompted students to explore the Main Idea within a related area and elicited their ideas.

Relate the Idea to Today's World

- The teacher prompted students to relate the Main Idea to today's world and elicited their ideas.

REVIEW

- The teacher asked questions and elicited answers to review the content of the QEG.
- The teacher asked questions and elicited answers to review the process of creating the QEG and how it aids learning.

test required students to write out their answers related to parts of the routine. The same test was used for each of the critical questions. Students were instructed to write the critical question that had been addressed by the lesson at the top of the test form, and then generic questions followed, pertaining to information related to any critical question. For example, students were asked to name and define key terms related to the critical question, to name and answer supporting questions related to the critical question, to provide an answer to the question, and to extend what they had learned to a situation beyond the classroom. The test was administered on the same day when the information related to a critical question had been presented.[9] It was administered twice: once before the teachers had received training, and once after the last observed class. Student responses were scored as correct or incorrect using written evaluation guidelines specifying acceptable responses for each item, based on the observer's notes. One point was earned for each correct answer, for a total of 13 points. A percentage score was calculated.

Routine Satisfaction Questionnaire

This questionnaire was developed to assess teachers' opinions about the QER after they had used it in the classroom, as well as how likely they would be to continue using the routine and recommend it to others. Each of the 20 questionnaire items included a 7-point Likert-type scale ranging from "strongly disagree" (1) to "strongly agree" (7), or ranging from "very unlikely" (1) to "very likely" (7). Teacher ratings were averaged as for Study 1.

Student Satisfaction Questionnaire

This questionnaire was administered to all the students in all the teachers' classes at the end of the last observed lesson after they had experienced the teacher using the QER and had completed the QEG. Data from only those students with consent who had taken both the pretest and the posttest were used. The eight Likert-type items focused on the students' reactions to the routine and the QEG. For example, students were asked whether they were satisfied with the way they could participate in the lesson and the way the QEG might help them study for tests. Again, a 7-point scale was utilized for each item, and mean ratings were calculated.

Interscorer Reliability

Interscorer reliability was determined using the same procedures as described for Study 1. For the Implementation

Checklist, the percentage of interobserver agreement was 93% (361 agreements out of 390 total opportunities for agreement). For the Teacher Knowledge Test, the scorers agreed 424 times out of 480 opportunities to agree (total percentage of agreement = 88.3%). For the Question Exploration Guides, the percentage of agreement was 92% (206 agreements out of 224 opportunities to agree). Finally, the percentage of agreement on the Student Knowledge Test was 95% (2,299 agreements out of 2,418 opportunities for agreement).

Procedures

The teachers completed their assigned workshop following the same procedures as those outlined for Study 1. Before beginning their respective workshop, however, all of the teachers were asked to identify critical questions students should be able to answer at the end of upcoming units of study. The teachers also indicated dates and times when they would teach lessons covering the content related to these questions. They were asked to complete a QEG for each lesson. Three of these lessons were observed, and teachers whose data displayed a stable trend were then trained. As long as the remaining teachers' data remained stable, they were trained after four observations. During baseline observations, trained project staff members used the Implementation Checklist and scored the QEG. After the training, the teachers were similarly observed, and the associated QEGs were scored.

After each participating teacher's first baseline observation, his/her class of students completed the Student Knowledge Test administered by a project staff member. Immediately following each participating teacher's last observed lesson, his/her class of students was administered the Student Knowledge Test and the Student Satisfaction Questionnaire.

Experimental Designs and Data Analysis

To determine the effects of the workshops on teachers' Implementation Scores and Question Exploration Guide Test scores, a multiple-baseline across-teachers design (Baer et al., 1968) was employed. Twenty teachers participated in this design, with two teachers participating in each iteration of the design. There were five iterations of the design for each teacher group (AW and VW), or 10 iterations in all. The teachers were asked to lead a discussion about a critical question at least once per unit of study. Training was provided at the appropriate time in the multiple-baseline structure. Each teacher was observed at least six times according to the teachers' own schedules. Scores were graphed for visual analysis. In addition, the scores earned during baseline were compared to the scores earned during the after-training condition, using a repeated-measures ANOVA for each group. To determine whether the AW workshop was superior to the VW workshop, an ANCOVA was used to compare the after-training scores of the two groups while using the pretest scores as the covariate. Additionally, the Nonoverlap of All Pairs (NAP) (Parker & Vannest, 2009) statistic was determined.

A pretest/posttest control-group design (Campbell & Stanley, 1963) was used to compare the Knowledge Scores and the Question Exploration Guide scores[10] of the two groups of teachers. To determine whether the pretest and posttest scores were significantly different, repeated-measures ANOVAs were used for each group. To determine whether the live workshop was superior to the VW workshop, an ANCOVA was performed, with the pretest scores serving as the covariate and the posttest scores serving as the dependent variable.

A posttest-only control-group design (Campbell & Stanley, 1963) was used to compare the satisfaction ratings of teachers participating in the VW and the AW for the Training Satisfaction Questionnaire and the Routine Satisfaction Questionnaire. To determine whether the AW workshop produced higher satisfaction, an ANOVA was performed to compare individual item ratings and the group mean rating for each group.

The pretest/posttest control-group design was also used for the students. First, means and standard deviations were calculated for descriptive variables potentially related to the outcome of interest (e.g., achievement scores, age) and compared across the groups to see whether differences existed. If differences were found between the groups, the plan was to control for those variables of interest where needed in subsequent analyses. With regard to the whole groups of students, a hierarchical linear model (HLM) analysis was conducted to test for differences between the posttest scores of the students in the AW classes and those in the VW classes. It was also used to test for differences between the pretest and posttest scores for each group. The HLM analysis was used to control for the dependency in the data because students were nested in classes.

Because there was a small number of students with LD in each class, a univariate ANCOVA was used to compare the posttest scores of the students with LD in the AW group to the posttest scores of the students with LD in the VW group. As in the HLM analyses, posttest scores served as the dependent variable, and pretest, reading, and math scores served as covariates. It was also used to compare the pretest scores to the posttest scores of students in each group separately.

STUDY 2: RESULTS

Workshop Fidelity Results

According to the collected data, all of the VW teachers completed 100% of the parts of the software program and all of the activities. The workshop leader displayed all of the required slides and video clips and conducted all of the required activities for the AW.

Teacher Knowledge Test Results

The third section of Table 1 shows the results for the Knowledge Test. Repeated measures ANOVAs indicated that the

posttest scores of VW teachers were significantly different from their pretest scores, as were those of the AW teachers. An ANCOVA did not show a significant difference between the posttest scores of AW and VW participants.

QER Implementation Checklist Results

Figures 2 and 3 show the performance of the 20 teachers on the Implementation Checklist as they implemented the QER in their classrooms. Each teacher's performance is shown both before (baseline) and after the assigned workshop (after training) as indicated by the vertical dotted line. The scores of VW participants are shown in Figure 2; those of the AW participants are shown in Figure 3. (See the fifth section of Table 1 for the means and other statistics.) Of the 30 VW lessons observed after training, all 30 exceeded the mastery level of 80%, and 25 scores of AW teachers exceeded the mastery level. The NAP revealed no overlap for 100% of the pairs for both groups. Thus, the effect size, Tau U, was 1.0 for each group. This is a large effect size. The ANOVAs indicated that the posttest scores of VW teachers were significantly different from their pretest scores, as were those of the AW teachers. No difference was found between the posttest scores of AW and VW participants.

QEG Test Results

Figures 4 (VW) and 5 (AW) display each teacher's performance both before (baseline) and after the assigned workshop (after training). After training, 30 out of 30 VW guides and 25 out of 30 AW guides met or exceeded the mastery criterion of 80%. No overlap was found for 100% of the pairs for both groups. Again, the effect size, Tau U, was 1.0 for each group. Repeated-measures ANOVAs indicated that the after-training scores of VW teachers were significantly different from their baseline scores, as were the after-training scores of AW teachers. No difference was found between the posttraining scores of AW and VW participants.

With regard to the results of the single-point QEG pretests and posttests where the participants created a QEG on an assigned topic (as in Study 1),[10] an ANCOVA did not reveal a difference between the posttest scores. According to the ANOVAs, however, each group earned significantly higher posttest scores than pretest scores. (See the fourth section of Table 1.) The effect size was large in each case.

Training Satisfaction Questionnaire Results

Mean ratings provided by VW teachers on the Training Satisfaction Questionnaire ranged from 2.60 to 6.85 ($M = 5.5$, $SD = 1.30$), and mean ratings provided by AW teachers ranged from 3.9 to 6.2 ($M = 5.5$, $SD = 0.61$). No significant differences were found between the overall mean satisfaction ratings or between the mean ratings for any individual items.[11]

Routine Satisfaction Questionnaire Results

Mean ratings for individual items on the Routine Satisfaction Questionnaire provided by AW teachers ranged from 5.0 to 6.0 ($M = 5.5$, $SD = 0.6$); for VW teachers, mean ratings ranged from 4.9 to 6.1 ($M = 5.5$, $SD = 1.3$). No significant differences were found.[12]

Software Satisfaction Questionnaire Results

The mean ratings of the teachers in the VW group related to the software program ranged from 4.63 to 6.0 ($M = 5.61$, $SD = .838$).[13] All of the mean ratings on individual items were in the satisfied range (between 5.0 and 6.0), except for one rating (4.63) related to the length of the program. The teachers indicated in written comments that the program was too long. (They had all spent the allotted 3 hours.)

Student Knowledge Test Results

When the initial analyses were conducted comparing standardized test scores and ages across the groups, significant between-group differences were found with regard to the standardized reading scores and ages for the whole groups of VW and AW students. Students in the VW group had earned significantly higher reading scores ($M = 62.21$, $SD = 28.28$) than the students in the AW group ($M = 43.98$, $SD = 27.17$) ($F\,(1, 149) = 15.58$, $p < .001$). Students in the VW group were also slightly older ($M = 14.83$ years, $SD = 1.19$) than students in the AW group ($M = 14.09$ years, $SD = 2.46$) ($F(1, 225) = 8.623$, $p = .004$). When the subgroups of students with LD were compared, differences were found with regard to standardized math scores. Those in the VW LD group earned significantly higher math scores ($M = 35.11$, $SD = 20.90$) than the students in the AW LD group ($M = 19.47$, $SD = 17.45$) ($F(1, 35) = 4.98$, $p = .032$). No significant differences were found related to gender in the whole-group comparison ($F(1, 240) = 1.101$, $p = .295$) or in the LD-subgroup comparison ($F(1, 67) = 1.631$, $p = .206$). Next, correlations between the significant predictors and posttest scores were examined. Both reading ($r = .413$, $p < .001$) and math ($r = .341$, $p < .001$) scores were significantly correlated with posttest scores, while age ($r = .047$, $p = .484$) was not. Thus, the reading and math scores were included as covariates in subsequent analyses.

Next, HLM tests were employed in which the posttest score was the outcome of interest, while Reading, Math, Pretest, Condition, and the interaction of Pretest and Condition were included as covariates. The results for the whole groups indicated that the posttest scores of students whose teachers participated in the AW were not significantly different from the posttest scores of students whose teachers participated in the VW. The ranges of scores earned by students taught by both AW teachers and VW teachers were identical, ranging from 0% to 46% on the pretest and 0% to 100% on the posttest for both groups. Mean scores and other statistics are shown in the sixth section of Table 1.

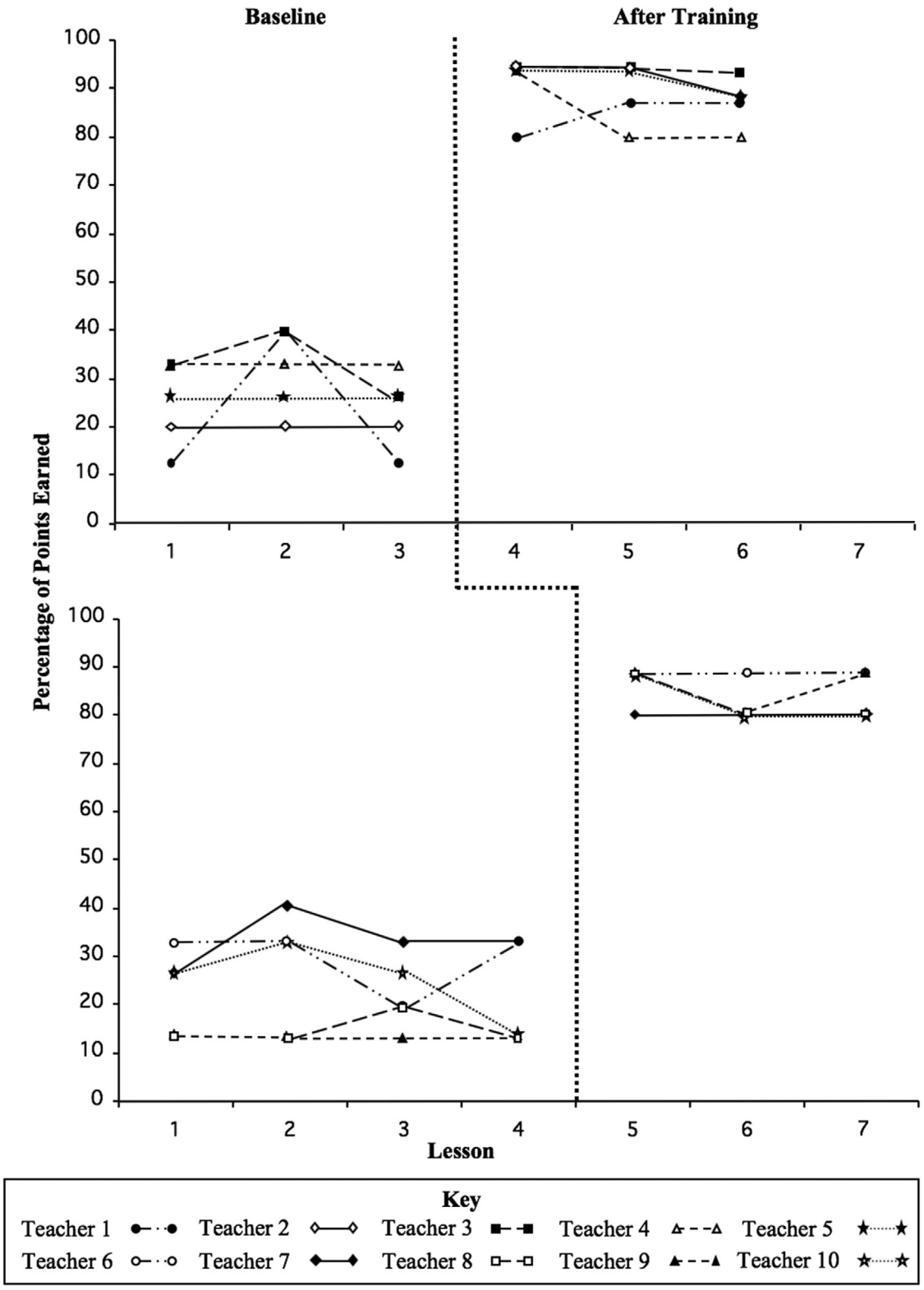

FIGURE 2 Implementation scores of teachers 1–10 (Virtual Workshop participants).

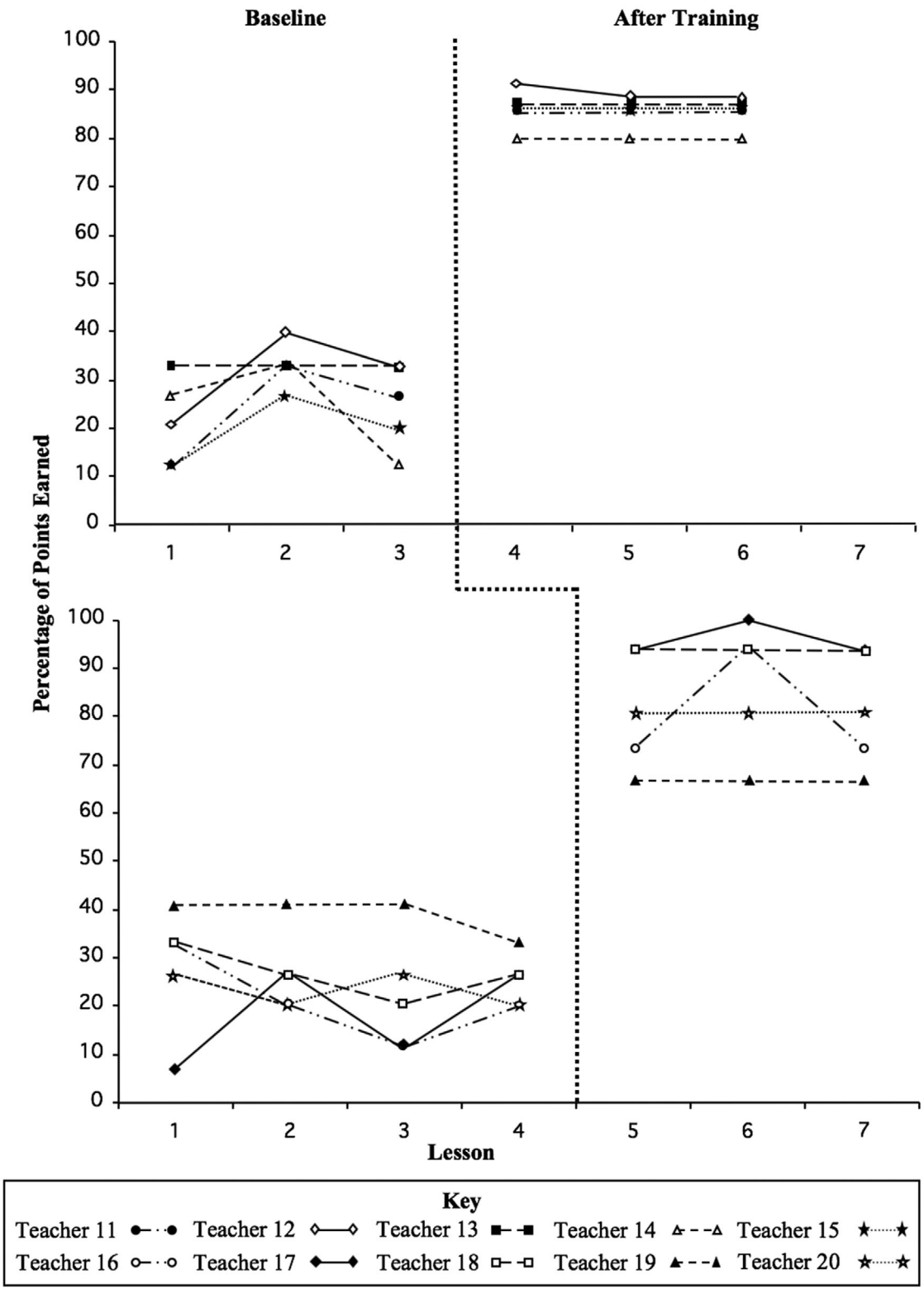

FIGURE 3 Implementation scores of teachers 11–20 (Actual Workshop participants).

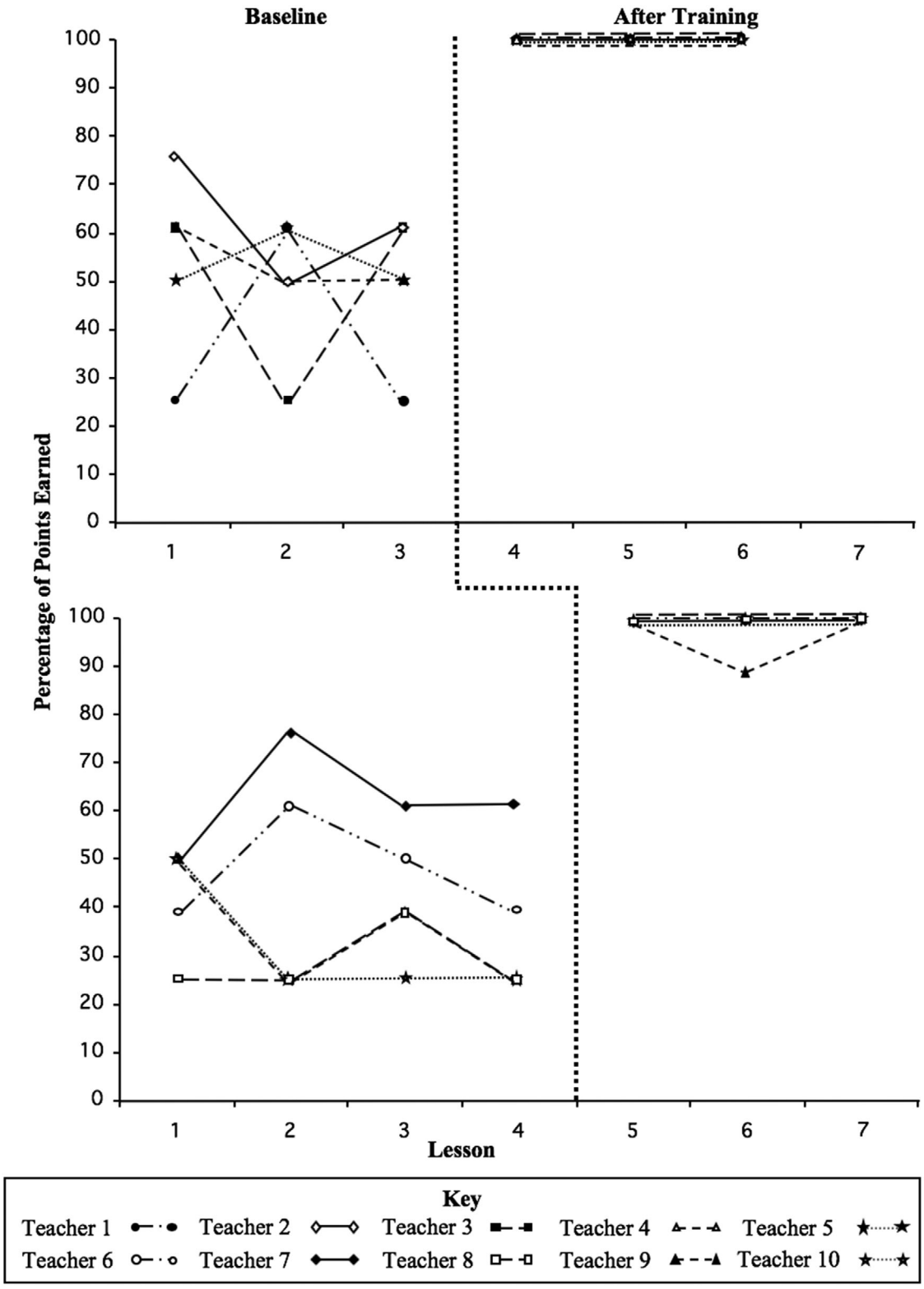

FIGURE 4 QEG test scores of teachers 1–10 (Virtual Workshop).

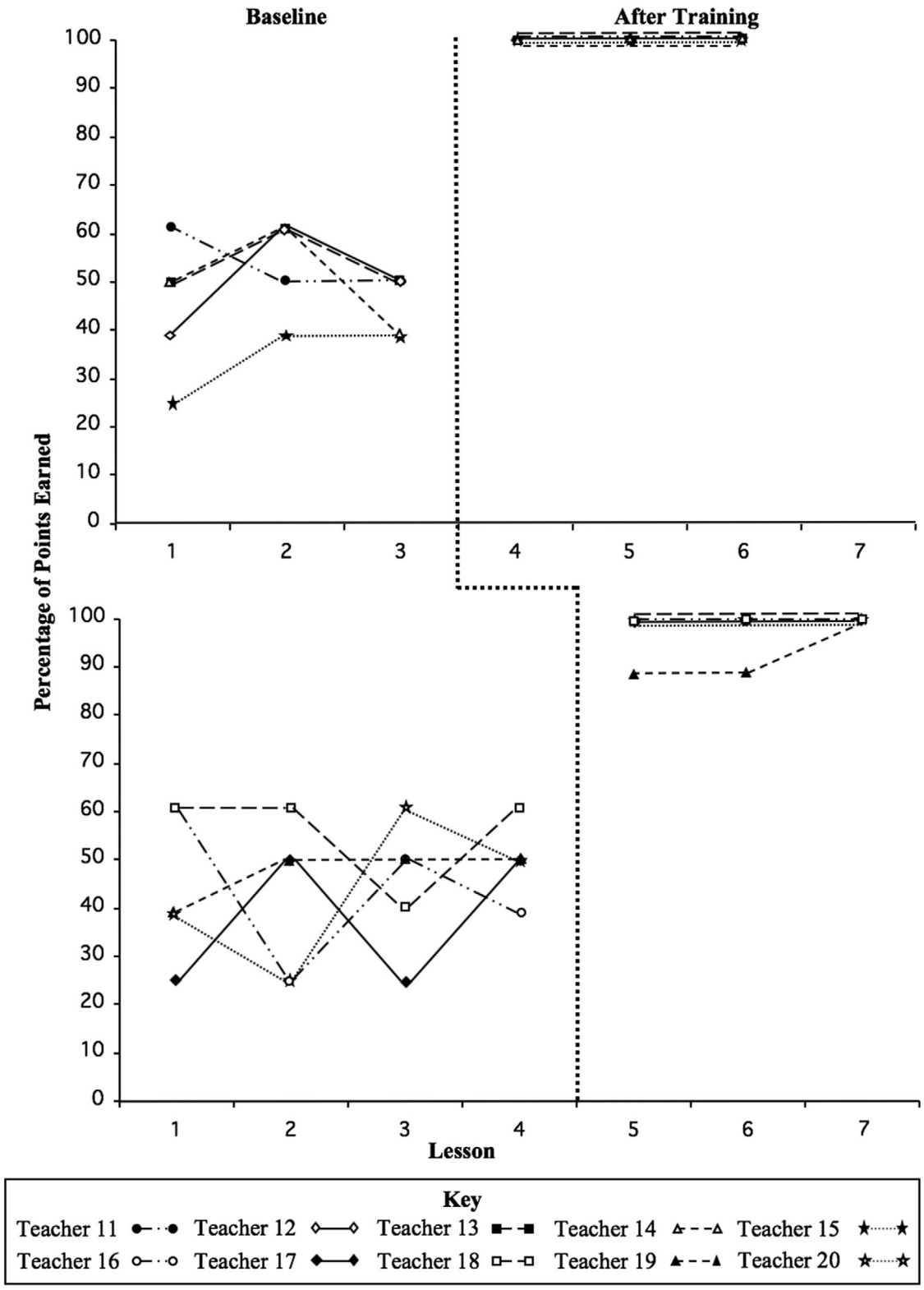

FIGURE 5 QEG test scores of teachers 11–20 (Actual Workshop).

When the groups' pretest scores were compared to their posttest scores, a significant difference was found for the AW students. Reading was a significant predictor of growth ($t(91.5) = 2.33$, $p = .022$, Cohen's $d = .487$), while math was not ($t(86.1) = 0.67$, $p = .503$). A similar difference was found for the VW students. In this case, however, math was the significant predictor of growth ($t(57) = 2.19$, $p = .033$, Cohen's $d = .580$), while reading was not ($t(57) = 1.10$, $p = .299$).

The univariate ANCOVA revealed no significant differences between the posttest scores of AW students with LD and those of the VW students with LD. Neither the reading nor the math score was related to the posttest score, although the pretest score was related to the posttest score. Students with LD made significant gains from pretest to posttest in each group. The effect size for the AW students with LD was medium; the effect size for the VW students with LD was large. (See the seventh section of Table 1 for the statistics.)

Student Satisfaction Questionnaire Results

Mean student satisfaction ratings ranged from 1.0 to 7.0 ($M = 4.8$, $SD = 1.40$) for students of AW teachers, and from 1.88 to 7.0 for students of VW teachers ($M = 5.2$, $SD = 1.24$).[14] An HLM analysis revealed no significant difference in overall satisfaction between the two groups ($F(1, 21) = 2.49$, $p = .13$). For students with LD, the mean satisfaction rating for those in the AW group was 5.1 ($SD = 1.38$), while the mean rating for those in the VW group was 5.4 ($SD = 1.18$). Due to the small number of students with LD in each class, a two-sample t-test was used for this analysis. No significant difference was found between the groups ($t = 1.130$, $p = .263$).

DISCUSSION

To summarize, the results of Studies 1 and 2 replicate and extend the results of previous studies. The present results show that when teachers are taught during a live workshop involving discussion, collaborative activities, and feedback, the results are not significantly better than the results of a multimedia learning situation with regard to teacher knowledge, teacher planning, teacher implementation in the classroom, teacher satisfaction with the training, and the learning of students with and without disabilities. The Study 2 results replicate the Study 1 results with regard to teacher knowledge, planning, and satisfaction. The pretest and posttest results for these measures mirror each other across the studies. With regard to all measures in both studies, socially and statistically significant gains were produced in all groups, representing large effect sizes for both teachers and students.

In addition, the results of these studies show that teachers can learn how to implement a multi-step instructional routine that targets complex higher-order thinking processes. The QER involves several higher order processes, including analyzing a question, defining key vocabulary, creating and sequencing subquestions and subanswers, summarizing the information to create a main-idea answer, applying knowledge to new situations, and creating an overall summary statement that pertains to the students' lives. These processes are much more abstract than those involved with naming characteristics and examples associated with a key concept, as one does when preparing for the Concept Mastery Routine (Bulgren et al., 1988, 1993) or the Concept Comparison Routine (Bulgren et al., 2002), which were the focus of previous studies.

Additionally, Study 2 represents the first time that teacher planning associated with a Content Enhancement Routine was measured repeatedly across time and across a variety of topics for teachers teaching a variety of grades and subject areas. It represents the first time teachers in a research study about the QER created their own QEGs. The study shows that teacher planning immediately improved after either live or multimedia instruction and maintained at an acceptable level throughout the remainder of the study. This result is important because it shows that the effects of the instruction did not "wear off" over time.

Moreover, this pair of studies represents the first time that teachers' reactions to multimedia software were measured. The teachers who received the software instruction indicated that they were satisfied with various aspects of the software program, except for the amount of time that was required to use it. Nevertheless, since their time was limited to 3 hours (and they could have chosen to leave earlier), and since the teachers who participated in the live training were required to remain in the classroom for 3 hours, the time the two groups spent in their respective workshops was equivalent. Thus, the specific reason for the VW teachers' lower satisfaction with the length of time spent on the software program is unclear. Perhaps sitting in front of a computer for 3 hours is more difficult than participating in a live workshop that includes interaction among the participants and an engaging live instructor. Perhaps if the teachers had been able to complete the software program on their own schedule and take breaks, they would have been more satisfied.

Another extension of this project was that the students whose teachers used the software program did not rate the QER less highly than the students of teachers who participated in the live training. Furthermore, there were no differences in satisfaction ratings across the groups of students with and without disabilities. Thus, an instructional software program can be built for teachers that not only enhances student learning but also produces student satisfaction.

An additional contribution of this line of Content Enhancement studies is that they demonstrate the components of PD that are sufficient to produce initial high-quality knowledge, planning, and implementation of a teaching routine as well as positive outcomes for students, fulfilling Kirkpatrick and Kirkpatrick's (2016) and others' models for evaluating PD (e.g., Desimone, 2009). In both the AW and VW, teachers heard an explanation of the parts of the routine, saw a variety of video-recorded models of the routine, reviewed other teachers' QEGs, and practiced planning to use the routine. None of the participants were required to read anything. The VW participants did not participate

in discussions, participate in a learning community (e.g., Treacy et al., 2002), or receive feedback. Although all of the teachers were observed teaching, they did not receive feedback on their planning or performance or instructional coaching (e.g., Knight, 2007). Nevertheless, they *were* observed; whether they would have used the routine as much as they did if an observer were not expected in their classrooms several times is not known. These findings raise some issues about the potential contribution of these highly recommended components of PD (some of which are quite expensive). Thus, further research on packages of these components and component analysis research seems warranted.

Clearly, other questions can be raised with regard to this type of professional development. First, developing software like the multimedia program used in the present project and others (e.g., Fisher et al., 1999; Fisher et al., 2010) is costly. Because professional programmers and videographers are required, along with writers and editors, the development costs can be considerable if the quality of the program is to be acceptable at professional levels. Furthermore, the software program has to be thoroughly tested and revised to ensure that it works well and is robust. How funds might become available for developing and testing these types of programs on a large scale, in order to nationally disseminate empirically validated interventions, is not known.

Second, the software program tested here was developed for an intervention that could be described in a relatively short period of time (three hours of live or computerized instruction). Other interventions might require more time to instruct, such as the curriculum that was used across a whole school year in the Fishman et al. (2013) study. Whether effective software programs can be funded and developed for more complex interventions is unknown.

Third, the long-term effects of the software program are currently unknown. Whether teachers will continue to use a routine after observers stop visiting them is unclear. Clearly, the added presence of the observers served as a prompt that each teacher needed to develop and present new lessons for the routine. Whether the teachers would continue to develop additional lessons and present them with such frequency or fidelity is unknown. Longitudinal research is certainly needed that continues to monitor teachers' use of a routine over the remainder of a school year and into the next school year.

Additionally, now that the effects of a software program alone have been determined for initial implementation of a routine, research is needed to determine whether the addition of extra elements (e.g., long-term coaching) results in longer maintenance of the intervention and more student learning across topics and units than when the software is used by itself. Component analyses are needed where some teachers receive a package of PD components (e.g., a software program + coaching) and others simply receive the software program while still others receive live instruction with and without coaching. Further, research is needed to determine student outcomes when the QER is used more often, for longer periods of time, at higher levels of quality, and across more years/courses by teachers of a given subject area (e.g., biology teachers), who work together to create coordinated QEGs and/or are coached regularly by a coach, versus when these elements are not included. Likewise, research is needed to determine the effects of including the software program and others like it into the larger frameworks of professional development programs (e.g., Desimone, 2009; Lembke et al., 2018) and teacher-training programs containing field experiences and other validated methods (e.g., Brownell et al., 2005).

Fourth, the teachers who participated in the current studies were volunteers. How teachers might react to the software program if they are being supervised by administrators in their district or if they do not have close supervision as they work through the software program is not known. Moreover, this was the teachers' first encounter with an instructional software program related to their behavior in class. Whether they might react differently if they were regularly receiving training in this way is not known. Additionally, since the teachers were teaching a variety of subjects, they chose different critical questions to be addressed as they implemented the routine. There was no way to control for the difficulty level of the content being delivered across the two groups of teachers, yet, because the teachers were randomly selected into the groups, the hope is that the content difficulty level was equalized across the groups. The questions that they chose appeared to be typical of the kinds of critical questions that might be addressed in middle-school and high-school courses. Future research should address issues related to administrator involvement in ensuring that an innovation is used and having teachers collaborate on the content that is being taught to students.

In conclusion, the results of the current studies, combined with the results of the Fisher et al. (1999, 2010) studies and the Schumaker et al. (2010) studies provide several implications for practice. One of the most difficult issues in the field of education involves disseminating empirically validated practices to educators in such a way that they are adopted and maintained. The current studies show that an innovative practice specifically designed and validated for the purpose of educating subgroups of students enrolled in general education courses (including those with disabilities) can be disseminated in this way. Since many schools are educating students with disabilities in general education courses, these studies show that the dissemination of an empirically validated practice for improving those endeavors could potentially be done in this way—that is, they show that teacher knowledge, planning, and practice can be changed in a way that results in enhanced student learning across subgroups of students. Thus, these software programs offer a new way to translate empirically validated programs into classroom practice. The time has come to move beyond comparing computerized instruction with live instruction (Fishman et al., 2014). Research is needed now to find the most effective components of computerized professional development and additional noncomputerized components that can be combined with it to enhance it. Furthermore, additional efforts are needed for translating more instructional practices and making them widely available across the nation.

ACKNOWLEDGMENTS

This research was supported by grant #R44 HD36173 awarded to Edge Enterprises, Inc., from the National Institute for Child Health and Human Development, Small Business Innovation Research Program. The authors received no financial support for the authorship of this article.

The authors wish to thank Drs. Janis A. Bulgren, Donald D. Deshler, and B. Keith Lenz for their work in the development and original testing of the Question Exploration Routine as well as the teachers, graduate school students, and students who kindly volunteered to participate in these studies. We also thank administrators in School Districts #229 and #500 in Kansas and School Districts #74 and #124 and Horizon Academy in Missouri for their schools' participation in this project.

NOTES

1. The two studies reported in this article and the procedures used to protect human subjects in both studies were approved by the Edge Enterprises Institutional Review Board, the school districts, and the granting agency.
2. For detailed information about these steps and the strategies built into them, see the instructor's manual for the Question Exploration Routine (Bulgren et al., 2001).
3. Figure 1 is printed with permission from the authors (Bulgren et al., 2001). See Bulgren et al. (2009, 2011, 2013) for more examples.
4. This person was certified as a professional development specialist by the University of Kansas Center for Research on Learning.
5. The session leader was a certified professional development specialist by the University of Kansas Center for Research on Learning.
6. A figure showing the mean ratings on individual items can be obtained from the first author.
7. A figure showing the mean ratings on individual items can be obtained from the first author.
8. According to the state guidelines in Kansas at the time of this study, a child may be determined as having a specific learning disability if the child does not demonstrate adequate achievement for the child's age or meet state-approved grade-level standards. Furthermore, a child will not be determined as having a specific learning disability if the lack of adequate achievement is primarily the result of: (a) lack of appropriate instruction, (b) limited English proficiency, (c) another disability, (d) cultural factors, or (e) environmental or economic disadvantage. In the state of Missouri, a child may be determined as having a learning disability if the child has a disorder in one or more of the basic psychological processes involved in understanding or in using language, spoken or written, a disorder which may manifest itself in an imperfect ability to listen, think, speak, read, write, spell, or to do mathematical calculations. The term includes such conditions as perceptual disabilities, brain injury, minimal brain dysfunction, dyslexia, and developmental aphasia. The term does not include learning problems that are primarily the result of the following: a visual, hearing, or motor disability; an intellectual disability; an emotional disturbance; cultural factors; environmental or economic disadvantage; or limited English proficiency.
9. Please note that all students in the participating teachers' targeted classes received the QER instruction. Nevertheless, data from only those students who were present on both the day of the pretest and the day of the posttest, and who had written consent from parents to participate, were included in the study.
10. These QEG scores were those derived from the pretest and posttest QEGs created by the teachers based on the *U.S. Civil War* and *Three Branches of Government* documents.
11. A figure showing the mean ratings for individual items can be obtained from the first author.
12. A figure showing the mean ratings for individual items can be obtained from the first author.
13. A figure showing the mean ratings for individual items can be obtained from the first author.
14. The ratings of only those students who took both the pretest and posttest are included in the satisfaction results.

REFERENCES

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, *1*, 91–97. http://doi.org/10.1901/jaba.1987.20-313

Bates, M. S., Phalen, L., & Moran, C. (2016). Online professional development: A primer. *Kappan*, *97*(5), 70–73. http://doi.org/10.1177/0031721716629662

Blanchard, M. R., LePrevost, C. E., Tolin, A. D., & Gutierrez, K. S. (2016). Investigating technology-enhanced teacher professional development in rural, high-poverty middle schools. *Educational Researcher*, *45*(3), 207–220. http://doi.org/10.3102/0013189X16644602

Brownell, M. T., Ross, D. D., Colon, E. P., & McCallum, C. L. (2005). Critical features of special education teacher preparation: A comparison with general teacher education. *The Journal of Special Education*, *38*(4), 242–252. http://doi.org/10.1177/00224669050380040601

Bulgren, J. A., Lenz, B. K., Deshler, D. D., & Schumaker, J. B. (1995). *The Concept Comparison Routine: Instructor's manual*. Edge Enterprises, Inc.

Bulgren, J. A., Lenz, B. K., Deshler, D. D., & Schumaker, J. B. (2001). *The Question Exploration Routine: Instructor's manual*. Edge Enterprises, Inc.

Bulgren, J. A., Marquis, J. G., Deshler, D. D., & Schumaker, J. B. (2013). The use and effectiveness of a Question Exploration Routine in secondary-level English language arts classrooms. *Learning Disabilities: Research and Practice*, *28*(4), 156–169. http://doi.org/10.1111/ldrp.12018

Bulgren, J. A., Marquis, J. G., Deshler, D. D., Schumaker, J. B., Lenz, B. K., Davis, B., & Grossen, B. (2006). Instructional context of inclusive secondary general education classes: Teachers' instructional roles and practices, curricular demands, and research-based practices and standards. *Learning Disabilities: A Contemporary Journal*, *4*(1), 39–66.

Bulgren, J. A., Marquis, J. G., Lenz, B. K., Deshler, D. D., & Schumaker, J. B. (2011). The effectiveness of the Question-Exploration Routine for enhancing the content learning of secondary students. *Journal*

of Educational Psychology, *103*(3), 578–593. http://doi.org/10.1037/a0023930

Bulgren, J. A., Marquis, J. G., Lenz, B. K., Schumaker, J. B., & Deshler, D. D. (2009). Effectiveness of question exploration to enhance student's written expression of content knowledge and comprehension. *Reading & Writing Quarterly*, *25*, 1–19. http://doi.org/10.1080/10573560903120813

Bulgren, J. A., & Schumaker, J. B. (2006). Teaching practices that optimize curriculum access. In D. D. Deshler & J. B. Schumaker (Eds.), *Teaching adolescents with disabilities: Accessing the general education curriculum*. Corwin Press.

Bulgren, J. A., Schumaker, J. B., & Deshler, D. D. (1988). Effectiveness of a concept teaching routine in enhancing the performance of LD students in secondary-level mainstream classes. *Learning Disability Quarterly*, *11*(1), 3–17. http://doi.org/10.2307/1511034

Bulgren, J. A., Schumaker, J. B., & Deshler, D. D. (1993). *The Concept Mastery Routine: Instructor's manual*. Edge Enterprises, Inc.

Bulgren, J. A., Schumaker, J. B., Deshler, D. D., Lenz, B. K., & Marquis, J. (2002). The use and effectiveness of a comparison routine in diverse secondary content classes. *Journal of Educational Psychology*, *94*(2), 357–371. http://doi.org/10.1037/0022-0663.94.2.356

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin.

Collins, L. J., & Liang, X. (2015). Examining high-quality online teacher professional development: Teachers' voices. *International Journal of Teacher Leadership*, *6*(1), 18–34.

Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). *Effective teacher professional development*. Learning Policy Institute. https://learningpolicyinstitute.org/product/teacher-prof-dev

Dede, C., Ketelhut, D. J., Whitehouse, P., Breit, L., & McCloskey, E. (2009). A research agenda for online teacher professional development. *Journal of Teacher Education*, *60*(1), 8–19. http://doi.org/10.1177/0022487108327554

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*(3), 181–199. http://doi.org/10.3102/0013189X08331140

Edinger, M. J. (2017). Online teacher professional development for gifted education: Examining the impact of a new pedagogical model. *Gifted Child Quarterly*, *61*(4), 300–312. http://doi.org/10.1177/0016986217722616

Elges, P., Righettini, M., & Combs, M. (2006). Professional development and recursive e-learning. *Computers in the Schools*, *23*(1), 45–57. http://doi.org/10.1300/J025v23n01_05

Fisher, J. B., Deshler, D. D., & Schumaker, J. B. (1999). The effects of an interactive multimedia program on teachers' understanding and implementation of an inclusive practice. *Learning Disability Quarterly*, *22*(2), 127–142. http://doi.org/10.2307/1511271

Fisher, J. B., Schumaker, J. B., Culbertson, J., & Deshler, D. D. (2010). Effects of a computerized professional development program on teacher and student outcomes. *Journal of Teacher Education*, *61*(4), 302–312. http://doi.org/10.1177/0022487110369556

Fishman, B., Konstantopoulos, S., Kubitskey, B. W., Vath, R., Park, G., Johnson, H., & Edelson, D. C. (2013). Comparing the impact of online and face-to-face professional development in the context of curriculum implementation. *Journal of Teacher Education*, *64*(5), 426–438. http://doi.org/10.1177/0022487113494413

Fishman, B., Konstantopoulos, S., Kubitskey, B. W., Vath, R., Park, G., Johnson, H., & Edelson, D. (2014). The future of professional development will be designed, not discovered: Response to Moon, Passmore, Reiser, and Michaels, "Beyond comparisons of online versus face-to-face PD." *Journal of Teacher Education*, *65*(3), 261–264. http://doi.org/10.1177/0022487113518440

Garet, M. S., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., & Yang, R. (2016). *Focusing on mathematical knowledge: The impact of content-intensive teacher professional development*. National Center for Education Evaluation and Regional Assistance. http://ies.ed.gov/ncee

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Walters, K., Song, M., & Hurlburt, S. (2011). *Middle school mathematics professional development impact study: Findings after the first year of implementation*. National Center for Education Evaluation and Regional Assistance. http://ies.ed.gov/ncee

Geiman, D. (2011). Online training: A high-quality, cost-effective solution. *Corrections Today*, *73*(2). 14, 16–17.

Griffin, C. C., & Brownell, M. T. (2018). The science of teacher professional development: Iterative design studies across content areas. *Teacher Education and Special Education*, *41*(2), 101–105. http://doi.org/10.1177/0888406417751032

Hill, H. C., Beisiegel, M., & Jacob, R. (2013). Professional development research: Consensus, crossroads, and challenges. *Educational Researcher*, *42*(9), 476–487. http://doi.org/10.3102/0013189X13512674

Kirkpatrick, D. L., & Kirkpatrick, W. K. (2016). *Kirkpatrick's four levels of training evaluation*. ATD Press.

Kennedy, M. J., Rodgers, W. J., Romig, J. E., Lloyd, J. W., & Brownell, M. T. (2017). Effects of a multimedia professional development package on inclusive science teachers' vocabulary instruction. *Journal of Teacher Education*, *68*(2), 213–230. http://doi.org/10.1177/0022487116687554

Knight, J. (2004). Instructional coaches make progress through partnership. *Journal of Staff Development*, *25*(2), 32–37.

Knight, J. (2007). *Instructional coaching: A partnership approach to improving instruction*. Corwin Press.

Lembke, E. S., McMaster, K. L., Smith, R. A., Allen, A., Brandes, D., & Wagner, K. (2018). Professional development for data-based instruction in early writing: Tools, learning and collaborative support. *Teacher Education and Special Education*, *41*(2), 106–120. http://doi.org/10.1177/0888406417730112

McDonald, M., Kazemi, E., & Kavanagh, S. S. (2013). Core practice and pedagogies of teacher education: A call for a common language and collective activity. *Journal of Teacher Education*, *64*(5), 378–386. http://doi.org/10.1177/0022487113493807

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy*, *40*, 357–367.

Peeples, K. N., Hirsch, S. E., Gardner, S. J., Keeley, R. G., Sherrow, B. L., McKenzie, J. M., Kennedy, M. J. (2018). Using multimedia instruction and performance feedback to improve preservice teachers' vocabulary instruction. *Teacher Education and Special Education*, *42*(3), 227–245. http://doi.org/10.1177/0888406418801913

Schumaker, J. B., Bulgren, J. B., Lenz, B. K., & Deshler, D. D. (2007). *Multimedia professional development program for the Question Exploration Routine*. Edge Enterprises, Inc.

Schumaker, J. B., & Deshler, D. D. (2010). Using a tiered intervention model in secondary schools to improve academic outcomes in subject-area courses. In M. R. Shinn & H M. Walker (Eds.), *Interventions for achievement and behavior problems in a three-tier model including RTI* (pp. 609–632). National Association of School Psychologists.

Schumaker, J. B., Deshler, D. D., & McKnight, P. (2002). Ensuring success in the secondary general education curriculum through the use of teaching routines. In G. Stover, M. R. Shinn, & H. M. Walker (Eds.), *Interventions for achievement and behavior problems* (pp. 791–824). National Association of School Psychologists.

Schumaker, J. B., Fisher, J. B., & Walsh, L. D. (2010). The effects of a computerized professional development program on teachers and students with and without disabilities in secondary general education classes. *Learning Disability Quarterly*, *33*(2), 111–131. http://doi.org/10.1177/073194871003300204

Shakespeare, W. (1992). *Romeo and Juliet*. Simon & Schuster Paperbacks.

Snow-Renner, R., & Lauer, P. A. (2005). *McREL insights: Professional development analysis*. Mid-Continent Research for Education and Learning.

Treacy, B., Kleiman, G., & Peterson, K. (2002). Successful online professional development. *Learning and Leading with Technology*, *30*(1), 42–47.

Wei, R. C., Darling-Hammond, L., & Adamson, F. (2010). *Professional development in the United States: Trends and challenges*. National Staff Development Council.

Warner, M. M., Schumaker, J. B., Alley, G. R., & Deshler, D. D. (1980). Learning disabled adolescents in the public schools: Are they different from other low achievers? *Exceptional Education Quarterly*, *1*(2), 27–35. Reprinted in the *Mainstreamed Library: Issues, Ideas, Innovations* (American Library Association), 1982. http://doi.org/10.1177/074193258000100207

About the Authors

Jean Bragg Schumaker is professor emeritus at the University of Kansas (KU) and president of Edge Enterprises, Inc. Her doctoral degree from KU is in developmental and child psychology. Retired from the KU Center for Research on Learning and the Department of Special Education, she devotes her time to developing new instructional programs for students and teachers. She is a developer of the *Content Enhancement Routines* and the *Learning Strategies Curriculum*. Her research interests focus on the use of strategic instruction for students who need intensive and explicit instruction.

Joseph B. Fisher, PhD, is a professor of special education in the College of Education at Grand Valley State University. He completed his doctoral training with Jean B. Schumaker and Donald D. Deshler at the University of Kansas Center for Research on Learning. He is a former middle-school special education teacher, and he continues to provide instruction to children with disabilities in K–12 schools alongside his graduate and undergraduate students. He has a particular interest in the research and development of instructional practices to improve the reading and writing abilities of struggling learners.

Lisa D. Walsh, PhD, is the owner and a licensed clinical psychotherapist at Autism Behavior and Psychological Services. She received her doctoral degree in developmental and child psychology at the University of Kansas and did her postdoctoral training as part of the Leadership Education in Neurodevelopmental and Related Disabilities Program at KU Medical Center in the Center for Child Health and Development. She specializes in the treatment and diagnosis of autism spectrum disorders and other behavioral disorders. In addition, she provides individual and family counseling and group therapy to children, adolescents, and adults.

Paula E. Lancaster, PhD, is the dean of the College of Education and Human Services at Central Michigan University. She earned her PhD in special education from the University of Kansas while serving as a doctoral fellow at the Center for Research on Learning. Her research has focused on effective instruction and instructional technology for adolescents with disabilities. More recently, her work is centered on the development and implementation of practice-based approaches to teacher preparation with a particular interest in collaboration between university-based preparation programs and P-12 school environments.