Learning: Research and Practice

ISSN: 2373-5082 (Print) 2373-5090 (Online) Journal homepage: https://www.tandfonline.com/loi/rlrp20

Investigating an intervention to support computer simulation use in whole-class teaching

Nico Rutten, Wouter R. van Joolingen & Jan T. van der Veen

To cite this article: Nico Rutten, Wouter R. van Joolingen & Jan T. van der Veen (2016) Investigating an intervention to support computer simulation use in whole-class teaching, Learning: Research and Practice, 2:1, 27-43, DOI: 10.1080/23735082.2016.1140222

To link to this article: https://doi.org/10.1080/23735082.2016.1140222

Learning: Research and Practice, 2016
Vol. 2, No. 1, 27–43, http://dx.doi.org/10.1080/23735082.2016.1140222

Investigating an intervention to support computer simulation use in whole-class teaching

Nico Rutten [a]*, Wouter R. van Joolingen [a] and Jan T. van der Veen[b]

[a]*Freudenthal Institute for Science and Mathematics Education, Utrecht University, P.O. Box 85170, 3508 AD Utrecht, The Netherlands;* [b]*ELAN Teacher Education and Science Communication, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands*

(*Received 4 August 2015; accepted 6 January 2016*)

Going beyond simply measuring the effectiveness of a teaching approach with computer simulations during whole-class science instruction, we investigated the interaction between teachers and their students as well as searched for mechanisms in the pedagogical context related to teachers' implementation of the intervention. Our quasi-experimental design involved having five teachers teach Newtonian mechanics with computer simulations to parallel classes of their upper secondary students. In the "Accustomed" condition the teacher decided how the lesson would unfold; in the experimental condition the lesson unfolded according to a pattern designed for "Peer Instruction". We investigated the pedagogical interaction between teachers and their students, which was expected to be affected by the intervention's support for the teacher as well by the teacher's support for the students. Learning effects as revealed by gains from pretest to posttest to delayed posttest did not consistently favour either condition. Identified mechanisms occurring in the pedagogical context that could explain our findings include: teacher's sense of ownership of the lesson, familiarity with the intervention conditions, and resistance to change. Suggestions for future research related to the identified mechanisms are offered.

Keywords: classroom studies; computer simulations; quasi-experimental research; secondary education; teaching/learning strategies

Introduction

Research over the past decade on the learning effects of computer simulations in science education has shown that computer simulations can improve the effectiveness of instruction (Rutten, van Joolingen, & van der Veen, 2012; Smetana & Bell, 2012). The use of computer simulations allows students to learn in the same way that scientists conduct their research: predict what will happen with a phenomenon in a certain situation; then investigate what actually happens in that case, and afterwards draw conclusions about why the phenomenon developed in the way that it did (de Jong & van Joolingen, 1998). The research of the past decade has primarily focused on learning with simulations by individual students or students in small groups (Rutten et al., 2012; Smetana & Bell, 2012). In such settings, it is relatively easy to influence and control students' learning processes while they are carrying out inquiry learning tasks. It is not a given that such results will transfer from these small-scale settings to learning during whole-class instruction, which typically involves one teacher and 20 to 30 students. Whereas the learning

*Corresponding author. Email: n.p.g.rutten@uu.nl

processes in small-scale settings are usually supported and controlled from within the simulation or by using additional teaching materials, during whole-class instruction an important role falls to the teacher. In particular, the teacher can act as a facilitator of knowledge construction, and can encourage the students to discuss and reflect on their learning (Khan, 2011; Maeng, Mulvey, Smetana, & Bell, 2013; Smetana & Bell, 2014). In line with constructivist and social constructivist perspectives (Krajcik et al., 1998), such a teaching approach allows for leveraging the effectiveness of learning by inquiry, which is a central component of science learning (Hofstein & Lunetta, 2004). A previous study has shown that teaching with computer simulations at the whole-class level using an inquiry-based approach relates positively to students' attitudes about the contribution of teaching with computer simulations to their motivation and insight (Rutten, van der Veen, & van Joolingen, 2015). Attitude towards a given instructional medium and instructional approach is an important factor influencing the degree of conceptual change (Pyatt & Sims, 2012; Trundle & Bell, 2010).

For these reasons, it is worthwhile to study in more detail how to support whole-class teaching for inquiry learning with computer simulations. In such studies, it is important to address both the teacher and the students. Implementing inquiry-based learning with simulations at the whole-class level requires commitment from the teacher, who plays a pivotal role in orchestrating the learning processes that are essential for this mode of learning. Teachers can help trigger the relevant learning processes by asking questions, by providing feedback, or by providing other means of support. Providing structures and examples for performing these triggering actions can assist teachers in their support of these inquiry processes. However, the properties of these types of support for teachers and students in the whole-class inquiry-based learning context are not well known.

In order to determine its effectiveness, any means of supporting teachers in their support of inquiry-based learning can be evaluated on three distinct levels related to the actors within the pedagogical context. First, the effect on teacher behaviour should be checked: does a support measure affect the behaviour of the teachers in terms of their performance of supportive actions directed towards the students? Second, the effects of the teachers' behaviour on the learning activities of the students need to be studied. These two aspects together form the pedagogical interaction in the classroom. Finally, the effects of this pedagogical interaction on the students' final learning outcomes in terms of test results, grades, or other measures should be determined. All three levels are important for understanding in full the impact that supports for inquiry-based instruction can have on the learning processes in classrooms. This means that in evaluating the effect of teaching methods, we need to address both the teacher and the students, in addition to the method used. After all, it is not the method itself that provides the instruction, but the teacher enacting the method. This means that evaluating a method as such is not possible. In evaluating instructional methods and approaches, one always evaluates teachers and methods in tandem. This is true for two reasons:

- The teacher usually takes, and should take, liberties with the way the method is implemented. The teacher is not a programmed agent simply performing a fixed script.
- The baseline to which the method is compared varies from teacher to teacher, as they vary in style, preference, interaction with the class, and other aspects that could also be relevant.

We present a study in which teachers are supported in stimulating their students to engage in inquiry processes using computer simulations, in the context of whole-class use of computer simulations during science instruction. The method we used to determine the effectiveness of the support provided takes into account both reasons mentioned above, by comparing each teacher's baseline instruction to a parallel lesson with a parallel class in the same context, in which we experimentally intervened. In our study we used Peer Instruction (Crouch & Mazur, 2001) as a method for engaging students in the processes relevant for inquiry and compared it to teachers' accustomed way of teaching with simulations. Peer Instruction can be considered as a suitable approach for having students participate more actively, as it contains built-in phases during which the students are in control of their learning processes (Crouch, Fagen, Callan, & Mazur, 2004). Active approaches to learning can increase student performance in science, engineering, and mathematics (Freeman et al., 2014). According to Zingaro and Porter (2014), the nascent *Peer Instruction* literature focuses on the actual Peer Instruction phase by measuring the learning effects of students convincing each other. In our study we explicitly included the teacher in the loop of inducing and evaluating episodes of Peer Instruction. By focusing only on learning measured directly after the Peer Instruction phase, it is possible to overlook the subsequent teacher-led discussion in which the teacher can fulfil a complementary role as facilitator of knowledge construction within the Peer Instruction learning process. Therefore, scores for the correctness of answers to questions posed directly after the Peer Instruction phase can be considered an underestimate of student learning (Zingaro & Porter, 2014). We assume that combining simulation-based inquiry learning with Peer Instruction will create lessons with alternating episodes of teacher control and student control, which can benefit the teachers' support of learning processes as well as the students' active engagement in inquiry learning processes.

In our study we used a quasi-experimental design for studying aspects of the pedagogical interaction and its effects on learning outcomes in six pairs of classes, taught by five different teachers. This made it possible to study the impact of the teacher support in different contexts, allowing us to understand its effectiveness or lack thereof not only at the level of outcome, but also at the level of the interaction between teacher and students. This study focused on improving our understanding about pedagogical interaction during whole-class teaching involving computer simulations, specifically on the following three levels:

(1) the impact of the intervention support on the performance of the teachers, as expressed in the kinds of questions that the teachers pose,
(2) the impact of teacher support on the pedagogical interaction between the students and their teacher, and
(3) the impact of participating in these pedagogical interactions on students' learning.

We were also interested in how mechanisms in the pedagogical context relate to teachers' implementation of the intervention.

Method

To investigate the impact of support for whole-class teaching with computer simulations according to an inquiry-based approach, we set up a series of pre-post quasi-experiments. Our procedures complied with our faculty's ethical standards. In each experiment a teacher presented a simulation-based lesson on Newtonian mechanics in two parallel classes, covering two days of instruction (about 80 minutes total). In one class teachers followed their usual (accustomed) way of teaching; in the other class they followed our script based on Peer Instruction for inquiry. We treat this as a series of separate quasi-experiments because we compared each teacher's accustomed way of teaching with a scripted condition, and the accustomed teaching modes may vary from teacher to teacher. Therefore we present each comparison between accustomed and scripted condition as a separate quasi-experiment. We will discuss the findings from the quasi-experiments in relation to each other in order to arrive at general conclusions.

Participants

Five science teachers volunteered to participate in our study, one of them female. Their average age was 40.2 (7.76) and ranged from 30 to 50 years. Each of them taught at least two parallel classes of the same level in upper secondary education. This allowed us to compare how each teacher taught when using two different approaches. Teachers received in advance a manual with information about the study design, participation criteria, preparatory steps, the planning of the lesson sequence, and a synopsis of physics concepts that are related to Newtonian mechanics. A total of 218 students participated, which includes only those who were present at the pretest, the intervention lessons, and also the posttest. They participated in their regular classes with their regular teachers. The students' average age was 15.7 (0.77) and ranged from 14 to 18 years. They were from two levels of the Dutch educational system: HAVO and VWO. HAVO has five grades and stands for "higher general continued education"; VWO has six grades and stands for "preparatory scholarly education".

Research design

Lesson sequence, simulations, and intervention

The total series of lessons on Newtonian mechanics spanned 4½ lessons for all participating students: during lesson 1 the pretest was administered; during lessons 2 and 3 the intervention lessons were conducted; during lesson 4 the posttest was administered; and one month later the delayed posttest was administered during lesson 5. Table 1 shows a schematic outline of the lesson sequence. Lessons 1, 4, and 5 were designed to be exactly the same for all participating students, as the pretest, posttest, and delayed posttest were administered during these lessons. The lessons that differed across the two conditions were lessons 2 and 3. We implemented our experimental intervention by having the same teacher teach with computer simulations during these lessons in two different ways in the two parallel classes. In the remainder of this article these two ways of teaching are referred to as the "Accustomed" condition and the "Peer Instruction" (PI) condition.

Table 1. Schematic outline of the sequence of lesson and assessments.

time	lesson 1	lesson 2	lesson 3	lesson 4	lesson one month later
10 20 30 40	FCI* (pretest)	experimental intervention	experimental intervention	FCI* (posttest)	FCI* (delayed posttest)

Note: *FCI: Force Concept Inventory.

We used simulations available from the PhET simulations website (Physics Education Technology, 2015). The simulations used in the lessons were selected based on the criteria that they aimed at resolving misconceptions about Newtonian mechanics, that a Dutch version was available, and that inquiry-based conceptual questions associated with the simulations were available (Loeblein, 2015). We chose the following PhET simulations: "Projectile motion", "Forces in 1 dimension", and "The ramp – forces and motion". Hestenes and colleagues (Force Concept Inventory (FCI), 2015, see "Revised Table II") clarify how specific Newtonian concepts are related to questions from the Force Concept Inventory (FCI), a test that assesses insight into Newtonian mechanics (Hestenes, Wells, & Swackhamer, 1992). By checking which specific Newtonian concepts are covered by the three chosen PhET simulations, we determined the overlap between the simulations and the questions from the FCI: questions 11 and 17 are most related to the concepts covered in these simulations; questions 5, 6, 7, 22, and 28 the least. Even though some questions are more related to the simulations than others, we administered and interpreted the FCI results as a whole, as recommended by Hestenes and Halloun (1995).

The teacher used the simulations on an interactive whiteboard (available in all classrooms). In the Accustomed condition, the teacher decided how the lesson would unfold. In the experimental (scripted) condition, the lesson unfolded along a pattern designed for Peer Instruction (see Figure 1). The teacher was asked to follow the script, instead of using his or her own preferred way of teaching. In this Peer Instruction condition the lesson unfolded using a PowerPoint presentation on a separate screen in front of the class. The researcher operated this presentation in collaboration with the teacher. This presentation was used to ensure that the general structure of lessons in this condition matched across the participating teachers. The built-in episodes of simulation-based Peer Instruction allowed for a shift of control over learning activities from the teacher to the students, which in turn better enables them to learn with simulations according to an inquiry-based approach (Salinas, 2008). The PowerPoint presentation consisted of information on which simulation to use and a series of 20 conceptual questions about Newtonian mechanics that were associated with the simulation being used by the teacher on the whiteboard. We based these questions (with permission) on the inquiry-based concept questions published by Trish Loeblein on the PhET website (Loeblein, 2015). Students could answer each multiple-choice question by using a personalized voting device. We implemented the voting process by showing the multiple-choice conceptual

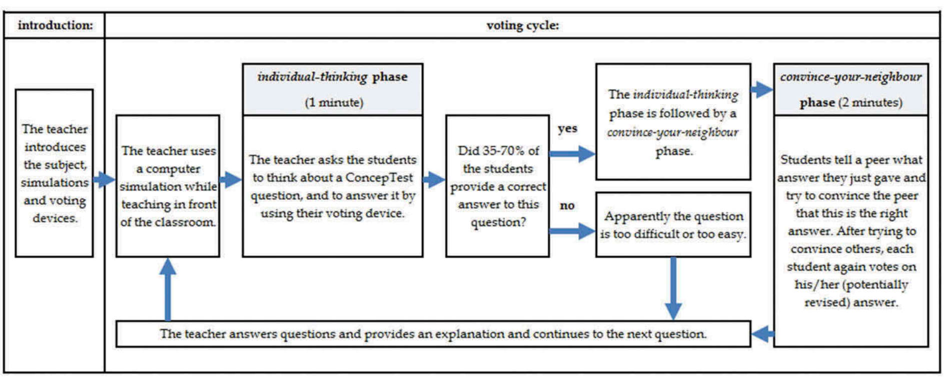

Figure 1. Peer Instruction implementation.

questions in the PowerPoint presentation and using a voting response system plug-in for PowerPoint. This allowed the students to see how many students had yet to vote, and to see the distribution of votes after the time for voting had elapsed.

Using this set-up we implemented Peer Instruction as described by Crouch and colleagues (Crouch, Watkins, Fagen, & Mazur, 2007); Figure 1 illustrates the connection of the different phases of instruction. The basic pattern ran as follows : in the "individual-thinking" phase the students were instructed to think individually about the question presented and vote. If the percentage of votes for the correct answer fell between 35% and 70%, a "convince-your-neighbours" phase followed. When the percentage of votes for the correct answer falls outside this range, the question is likely to be too difficult/ambiguous or too easy to make student discussions worthwhile (Crouch et al., 2007). During the "convince-your-neighbours" phase students were instructed to discuss the question in groups of two or three. They were instructed to tell each other what answer they gave and try to convince the other(s) that that answer was correct. After this phase students voted again. Based on the literature on Peer Instruction (Crouch et al., 2007) we set the duration of the "individual-thinking" phase at one minute and that of the "convince-your-neighbours" phase at two minutes.

Measures

We measured learning effects by administering questionnaires during a pretest, a posttest, and a delayed posttest. We used delayed posttests to check whether short-term increases in understanding also led to meaningful learning over the long term. We used the Force Concept Inventory (Hestenes et al., 1992) to assess improvements in conceptual insight in Newtonian mechanics. This instrument is available online in many languages at http://modeling.asu.edu/R&E/Research.html (2015) and has been widely used. Because the FCI data were collected at different timepoints and the students were nested within their teachers' classes, it would seem logical to analyse the data with a multi-level growth model. However, sufficient power to detect cross-level effects would require 20 or more groups (Kreft, 1998).

Investigating teacher performance and pedagogical interactions

Lesson observations

Table 2 shows the scheme (Rutten et al., 2015) that we used to code student–teacher interactions related to questions asked by the teacher in order to find out whether the teaching approach reflected important aspects of inquiry-based teaching. Any episode during which the teacher addressed the whole class was eligible for coding. We then used these codings to reveal the impact of our intervention on the pedagogical interactions between the students and their teacher and, in turn, the impact of the pedagogical interactions in the classroom on student learning outcomes.

All transcribed lesson observations were coded by the first author. Eight out of the 24 total transcripts were independently double-coded by a PhD student. The selection of these transcripts was balanced across conditions, teachers, and timepoint of data collection. We calculated the reliability of our approach for coding the teacher questions in three ways: our ability to discriminate consistently between actual physics content questions ("recall", "prediction", "observation", and "explanation") and "other" questions; our ability to discriminate who answered questions ("answered by the teacher", "answered

Table 2. Coding scheme for lesson observations.

	codes	application	examples
teacher questions that are related to physics			
What kind of question is it?	Recall	Questions that students should be able to answer with the knowledge they already have.	"In what unit is this variable measured?"
	Prediction	Students are asked to predict how a phenomenon will develop further before this has actually happened.	"What happens if that variable is doubled?"
	Observation	The teacher inquires about what students are observing at that moment.	"And what do you see right now?"
	Explanation	Students are asked to explain why a phenomenon has developed in a certain way.	"Now how do you explain this result?"
Who answers the question?	teacher	The teacher's question is answered by the teacher himself/herself.	
	student	The teacher's question is answered by a student.	
teacher questions that are not related to physics or fall within the categories below			
What kind of question is it?	other	Students are personally addressed.	"Alison?"
		Student answers are repeated back in the form of a question.	"You're saying a lower frequency?"
		The teacher checks whether subject-matter is understood.	"Is that clear?"
		The learning process is regulated.	"What have we seen today?"

by the student", and "other"); and our ability to discriminate between the different types of questions ("recall", "prediction", "observation", "explanation", and "other"). Calculations of inter-rater reliability based on Cohen's kappa revealed that the reliability of our discrimination between actual physics content questions and "other" questions was .87, for who answered questions it was .86, and for the different kinds of questions it was .73. According to the criteria proposed by Fleiss, Levin, and Paik (2003), these reliability results can be interpreted as excellent, excellent, and good, respectively.

For each lesson observation we calculated two scores: a Student-Response-Rate (SRR) and an Inquiry-Cycle-Score (ICS). Active student participation as indicated by the SRR was defined as the percentage of physics-related teacher questions that were answered by the students (and not by the teacher himself or herself). The ICS was computed based on teachers' actions related to a cycle of "predict-observe-explain". Hennessy and colleagues (2007) argued that this P-O-E cycle is one of the pedagogical principles upon which research on the use of technology in science has been based. Each question teachers asked was coded as either "recall" (r), "prediction" (P), "observation" (O) or "explanation" (E). Sequences of questions were assigned Inquiry Cycle scores as follows: E = 1; O-E = 2; P = 3; P-O = 4; P-E = 5; P-O-E = 6. This scoring system was based on the following rationale: without the phase of "prediction" we cannot speak of inquiry-based teaching. "Observation" makes an inquiry cycle more complete compared to a cycle in which explicit "observation" is lacking. In determining these Inquiry Cycle scores, "recall" and "other" questions were ignored and consecutive repetitions of the same question type were combined, such as an "observation" followed by an "observation". Shorter sequences only counted when they were not part of a longer sequence of higher weight. For example, the sequence P-P-P-O-E-E did not count as separate sequences of P, P-O, O-E, or E, because in this case these were all overlapped by one P-O-E sequence. We acknowledge that the information regarding what the questions were actually about was lost in this scoring system. Therefore, it is possible, for example, that a sequence that we consider P-O-E could span different conceptual domains, because of a switch to a different topic between an "observation" question and an "explanation" question.

Teacher predictions

Once a teacher had finished teaching lessons 1–4 for both conditions, the teacher was asked to predict which class had learned most about Newtonian mechanics during the series of lessons. They were also asked to describe what they thought were the most important contributing factors. Possible factors could address differences between classes and their circumstances, but could also be about differences in individual student characteristics. By asking for these predictions and contributing factors we could supplement our observational and questionnaire measures with teachers' personal reflections.

Results

Teacher performance and pedagogical interactions

Table 3 provides an overview of inquiry-based question sequences during each of the observed lessons. In total, we identified 749 physics content questions posed by teachers during 834 minutes of video recordings. The average duration of recordings transcribed for a lesson was 34.7 minutes, 95% CI [32.4-37.1]. Each lesson was scored for the extent to which the inquiry cycle was evident (ICS) and active student participation (SRR). The

Table 3. Teacher questions.

teacher	level	condition	lesson	Inquiry-Cycle-Score (ICS)	% answered by… student (SRR)	% answered by… teacher	physics content questions posed by each teacher in chronological order*
A	VWO	Accustomed	2	40	73	27	rrrrrrrPPPrOOrrrrrOOrrrErrrrrrOrrPrOOOOPrrrPErPPrOrPrPPEOr
			3				EPrrPPrrrrOOOPPOOPrPrrrrrrPPPPEOOEEO
		PI	2	17	67	33	EEPrPPrPrPPP
			3				rrrOEPPEOrOPPOrrrrEEE
B	VWO	Accustomed	2	33	94	6	OPPPPPPErrrrOOEOrrrPrrrrrrErrrPEOOPrrrrrOOOOPPPrPPPPPOrOrrrrrr
			3				OPrrrrErrrrrOrOOOOOPrrrrrrrrrrrrrrr
		PI	2	19	92	8	rPOrP
			3				rrPPPPPErrrrPPOOrrrrP
C	VWO	Accustomed	2	31	75	25	rrrEPEErrrOrrrrrPOPrrErrrrErrErrrErrOPPPrrrOEErPOrrErErrrErrEO
			3				OrrrEErrrrrrrErOErrrrrrOOOrrOOOrrrrrrrrr
		PI	2	11	47	53	PrOEEErrrr
			3				rrEPrPrrO
D	HAVO	Accustomed	2	45	79	21	PPPrEPPPPOOPPPOEOOOOPPPOPrOOPPOOOOrPOO
			3				PPOOrOOPOOPPPrrOOrrOEOOOOOrrOrrOrrOrrrrOrOOO
		PI	2	8	60	40	OOOr
			3				rOOrrrPOPOr
E	HAVO	Accustomed	2	39	88	12	rrrOrEErOrrEEOOEOrrrOrPrEOOOEOPEOPEOErPrrrrrrr
			3				rrEErrrPrErErrrEEOOOrPPPPPrErrrrEr
		PI	2	34	92	8	PEErrrrErErrrPPEErrOEPEEEEr
			3				EErOrEOEOOrrOEEPrrrrrrrrrOEOOEOEr
	VWO	Accustomed	2	38	91	9	OOrOOOOOOEOOOEEOrrEOOrrEEPOrEPEOEOOOOOOOOO
			3				rrrOEErOOrrErrOrrEErEErrrOrrErOrPrOOrrrOrrPPPE
		PI	2	40	89	11	PPEErrrrrrOErrPOEPPEPPEEOrPEEEOP
			3				EErrrOrPrrErrErrrrErP

Note: *Each code refers to a type of teacher question abbreviated as follows: r = recall, P = prediction, O = observation, and E = explanation. Questions in bold refer to teacher questions answered by a student; teacher questions that are not bold are answered by the teacher him-/herself. A darker shade of grey represents higher resemblance to the inquiry cycle.

ICS provides insight into the impact of the intervention support on the teachers' performance, as it does not depend on who answered the teacher's questions. The SRR takes into account whether a question is answered by the teacher or a student, and therefore allows for examination of the impact of teacher support on the pedagogical interaction between the students and their teacher.

Teacher E's teaching approach seemed to be quite consistent regarding both types of score, no matter whether he was teaching at the HAVO or the VWO level, or in the Accustomed or the PI condition. Teachers C and D in particular scored higher on SRR in the Accustomed condition (teacher C: 75; teacher D: 79) as compared to the PI condition (teacher C: 47; teacher D: 60). The support provided in the PI condition appeared to have a negative impact on the ICS of teachers A, B, C, and D, with the most extreme example being teacher D, with an ICS of 45 in the Accustomed condition and eight in the PI condition.

Table 4 shows teachers' predictions of which condition would have the highest learning gains, directly after the posttest was administered. It also elaborates on what factors the teachers deemed to be most influential regarding these learning effects. The participating teachers generally recognize the activating function of Peer Instruction: "… forces every student to participate, there is no 'opt-out'" (teacher E); "The peak of these lessons was at the times the students had to convince each other" (teacher B); and, "… they were alert, as they wanted to know whether they were right" (teacher A). However, teacher D, on the contrary, does not appreciate the PI condition, but favours the Accustomed condition for being easily adaptable to his own way of teaching. Teachers' experiences with the scripted structure of the PI condition were mixed: "The good tempo might also contribute to captivating them" (teacher A) as well as "I considered this lesson to be very passive and pre-programmed" (teacher B).

Learning outcomes

Table 5 shows the analysis of students' responses to the FCI at pretest, posttest, and delayed posttest. The *t*-values and Cohen's *d* effect sizes in this table seem to suggest that this series of pre-post quasi-experiments resulted in contradictory effects with respect to learning gains. Effect sizes can be considered large when they exceed 0.80, and medium when between 0.50 and 0.80 (Thalheimer & Cook, 2002). Significant effects with a large effect size as well were measured for teachers A and D at the posttest. However, the effect for teacher A favoured Peer Instruction ($t(39) = 2.62$, $p = .01$, $d = 0.89$), while that for teacher D favoured the Accustomed condition ($t(45) = -2.98$, $p = .01$, $d = -0.89$). Learning gains for teacher E's class at the VWO level were also significant and in favour of the Accustomed condition, but had a medium effect size: $t(36) = -2.19$, $p = .04$, $d = -0.73$. The advantage of the Accustomed condition for teacher D appeared to be robust over the long term: $t(43) = -3.04$, $p = .01$, $d = -0.93$.

Even though the results in Table 5 show that the learning effects are not consistently in favour of one condition, a comparison between Tables 4 and 5 reveals that where effects are large ($|d| > 0.8$) as well as significant, they coincide with the teachers' predictions. This suggests that after having finished the experiment, teachers' understanding of which teaching approach resulted in higher learning gains matched what could be measured by questionnaires alone.

Table 4. Teachers' predictions of which condition has the highest learning gains.

teacher	teacher's prediction	condition		factors considered by the teachers to influence learning gains
A	PI	Accustomed	helpful	I could focus on what I consider to be important.
				This is the "softer" group that is scared of formulas.
				They were glad to receive the material visualized.
				The "discovery" of F = ma was a success.
			detrimental	They had to think less about the concepts than the other group.
		PI	detrimental	My feeling says that it went too fast and superficially.
			helpful	It appealed to them and they were alert, as they wanted to know whether they were right.
			helpful	The good tempo might also contribute to captivating them.
B	no difference	Accustomed	helpful	This group now has another teacher and will therefore have been a bit more attentive to what happened in class.
		PI	helpful	This group seems more attentive to me as something new happens in class: the voting devices.
			detrimental	I considered this lesson to be very passive and pre-programmed.
			helpful	The peak of these lessons was at the times the students had to convince each other.
C	PI	Accustomed	detrimental	Besides focusing on the concepts, I had this group focus on constructing and calculating, which results in their being educated more in line with what is expected from them later on at the exams. They were, however, less well-prepared to answer the posttest questions.
			detrimental	I almost entirely taught in a frontal manner.
			detrimental	There was no opportunity to check and discuss their homework, resulting in less reflection than normal.
		PI	helpful	As this group received two lessons that were completely focused on understanding Newtonian mechanics, they were better prepared to answer the posttest questions.
D	Accustomed	Accustomed	helpful	After having conducted a lesson with the voting system,
				I could quickly adapt the lesson to teach in the way my students are accustomed to.
		PI		-
E	PI	HAVO A		-
		HAVO PI	helpful	As the voting system forces every student to participate, there is no "opt-out".
		VWO A		-
		VWO PI	helpful	This group has better students.
			helpful	I believe these lessons went more smoothly.
			helpful	They seem to be less affected by puberty, causing the yields of lessons to be higher.

Table 5. Analysis of FCI responses.

teacher		A		B		C		D		E			
year and educational level		4 VWO		5 VWO		4 VWO		4 HAVO		4 HAVO		4 VWO	
experimental condition		A	PI	A	PI	A	PI	A	PI	A	PI	A	PI
(A = *Accustomed*; PI = *Peer Instruction*)													
N students*		14	27	10	14	18	16	23	24	16	18	17	21
pretest	mean	10.14	10.56	13.10	13.64	13.39	12.88	8.74	8.33	8.31	8.33	6.53	7.05
	SD	2.41	3.56	4.01	5.90	4.50	3.14	4.26	2.14	3.24	2.91	2.83	2.52
posttest	mean	11.43	14.48	14.30	14.86	13.22	12.19	13.22	10.54	9.69	8.67	8.94	7.57
	SD	2.95	4.14	3.56	5.91	4.70	3.27	4.56	2.98	5.36	3.43	4.13	2.42
delayed posttest	mean	13.46	14.48	12.22	15.00	13.50	13.93	13.55	10.30	9.43	10.61	9.65	9.38
	SD	1.85	3.54	3.07	6.32	4.96	4.45	5.18	3.38	6.66	3.94	4.99	2.36
independent samples test: *t* (sig.)**	pre-post	2.62 (.01)		0.01 (.99)		–0.64 (.53)		–2.98 (.01)		–0.95 (.35)		–2.19 (.04)	
	pre-d. post	1.14 (.26)		1.54 (.14)		0.69 (.49)		–3.04 (.01)		0.80 (.43)		–0.78 (.44)	
effect size**	pre-post	0.89		0.01		–0.23		–0.89		–0.34		–0.73	
	pre-d. post	0.40		0.69		0.25		–0.93		0.30		–0.26	

Note: A darker shade refers to a higher score across the row for that measure; *Data for students who missed any part of the study except for the delayed posttest were excluded from the analysis; **Calculations of *t*-tests and Cohen's *d* effect sizes are based on gain scores. Positive *t*-values and effect sizes refer to effects in favour of the Peer Instruction condition.

Conclusions and discussion

The research questions of this study on improving our understanding about supporting teachers during whole-class use of computer simulations are focused on investigating (1) the impact of intervention support on the performance of the teachers, as expressed in the kinds of questions that the teachers pose, (2) the impact of teacher support on the pedagogical interaction between the students and their teacher, and (3) the impact of participating in those interactions on students' learning. From the data collected in the different classes we can see that, although we carefully supported each teacher with the implementation of the PI condition, there is wide variation in the teacher performance and pedagogical interaction that resulted from this implementation. Table 3 illustrates how diversely teachers can respond to an experimental implementation. With respect to ICS, teacher E taught more or less the same way in both conditions at both educational levels, whereas teacher D showed the most extreme difference: 45 in the Accustomed condition and only eight in the PI condition. In our PI condition, a bit of time is occupied by the researcher's PowerPoint questions and the students' processes of Peer Instruction and voting. If both conditions are equally effective at supporting teacher performance, it makes sense to have a higher ICS in the Accustomed condition than in the PI condition, because there simply is a bit more time available. Nevertheless, Table 3 clearly shows great disparity concerning the impact of experimental condition on teacher performance and interaction between teachers and students. A key role could be played by sense of ownership or control over the lesson. Teachers need to feel the urge and necessity to change their pedagogical approach, before they are willing to invest in an educational intervention (van den Berg & Geurts, 2007). This willingness to invest is a prerequisite for developing a sense of ownership (Breiting, 2008). According to such a view, teacher E retained control over the lesson no matter the condition in which he was teaching, resulting in relatively comparable instructional moves, interaction patterns, and learning effects. However, even though teacher D also had a relatively high ICS and SRR in the Accustomed condition, he did not exert control over Peer Instruction lessons, apparently considering the researcher to be in charge. This sense of ownership could be crucial for how the pedagogical interaction unfolds, impacting its learning outcomes. Investigations of teaching approaches by using scripted lessons should take this possibility into account, as such a scripted lesson might not lead to the desired behavioural changes by the teacher, because of losing this sense of ownership.

We consider the reasonable accuracy of the teacher predictions in Table 4 as support for supplementing the questionnaire results with our examination of lesson observations in order to extend the analysis of products at different timepoints to analysis of the instruction/learning process itself. These teachers' predictions are elaborated by the factors that the teachers considered to influence learning gains. Several factors mentioned might be generalizable beyond this study to studies in which learning effects are investigated with a researcher-imposed structure: teachers feeling less able to focus on what they consider important and experiencing the researcher-imposed condition as pre-programmed. A researcher who is imposing a certain lesson structure can also be experienced as a kind of invader, which possibly results in a teacher feeling tempted to oppose the intervention out of a resistance to change (Reid, 2014). Teacher D's significant *t*-values and large effect sizes in favour of the Accustomed condition could be explained by his approach of combining the best of both worlds, as he explains: "After having conducted a lesson with the voting system, I could quickly adapt the lesson to teach in the way my students are accustomed to". Apparently, teacher D preferred to transform the less familiar educational intervention into an approach to which he and his students were more accustomed. While teacher E taught consistently across conditions and educational levels, teacher D exerted the least control in Peer Instruction lessons, which might also be related to a resistance to change (Reid, 2014).

Our findings presented in Table 5 illustrate the expression that "one experiment is no experiment at all": differences found between conditions for certain teachers are the opposite for others. These findings suggest that our questionnaire data by themselves provide insufficient information for understanding the students' learning gains or lack thereof. This pre-post quasi-experimental research design is often used for investigating the learning effects of computer simulations in science education (Rutten et al., 2012): measuring the effects of an intervention by comparing several classes of students at different timepoints. This design is used in a variety of instructional settings: working with computer simulations individually, in small groups, or in a whole-class setting, where the teacher or a student operates the simulation in front of the class. The present study shows that supplementing such a pre-post design with process analyses focused on teacher performance and teacher–student interaction can help yield better understanding of the pedagogical mechanisms that are at play.

In our review study (2012, p. 151), we concluded that most of the studies that we reviewed "investigated the effects of computer simulations on learning *ceteris paribus*, consequently ignoring the influence of the teacher, the curriculum, and other such pedagogical factors". Our present study shows that it is important to go deeper than just looking at learning outcomes for understanding how to best support whole-class teaching with computer simulations. When viewing the learning situation at this abstract level too many factors of the context at hand are not taken into account, resulting in an abstract principle about computer simulations that cannot be concretized in other contexts (Rol & Cartwright, 2012). We therefore recommend that further research on the learning effects of computer simulations, and technology in general, take into account contextual, pedagogical factors, and incorporate these into research designs.

According to Crasnow (2012), the socio-scientific research of the past decades shows a methodological shift: switching from investigating specific cases to a more statistical approach, based on the idea that this allows for finding principles that are generally applicable. The assumed usefulness of such a statistical approach is that finding "effects" also brings into better focus what are the possible "causes" of these effects. However, the decontextualization of the research participants, which is necessary for the determination of such general principles, narrows down the insightfulness of such principles. Furthermore, the possibility exists that the effects found do not even exist as decontextualized principles, when these are partially caused by the specific characteristics of the context. Crasnow (2012) therefore argues that to understand reality it is not only necessary to search for general, decontextualized principles, but also to supplement this search with observations of participants within their context. Paraphrasing Crasnow (2012), this means not only searching for the "effects" of "causes", but also for the "causes" of "effects". The widely used research design on which the pre-post quasi-experiment of our study was based is used by researchers to find "effects", often in the sense of generally applicable principles. By replicating this research design several times, we showed that the effects found per experiment do not qualify for general applicability. Our analyses of the teachers' behaviours and pedagogical interaction allowed us to search for possible "causes" of "effects": searching for mechanisms that could have influenced the effects in the contexts at hand. In this case, that meant looking further for why teachers might have responded to the support provided in the Peer Instruction condition as they did.

Suggestions for future research

In the present study, we supplemented the pre-post research design that is often used in this field in two ways: by performing it six times, and by not only analysing conceptual gains by measuring at different timepoints, but also zooming in on teacher behaviours and the pedagogical interactions between the students and their teacher. For testing the effectiveness of pedagogical approaches, it is beneficial to conduct such process analyses of how teacher–student interaction unfolds as a supplement to measuring learning outcomes with pre- and posttests. We suggest the following working hypotheses as possible mechanisms in the pedagogical context that influenced the instruction/learning processes in the present study:

(1) When imposing more structure in one condition compared to the other, this can lead to the teacher losing a sense of ownership in the structured condition, possibly causing the teacher to take a more passive stance.
(2) When the approach in one condition is less familiar to the teacher compared to the other, then this sense of familiarity can differentially influence the execution of the less familiar pedagogical approach.
(3) Both imposed structure and unfamiliarity can cause the teacher to refrain from implementing a pedagogical approach out of a resistance to change.

These working hypotheses relate to the first two levels of effectiveness that we identified in the Introduction section when discussing evaluating the effectiveness of an instructional intervention: teacher performance support by the intervention, and student learning support by the teacher. We recommend that researchers conducting similar studies not only describe the pedagogical intervention itself, but also thoroughly elaborate on how their implementation strategy takes into account influential mechanisms in the pedagogical context. We consider it particularly worthwhile to investigate these working hypotheses by having teachers collaboratively design pedagogical interventions, implement these in their teaching practices, and collaboratively adapt these interventions afterwards, as inspired by their experiences.

Acknowledgements

The authors would like to thank of all the teachers and their students who participated in this study. We would also like to thank Dr M. E. G. M. (Menno) Rol and Dr E. (Emily) Fox for their valuable feedback on earlier versions of this article.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Nico Rutten http://orcid.org/0000-0002-1854-4546
Wouter R. van Joolingen http://orcid.org/0000-0002-4271-2861

References

Breiting, S. (2008). Mental ownership and participation for innovation in environmental education and education for sustainable development. In A. Reid, B. Jensen, J. Nikel, & V. Simovska (Eds.), *Participation and learning* (pp. 159–180). Dordrecht: Springer Netherlands.

Crasnow, S. (2012). The role of case study research in political science: Evidence for causal claims. *Philosophy of Science*, *79*(5), 655–666.

Crouch, C. H., Fagen, A. P., Callan, J. P., & Mazur, E. (2004). Classroom demonstrations: Learning tools or entertainment? *American Journal of Physics*, *72*(6), 835–838.

Crouch, C. H., & Mazur, E. (2001). Peer Instruction: Ten years of experience and results. *American Journal of Physics*, *69*(9), 970–977.

Crouch, C. H., Watkins, J., Fagen, A. P., & Mazur, E. (2007). Peer Instruction: Engaging students one-on-one, all at once. *Research-Based Reform of University Physics*, *1*(1), 40–95.

de Jong, T., & van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, *68*(2), 179–201.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Force Concept Inventory. (2015). Retrieved from http://modeling.asu.edu/R&E/Research.html

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410–8415.

Hennessy, S., Wishart, J., Whitelock, D., Deaney, R., Brawn, R., la Velle, L., McFarlane, A., Ruthven, K., & Winterbottom, M. (2007). Pedagogical approaches for technology-integrated science teaching. *Computers & Education*, *48*(1), 137–152.

Hestenes, D., & Halloun, I. (1995). Interpreting the force concept inventory. *The Physics Teacher*, *33*(8), 502–506.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*(3), 141–158.

Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. *Science Education*, *88*(1), 28–54.

Khan, S. (2011). New pedagogies on teaching science with computer simulations. *Journal of Science Education and Technology*, *20*(3), 215–232.

Krajcik, J., Blumenfeld, P. C., Marx, R. W., Bass, K. M., Fredricks, J., & Soloway, E. (1998). Inquiry in project-based science classrooms: Initial attempts by middle school students. *Journal of the Learning Sciences*, *7*(3–4), 313–350.

Kreft, I. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

Loeblein, P. (2015). Concept questions for Physics using PhET (Inquiry Based). Retrieved from http://phet.colorado.edu/files/activities/3112/Loeblein%20physics%20clicker%20questions.pptx

Maeng, J. L., Mulvey, B. K., Smetana, L. K., & Bell, R. L. (2013). Preservice teachers' TPACK: Using technology to support inquiry instruction. *Journal of Science Education and Technology*, *22*(6), 838–857.

Physics Education Technology website. (2015). Retrieved from http://phet.colorado.edu

Pyatt, K., & Sims, R. (2012). Virtual and physical experimentation in inquiry-based science labs: Attitudes, performance and access. *Journal of Science Education and Technology*, *21*(1), 133–147.

Reid, P. (2014). Categories for barriers to adoption of instructional technologies. *Education and Information Technologies*, *19*(2), 383–407.

Rol, M., & Cartwright, N. (2012). Warranting the use of causal claims: A non-trivial case for interdisciplinarity. *Theoria*, *27*(2), 189–202.

Rutten, N., van der Veen, J. T., & van Joolingen, W. R. (2015). Inquiry-based whole-class teaching with computer simulations in physics. *International Journal of Science Education*, *37*(8), 1225–1245.

Rutten, N., van Joolingen, W. R., & van der Veen, J. T. (2012). The learning effects of computer simulations in science education. *Computers & Education*, *58*(1), 136–153.

Salinas, M. F. (2008). From Dewey to Gates: A model to integrate psychoeducational principles in the selection and use of instructional technology. *Computers & Education*, *50*(3), 652–660.

Smetana, L. K., & Bell, R. L. (2012). Computer simulations to support science instruction and learning: A critical review of the literature. *International Journal of Science Education*, *34*(9), 1337–1370.

Smetana, L. K., & Bell, R. L. (2014). Which setting to choose: Comparison of whole-class vs. small-group computer simulation use. *Journal of Science Education and Technology*, *23*(4), 1–15.

Thalheimer, W., & Cook, S. (2002). How to calculate effect sizes from published research articles: A simplified methodology. Retrieved from http://www.bwgriffin.com/gsu/courses/edur9131/content/Effect_Sizes_pdf5.pdf

Trundle, K. C., & Bell, R. L. (2010). The use of a computer simulation to promote conceptual change: A quasi-experimental study. *Computers & Education*, *54*(4), 1078–1088.

van den Berg, J., & Geurts, J. (2007). Leren van innoveren: Vijf sleutels voor succes [Learning through innovation: Five keys to success]. *CINOP, 's-Hertogenbosch*.

Zingaro, D., & Porter, L. (2014). Peer Instruction in computing: The value of instructor intervention. *Computers & Education*, *71*(0), 87–96.