J Sci Educ Technol (2018) 27:215–235
https://doi.org/10.1007/s10956-017-9719-8

Tools for Science Inquiry Learning: Tool Affordances, Experimentation Strategies, and Conceptual Understanding

Engin Bumbacher[1] · Shima Salehi[1] · Carl Wieman[1] · Paulo Blikstein[1]

Published online: 26 December 2017

Abstract Manipulative environments play a fundamental role in inquiry-based science learning, yet how they impact learning is not fully understood. In a series of two studies, we develop the argument that manipulative environments (MEs) influence the kind of inquiry behaviors students engage in, and that this influence realizes through the affordances of MEs, independent of whether they are physical or virtual. In particular, we examine how MEs shape college students' experimentation strategies and conceptual understanding. In study 1, students engaged in two consecutive inquiry tasks, first on mass and spring systems and then on electric circuits. They either used virtual or physical MEs. We found that the use of experimentation strategies was strongly related to conceptual understanding across tasks, but that students engaged differently in those strategies depending on what ME they used. More students engaged in productive strategies using the virtual ME for electric circuits, and vice versa using the physical ME for mass and spring systems. In study 2, we isolated the affordance of measurement uncertainty by comparing two versions of the same virtual ME for electric circuits—one with and one without noise—and found that the conditions differed in terms of productive experimentation strategies. These findings indicate that measures of inquiry processes may resolve apparent ambiguities and inconsistencies between studies on MEs that are based on learning outcomes alone.

Keywords Science inquiry · Experimentation strategies · Physical & virtual manipulative environments

Introduction

Inquiry-based instruction plays a prominent role in contemporary models of science teaching and learning (Quinn et al. 2012). Novel virtual and physical tools such as interactive computer simulations or remote laboratories have created new opportunities for inquiry-based experimentation in schools beyond traditional physical laboratories (Heradio et al. 2016). These tools enable learners to observe scientific phenomena, manipulate variables, set up experiments, and collect data; hence, we refer to them as manipulative environments (ME) for inquiry. The design space for such manipulative environments is large and diverse, and not all ME are effective at fostering science learning through inquiry (Zacharia et al. 2015). How can we help educators choose environments best suited for their needs, and inform designers to develop effective environments? In order to answer these questions, we need to understand (1) what makes an environment effective for a learning objective, (2) what the critical design features are for successful learning, and (3) how can we leverage these features.

The past decades have seen great progress in research on MEs for inquiry-based learning. A large body of research has focused on *medium* of MEs and have compared the relative effectiveness of virtual and physical ME and of combinations thereof; more recently, researchers have shifted attention to the affordances of MEs besides their medium; affordances such as the capability to manipulate normally

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10956-017-9719-8) contains supplementary material, which is available to authorized users.

✉ Engin Bumbacher
buben@stanford.edu

[1] Graduate School of Education, Stanford University, Stanford, CA 94305, USA

unobservable variables or the presence of measurement error (de Jong et al. 2013; Zacharia et al. 2008). Ongoing research on the impact of different MEs on learning, and on the role of various affordances of ME, examined different technologies, subject domains and age groups; but results are still inconclusive (de Jong et al. 2013).

These ambiguities need to be resolved to properly inform the evaluation of existing MEs and the design of new ones. In this paper, we argue that the reason for these ambiguities is more of a methodological nature than of a conceptual one, and suggest two ways to modify the predominantly used experimental designs to address the problem. We present this argument in two parts, each of which is based on a different study.

The first part of the paper presents the argument that studies comparing—or combining—virtual and physical MEs appear to be inconclusive because they focused on learning *outcomes* with little consideration of inquiry *process*, such as learners' experimentation strategies or data interpretation. Causal mechanisms for the impact of ME on learning outcomes are likely to be found in how learners engage with a ME. Therefore, not examining this engagement means not examining the reasons for differences in learning outcomes between MEs (Parnafes 2006). In study 1, learners used either a virtual or a physical ME (VME or PME) in two inquiry tasks on different physics phenomena—mass and spring systems and electric circuits. We measured both their use of experimentation strategies, in particular the control of variable strategy, and their conceptual understanding before and after each activity. Results indicate that in both tasks, learners' use of experimentation strategies differed by *medium* of ME (virtual or physical), and this difference could partially account for the different learning outcomes between VME and PME. Furthermore, measures of experimentation strategies revealed distinctions between MEs even when the average changes in conceptual understanding were the same.

The second part of the paper argues that the inconsistencies between studies on the impact of affordances of MEs on learning arise because the studies mainly compared MEs across medium, i.e., virtual MEs (VME) with physical MEs (PME). *MEs that differ in medium tend to differ in more than just one affordance*. For instance, the VME and PME used in study 1 for the electric circuits task differed in ease of variable manipulation, complexity of the scientific phenomenon, measurement uncertainty, and observational feedback. It is almost impossible to match VME and PME with respect to all but the target affordance (e.g., Triona and Klahr 2003). Potential confounds remain due to intrinsic differences between the media, such as the amount of visual information, the dimensionality of the phenomenon, or the ways of interacting with each medium.

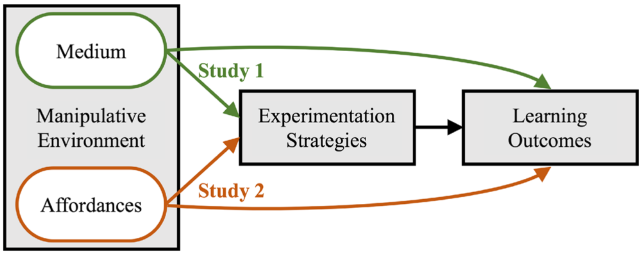

Fig. 1 Diagram of study 1 and study 2. Arrows indicate the targeted relationships in each study

Instead of evaluating the causal impact of single affordances by contrasting VME and PME, we suggest to compare MEs of the same medium that differ in only the target affordance. In study 2, we used this approach to explore what affordances caused the differences between learners using the VME and PME in the electric circuits task in study 1. Specifically, we targeted system-inherent noise, which is one of the affordances considered to be a key differentiator between VME and PME (de Jong et al. 2013). We found that the presence of noise had an impact on learners' experimentation strategies and reasoning, but that this affordance alone could not account for the results of study 1.

The overall goal of these studies (see Fig. 1) was to provide experimental evidence for, first, the necessity of using measures of inquiry strategies, in this case experimentation strategies, when studying how MEs impact learning outcomes (study 1); and second, the necessity of examining ME affordances separately and independent of medium, when studying how MEs impact learning outcomes (study 2).

Theoretical Background

Inquiry-based learning approaches build on constructivist models of learning that consider learning as an active, interpretive and iterative process, where learners construct their ideas and understanding of new experiences based on prior ones (Hofstein and Lunetta 2004). Activities for inquiry-based experimentation in classrooms typically require students to investigate multivariate systems using the tools for experimentation and data analysis (Chinn and Malhotra 2002b; de Jong and van Joolingen 1998).

Research communities hold different views about what science inquiry is. These have implications for how inquiry activities are designed, what factors are considered as important, and what measures of learning are used (e.g., compare Duschl and Grandy 2008 and Pedaste et al. 2015). Research on ME and on inquiry processes predominantly hold the view of science inquiry as a cyclic process of hypotheses generation, planning and execution of experiments and data analysis (Pedaste et al. 2015), while other

Table 1 Broad categorization of affordances of ME, with examples

Type of affordance	Examples of affordances
Representation	Dynamics of representations (Ainsworth and VanLabeke 2004)
	Multiple and multimodal representations (Hsu and Thomas 2002)
	Visualization of objects and processes, abstract, or beyond perception (Winn et al. 2006)
	Dynamic linking across representations (van der Meij and de Jong 2006)
Interaction	Physicality/tangibility (ability to touch concrete material) (Lazonder and Ehrenhard 2014)
	Possibility to repeat and modify experiments (Zacharia et al. 2008)
	Manipulation and customization of measurement equipment (Triona and Klahr 2003)
	Manipulation of reified objects, e.g., global temperature (Windschitl 2000)
	Degrees of freedom of manipulation (Renken and Nunez 2013)
Built-in support (scaffolds and feedback)	Immediate feedback and error correction (Huppert et al. 2002)
	Directing attention of students (de Jong and van Joolingen 1998)
	Cognitive or collaborative scaffolds (Zacharia et al. 2015)
Complexity of model	Degree of complexity (by adding friction, gravity, resistance, etc.)
	Measurement uncertainty and system-inherent noise (Renken and Nunez 2013)
Infrastructure	Portability, safety, cost-efficiency, reliability, etc.

Notes: We have listed affordances that have been reviewed by Marshall and Young (2006), Olympiou and Zacharia (2012) and de Jong et al. (2013)

work is more aligned with argumentation-driven and dialectic views of science inquiry (Duschl and Grandy 2008; Sandoval and Reiser 2004). For this paper, we draw on the former view of science inquiry in the design of the inquiry activities. However, we think that our conclusions are generally applicable to inquiry-based learning approaches using manipulative environments, independent of the underlying view of science inquiry.

Manipulative Environments for Inquiry-Based Experimentation

A large body of research on manipulative environments has focused on the effect of medium of ME on inquiry-based learning. The aggregate findings are that both virtual and physical ME generally promote conceptual understanding, but the studies disagree on their relative effectiveness (Finkelstein et al. 2005; Gire et al. 2010; Jaakkola and Nurmi 2008; Klahr et al. 2007; Marshall and Young 2006; Olympiou and Zacharia 2012; Winn et al. 2006; Zacharia et al. 2008).

The prevalent explanation for the lack of consistency is that the VMEs and PMEs in the studies carried different *affordances* (Klahr et al. 2007; Olympiou and Zacharia 2012). We conceptualize affordances of a ME as: 1. what you can possibly do in the ME and how you interact with it (interaction); 2. what you can possibly see in the ME and how that information is represented (representation); 3. how all of this is made available to the user (structure and scaffolds). We include representational aspects of a ME into the concept of affordances because the information a ME provides builds the basis for decisions of how to interact with it. Researchers have been compiling a growing list of affordances of ME that range from safety-related affordances to the types of variables that can be manipulated. Table 1 shows a list of affordances grouped into five broad categories. These categories might not capture all possible affordances of MEs, but they capture essential affordances that are independent of the medium of a ME. In this paper, we will not talk about affordances for infrastructure nor about affordances for built-in support, which have been extensively reviewed in Zacharia et al. (2015).

While VME and PME share some affordances for inquiry-based learning, like the possibility for active manipulation or collecting data from experiments, other affordances have been treated as inherently tied to a medium (Winn et al. 2006). For instance, VME are more likely to enable direct interaction with objects that are not manipulable in real life, whereas PME naturally provide the affordance of physical touch (Lazonder and Ehrenhard 2014). Recent work has shown that any combination of VME and PME leads to better conceptual understanding than either one alone, likely because combinations best leverage the different affordances of each ME[1] (de Jong et al. 2013; Jaakkola 2012; Olympiou and Zacharia 2012; Zacharia 2007).

In large parts, MEs have been combined based on the match of their affordances with specific learning objectives.

[1] Alternative explanations are that combinations of ME provide multiple representations, which leads to increased learning benefits, irrespective of affordances. The current state of research cannot exclude these alternative explanations (Lazonder and Ehrenhard 2014).

Choosing MEs that way can resolve the inconsistencies mentioned previously to the extent that the links between affordances and learning objectives are understood. For some learning objectives, it is clear which affordances match them well. For instance, the affordances of PME are generally better aligned with learning goals of developing practical laboratory skills, such as troubleshooting of experimental apparatus, or understanding safety procedures (de Jong et al. 2013). Conversely, affordances of VME might be better for learning objectives related to phenomena that are impossible to examine with PME at school (Olympiou and Zacharia 2012).

It is much less clear what affordances lend themselves to learning objectives related to conceptual or epistemological understanding (Chen et al. 2014; Chini et al. 2012). For example, the affordance of manipulating variables rapidly and easily (normally attributed to VME) is either seen as beneficial because learners get exposed to more examples in the same amount of time (Huppert et al. 2002; Zacharia and de Jong 2014), or as detrimental because it can encourage learners to carry out "play-like," undeliberated interactions (Renken and Nunez 2013). Another example is the affordance of measurement uncertainty due to noise that is naturally present in PME. On the one hand, measurement uncertainty can make it difficult to interpret the data, as it decreases the signal-to-noise ratio in the data (Chinn and Malhotra 2002a). On the other hand, the absence of measurement uncertainty can induce oversimplified views of science inquiry (Chen 2010; Chen et al. 2014).

One reason for these contradictions is that the interplay between affordances and conceptual understanding is inherently rather complex and thus difficult to study (e.g., physicality, see Gire et al. 2010; Lazonder and Ehrenhard 2014; Triona and Klahr 2003; Zacharia et al. 2012). Klahr et al. (2007) argue that conflicting results could also arise from the lack of proper control for confounding variables in these studies, such as the method of instruction or curriculum materials. More fundamentally, we think that comparisons of VME and PME inherently confound the effect of single affordances. For example, Finkelstein et al. (2005) found differences in conceptual understanding between learners using VME and PME on an inquiry task for electric circuits; but we cannot determine from this study whether the difference in performance was caused because the VME provided richer visual cues, allowed for easier manipulation of the circuits, constrained the manipulation of irrelevant variables (e.g., the color of the wires), or because of the interplay between these affordances.

Several studies, in particular on representational affordances, have shown that it is possible to isolate the effect of specific affordances by using MEs of the same medium that differ only in the targeted affordance. These studies found that conceptual understanding can be fostered when abstract objects are represented in addition to concrete objects (Olympiou et al. 2013), when multiple representations are used rather than single representations only (Ainsworth and VanLabeke 2004), or when representations are dynamic rather than static (see McElhaney et al. 2015 for a thorough review).

Inquiry Strategies and Conceptual Understanding in Inquiry-Based Experimentation

Developmental and cognitive psychologists studying inquiry-based learning have particularly focused on inquiry strategies and their interplay with prior knowledge (Klahr and Dunbar 1988; Schauble 1996). They have predominantly used microgenetic methods (Siegler and Crowley 1991), i.e., tracking the process of change by observing participants as they engage in an inquiry-based activity (Zimmerman 2000). They show that students' learning outcomes from inquiry-based activities depend on their inquiry strategies, such as their experimentation or data interpretation strategies (for extensive review, see Zimmerman 2000, 2007).

In this paper, we focus on experimentation strategies as a subset of inquiry strategies, because how students learn from the experiments hinges on the quality of the data they gathered. Data from unconfounded experiments, i.e., experiments that target isolated variables by controlling other variables, are more likely to be interpreted correctly and more likely to promote conceptual understanding than data from confounded experiments (Chen and Klahr 1999; Klahr and Nigam 2004). This is why the *control of variable strategy* (CVS), the strategy to set up unconfounded experiments, is one of the most extensively studied, systematic experimentation strategies (Zimmerman 2000).

A robust finding is that the use of CVS is subjected to large inter- and intra-individual variability, depending on a variety of factors; for example, the frequency and utility of CVS depends on students' domain-specific knowledge, students with higher prior knowledge tend to use CVS more often (Kanari and Millar 2004; Schauble 1996; Schauble et al. 1992).

There is skepticism of the focus on CVS in light of its limitations for exploring multivariate relations or developing an epistemologically rich understanding of science (Conlin 2014; Kuhn et al. 2008; McElhaney 2010). However, as a basic domain-general search strategy, CVS is valid and essential for investigation across different content areas (Chen and Klahr 1999), as it produces evidence that is interpretable and facilitates inferential skills (Zimmerman 2007). The use of CVS also can easily be extracted from the experiments students run during an inquiry activity. As such, CVS provides a simple but robust measure of experimentation strategies that are *productive* for learning, i.e., are correlated with learning outcomes, and that we could use to study how ME impact learning at a fine-grained level.

Overall, there has been relatively little work connecting the body of research on inquiry strategies to research on manipulative environments. Some experimental studies on manipulative environments did look at how students engaged with virtual and physical MEs during an inquiry activity. They focused on measures of experimentation like the number of experiments (Chien et al. 2015; Lazonder and Ehrenhard 2014; Renken and Nunez 2013), or the problems that emerged during the setup or the evaluation of experiments (Finkelstein et al. 2005; Zacharia and de Jong 2014; Zacharia and Michael 2016). These studies found differences between the virtual and physical MEs, which suggests that the manipulative environment can impact how students go about an inquiry activity.

However, only a few studies on MEs have examined experimentation strategies and the quality of experiments. Triona and Klahr (2003) have looked at how virtual and physical MEs compare as tools for teaching CVS. Renken and Nunez (2013) compared middle school students' experimentation and conceptual understanding in an activity on pendulum motion when using either a VME or a PME.[2] They found that students using the VME ran more trials and fewer controlled experiments; but the study lacks conclusive evidence for how these differences came about because the MEs differed in multiple affordances, such as the ease of manipulation and range of variable values.

A further limitation of Renken and Nunez (2013) is the simplified task structure with a rather small space of possible experiments, and a fixed sequence of inquiry phases for each student. Such structural simplifications are common in studies on inquiry-based learning (for review, see Zimmerman 2000). However, there is a risk of oversimplification. For example, when the range of choices is too small (for example by dichotomizing variables, Kuhn and Dean 2005), trial and error methods become sufficiently efficient to cover the range of possible experiments; or when the inquiry process is structured too heavily, students' cognitive involvement gets reduced to mere execution of steps.

Target Concept and Intentionality: Adapting Simple Measures of Experimentation Strategies to more Complex Inquiry Tasks

We employed less structured, more complex inquiry tasks than Renken and Nunez (2013). The target phenomena were multivariate, with multiple non-dichotomous dependent and independent variables. Furthermore, the tasks were goal-oriented in that learners had to find relevant relationships between the variables, but open-ended in that learners did not get any further guidance or procedural constraints.

Such tasks give learners more flexibility both in what experiments to run and in how to run them. In order to account for the increased flexibility and complexity, we propose two modifications to simple frequency counts of control of variable strategy (CV) as measures of productive experimentation strategies:

First, we take into account the time between experiments as a measure of how *intentional* learners are in their actions (Fischer 1980). A learner might set up a controlled experiment intentionally, after deliberate planning and reflection, or accidentally by non-goal directed manipulation of variables (de Jong and van Joolingen 1998). Levy and Wilensky (2006) proposed to distinguish between these behaviors by the duration between consecutive experiments. Longer durations likely indicate intentional behavior due to reflection on or planning of experiments. Hence, we define CV that occur "long enough" after the previous experiment as *intentional* CV, and as non-intentional otherwise.

Second, we distinguish between CV experiments that target different concepts. In multivariate systems, certain relations between variables might be more familiar to learners than others. Accordingly, experiments targeting the unfamiliar relations can benefit conceptual understanding more than experiments of familiar variables (Salehi et al. 2015), but this is not necessarily always the case (Renken and Nunez 2013).

In sum, we operationalize experimentation strategies as depicted in Fig. 2, with emphasis on the deepest level in this decision tree. Besides focusing on CV experiments, we also keep track of experiments that are confounded. Under "other," we capture experiments that cannot be classified as either CV or confounded, such as repetitions of experiment configurations. Importantly, this operationalization implies that we need to consider all dimensions when evaluating the experimentation strategies, and not just one.

As we considered intentionality of experiments, we also wanted to see whether we could support learners be more reflective and hence more intentional in their inquiry. There are instructional interventions for supporting students in productively designing, observing, and interpreting

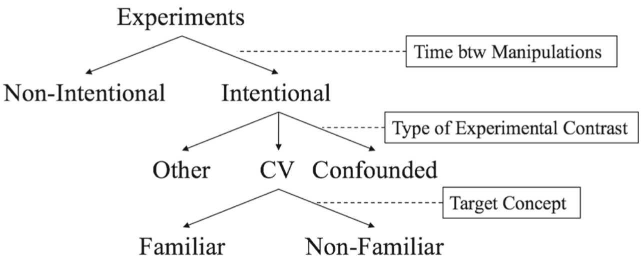

Fig. 2 Operationalization of experimentation strategies

[2] In Lazonder and Ehrenhard (2014), the researcher ensured that each experiments students designed was unconfounded.

experiments (Chinn and Malhotra 2002a). Predict-Observe-Explain (POE) is one such approach that requires learners before each experiment to make predictions about the outcome, to design the experiment, and to re-examine the predictions in comparison to the new observations (Rickey and Stacy 2000). Despite its caveats (Kearney et al. 2001), POE is a simple intervention that has been conducive to conceptual understanding, for example in physics. We hypothesize in study 1 that the POE intervention gets learners to think more carefully about how to run the experiments, which manifests in an increased proportion of intentional and unconfounded experimentation (Gunstone and Mitchell 2005; Paulson et al. 2009).

Summary of the Literature Review and Purpose of Study 1 and Study 2

For conceptual understanding to develop in inquiry activities, productive inquiry strategies are essential, and not just the exposure to one medium of ME. Manipulative environments and influences of their affordances on inquiry strategies have been under-researched. Therefore, studies on MEs have been inconclusive about how the MEs and their affordances impact conceptual understanding. We argue that measures of inquiry strategies, including experimentation strategies, can resolve some of the inconsistencies by increasing the "resolution" of the difference between MEs. The present work lies at the intersection of research on ME and research on inquiry strategies to address these issues.

Study 1

Research Questions

The main research questions addressed are the following:

1.A. Does the use of productive experimentation strategies differ between the virtual or physical MEs?

1.B. To what extent does the use of these strategies mediate potential differences in conceptual understanding?

1.C. How does the relation between medium, experimentation strategy, and conceptual learning outcomes compare between activities?

2. Does accounting for intentionality improve the measures of productive experimentation inquiry strategies?

3. Does the POE intervention moderate the effect of ME on learning outcomes by improving the use of productive experimentation strategies?

General Study Design

We applied a 2 x 2 study design, and each participant did two inquiry activities during the study. The first between-subject factor was medium of the ME. Participants used either the physical system (PME) or the interactive computer simulation (VME) for both activities. The second factor was the POE instruction. In the first activity, participants either received an instructional intervention on Predict-Observe-Explain (POE) or no guidance. In the second activity, participants did not receive any instructional guidance.

The first activity was on mass and spring oscillation, and the second one on basic DC electric circuits (activity as a within-subject factor). We chose these two subject domains because of the following reasons: First, these topics were new enough to the college student population, yet easy enough to be explored in the short time span of the activities. Second, the nature of phenomena differed significantly in terms of the physical perception of the dependent variables. Mass and spring systems provide physically grounded experiences through the perceptual salience of weight, tension, and forces. The variables of interest are visually perceptible without the need of additional measurement tools (similar to Lazonder and Ehrenhard 2014). On the other hand, current and voltage are indirectly perceptible only by using light bulbs or measurement tools.

Materials and Methods

Participants

Sixty-eight students of the same community college participated in the experiment for credit of psychology courses. Average age of participants was 20.5 ($SD = 3.49$). The study was run individually. Participants were randomly assigned to the experimental conditions. The unbalanced design resulted from some registered participants not attending their study session (see Fig. 4).

Manipulative Environments

The manipulative environments are shown and described in Fig. 3a and b.

Procedure

For the overall procedure, see Fig. 4.

Mass and Spring Activity Introduction: Participants were familiarized with the concepts and the ME. Prompt of Activity: "How do the mass and the spring constant influence

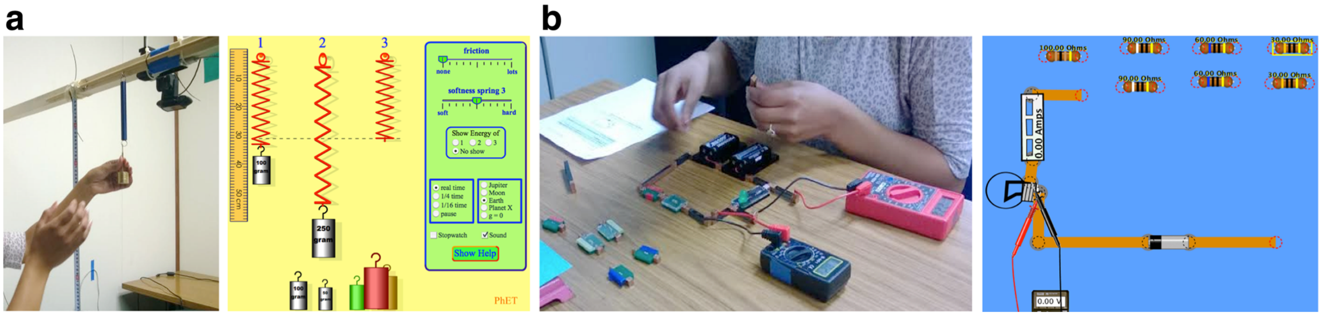

Fig. 3 Study 1: Manipulative environments for the Mass and Spring activity and the Electric Circuits activity. **a** PME (left): The system consists of two hooks to hold springs; there are four different pairs of springs. Several disks of different weights can be attached to hangers as the mass; VME (right): PhET interactive computer simulation (Perkins et al. 2005). The only parameter participants are allowed to change is "softness of spring 3." There are seven different weights. In both MEs, amplitude and frequency can be measured with a measuring tape and stopwatch. **b** PME (left): The electric circuits kit was a prototype developed by Chan et al. (2013). Resistors, batteries, wires, and LEDs can be quickly snapped together and disconnected again through magnetic connectors. VME (right): PhET Circuit Construction Kit (Perkins et al. 2005) is a drag-and-drop ME for simple DC circuits. We de-activated the feature that explicitly visualizes the current flow in closed circuits to not provide users of the VME a conceptual advantage. Both MEs provide an ammeter and a voltmeter to measure current and voltage. Participants were given seven resistors to work with. There was no restriction on the number of wire elements

the frequency of oscillation and amplitude of oscillation?" In the Predict-Observe-Explain condition, participants were guided as follows: During the activity, the researcher asked a participant before each experiment to first decide on the target variable, and to predict the results of the intended experiment. After each experiment, they had to decide whether the prediction was confirmed or not, and explain why. The other condition received no guidance.

Writing Intervention The POE condition had to explain the POE framework in their own words, and apply it to another hypothetical activity. The other condition had to summarize what they did during the activity, and what their findings were. The purpose of this task was to make sure participants in the POE condition engaged cognitively with the idea of the POE framework again before starting the second activity.

Electric Circuit Activity Introduction: Participants were familiarized with the concepts and the ME. Participants were shown three basic electric circuits containing a single resistor, two resistors in series, and two resistors in parallel and told they could extend them by any combination of resistors available. Prompt of Activity: "Explore the relationship between how bright the bulb shines, the voltage across the bulb and the current through it. How is it affected by resistors? Explore by finding the resistor configurations that maximize and minimize the brightness of the light bulb." In each ME, participants started from an incomplete circuit as shown in Fig. 3b.

In both activities, participants could take notes. Participants were allowed to stop at any time they thought they had completed the task.

Subject Knowledge Assessment

Mass and Spring Activity The pre-test and the post-test had four multiple-choice questions, each with two subquestions. The first two questions addressed the impact of changing either the spring constant or the mass on the amplitude and frequency of oscillation. The third question targeted the understanding of force and speed in an oscillating spring-mass system. The fourth question was a near-transfer question inspired by the generalization questions of Renken and Nunez (2013), asking learners to apply their knowledge to a bungee jumping scenario.

Electric Circuit Activity The pre-test and the post-test consisted of two questions. The first question asked participants how adding a resistor in series (pre-test) or in parallel (post-test) to a circuit with one resistor would affect: the brightness of the light bulb; the current in the circuit; and the voltage across the bulb. The second question was a modified version of the assessment developed by McDermott and Shaffer (1992). Participants had to do pairwise comparisons of the brightness of light bulbs in five circuits with different

Int.	Medium	Mass & Spring					Electric Circuits			
		Pre -Test	Activity Introduction		Post - Test	Writing Task	Pre -Test	Activity Introduction	Activity without Guidance	Post-Test
No Guid.	VME n=18			Activity without Guidance						
	PME n=18									
POE	VME n=14			Activity with POE training						
	PME n=18									
Duration		5’	2’	10’	5’	5’	10’	2’	15’	15’

Fig. 4 Procedure and conditions of study 1. For more detail, see text

resistor configurations. There were nine such comparisons. See Supplementary Materials S.3 for the post-test items.

Pre-test and post-test items received a score of 1 for a correct answer, and 0 otherwise. Questions that required participants to explain their reasoning were given 0.5 for partially correct answers. The maximal scores were 8 (mass and spring) and 12 (electric circuits). Besides the overall aggregate score, we calculated also sub-scores for the two target concepts of each activity (spring constant and mass; series and parallel circuits).

Coding of Experimentation Strategies

We video-recorded or screen-captured each activity. From these recordings, we extracted for each participant each experimental configuration. We defined an experiment based on what variables participants manipulated: a new experiment started when one or more variables were changed, or it was repeated when the variable configurations did not change, but a new run was initiated. We coded the type of an experimental manipulation from the contrast between two succeeding experimental configurations (and the contrast of simultaneous experiments in the mass and spring activity).

For each activity, we defined the 25th percentile of the histogram of all dwell times between manipulations as the threshold time for intentionality. We coded any manipulation with a dwell time shorter than the 25th percentile as non-intentional. We use the same threshold of intentionality for both VME and PME instead of using different thresholds for each medium. As a consequence, if one ME requires by nature more time for setting up experiments, manipulations of that ME are more likely to be coded as intentional compared to the other ME. That way, we can capture potential learning or reflection processes that happen as participants are setting up experiments. The threshold times were 11 and 19 s respectively for the mass and spring and the electric circuits activity. In the mass and spring activity, the Median of dwell times was 20 s, with a maximum of 280 s. For the electric circuits activity, the Median was 32 s, with a maximum of 253 s.

As explained in Section "Target Concept and Intentionality: Adapting Simple Measures of Experimentation Strategies to more Complex Inquiry Tasks," we differentiated experimental manipulations based on the targeted concepts. In the Mass and Spring activity, we distinguished between manipulations of the spring constant and of the mass; in the Electric Circuits activity, we distinguished between comparisons of circuits with only non-parallel resistor configurations from comparisons with at least one circuit with parallel resistor configuration. See Supplementary Materials S.1 for more details of how we coded experimentation strategies.

Analysis

As illustrated in Fig. 2, we focused on experiments, resp. experimental manipulations, that were either (1) non-intentional manipulations, (2) intentional but confounded manipulations, or (3) intentional CV manipulations targeting different concepts. Experiments that were neither confounded nor CV were categorized as "other." For each participant, we calculate the proportion of occurrences of each of the tracked manipulations. The four main manipulation types accounted in average for 87.2% resp. 92.7% of a participants' manipulations in the Mass and Spring resp. Electric Circuits activities.

Multiple linear regression cannot capture the regularities in this multidimensional space of experimental manipulations. Rather, we used cluster analysis to find clusters of participants that look similar in terms of these dimensions of experimental manipulations. We took the following approach to examine the relation between MEs, experimentation strategies and learning outcomes as depicted in Fig. 1: We clustered participants into two groups based on the aggregate measures of experimentation strategies. We then examined how the clusters compared in terms of learning outcomes, to find the patterns of experimentation strategies that were productive, and then analyzed how participants in the different conditions split up between the clusters.

We used Portioning Around Medoids (PAM) for clustering (Reynolds et al. 2006), with cosine distance as the distance metric. We evaluated the quality of the clusters with the silhouette score (Rousseeuw 1987), a measure of similarity between points and the clusters they are assigned to.

For pairwise comparisons between variables that violated the normality assumptions, we report results from the nonparametric Mann-Whitney *U* test.

Results

In the following sections, we present the analyses of the conditions with respect to conceptual understanding and experimentation strategies. For comparisons of different operationalizations of experimentation strategies, across inquiry activities and studies, see Supplementary Materials S.2.7.

Mass and Spring Activity

This analysis includes only data for the participants that were not in POE because the intervention itself provided significant scaffolding. This leaves us with 34 participants, after excluding 2 participants with perfect pre-test scores.

We found that the conceptual understanding was higher for the PME, but that the difference was not significant: Linear regression of post-test scores on medium shows

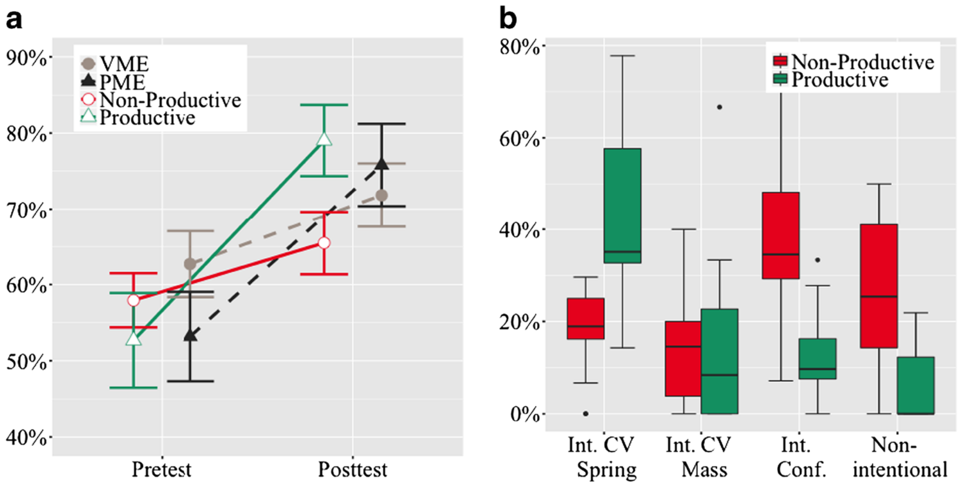

Fig. 5 Study 1 Mass and Spring Cluster Analysis: **a** Pre-test and post-test by cluster and condition of ME. Error bars indicate standard error. **b** Box plot of all manipulation types used to calculate the clusters: Intentional CV with at least one parallel circuit, or no parallel circuit, intentional confounded, and non-intentional manipulations. Numbers show the proportions of all manipulations per participant that are of a manipulation type

a positive but small effect, β_{medium}=.06, $t(31)$=1.0, p=.33, when controlling for pre-test scores, β_{pre}=.32, $t(31)$=1.9, p=.06, see Fig. 5a; i.e., the main effect was a 6% difference in post-test scores by medium. Including measures of experimentation strategy into the regression model, we find that intentional spring-only manipulation (ICVS) was a significant predictor of post-test scores, β_{ICVS}=.34, $t(29)$=2.1, p=.05, but not intentional mass-only manipulations (ICVM),β_{ICVM}=.03, $t(29)$=.2, $p>.5$.[3] This reflects also the difference in prior knowledge about the impact of the spring constant versus the mass on harmonic oscillations (M_{spring}=41.2%, SD=31.3%; M_{mass}=52.9%, SD=30.0%),[4] paired $t(33)=-1.54$, d=.4, p=.13. Adding CV manipulation to the baseline regression model was not significant, β_{CV}=.03, $t(30)$=.02, $p>.5$.

While these results suggest that VME and PME were equally effective in terms of conceptual understanding, we found differences in the use of experimentation strategies. As shown in Table 2, participants in PME used intentional CV strategies significantly more often, except for ICV manipulations of the mass (ICVM), and were more intentional overall.

We did not detect a direct effect of medium on conceptual understanding despite differences in productive experimentation strategies because of the large variation in strategies within condition. We addressed this issue by clustering participants as explained in Section "Target Concept and Intentionality: Adapting Simple Measures of Experimentation Strategies to more Complex Inquiry Tasks" (the measures shown in Fig. 5b). Cluster analysis finds groups of participants that look similar, i.e., that are close to each other in the multidimensional space that characterizes experimental manipulations. That way, cluster analysis can find regularities in the data despite the large variations in each variable within condition.

The cluster analysis gave rise to two distinct clusters, a *Productive* (n = 18) and a *Non-Productive* (n = 16) Cluster (avg silhouette score = .47). The naming of clusters is based on the test scores of each cluster. As shown in Fig. 5a, the Productive Cluster had significantly a better conceptual understanding post activity, as confirmed by regressing post-test scores on pre-test scores and cluster, β_{cluster}=.13, $t(31)$=2.6, p=.02.

Table 2 Study 1 mass and spring: experiment manipulations and target concepts by ME

	VME		PME		z	Sig	Eff
	Mdn	IQR	Mdn	IQR			
Experiment manipulations [%]							
CV	54.2	20.3	61.5	22.2	*−0.5*	*.6*	*.2*
ICV	29.7	22.4	56.5	24.3	−2.2	.02*	.8
ICVS	21.4	11.9	33.3	30.6	−2.3	.02*	.4
ICVM	10.5	20.0	14.3	20.0	−.3	.8	.05
Non-int.	17.7	28.6	0	16.7	2.6	.01*	.4
Target concepts [%]							
Spring	25.0	13.3	33.3	26.5	−1.9	.05*	.3
Mass	21.4	19.0	14.3	20.0	1.8	.08†	.3

Notes: †($p\le$.1), *($p\le$.05) , **($p\le$.01); *ICV* intentional CV, *ICVS/M* intentional CV of spring/mass, *Non-int.* non-intentional, *Mdn* median, *IQR* inter-quartile range; z = z-score of the Mann-Whitney U tests between the samples—when the samples were normally distributed, we report results of a t test instead, with 32 degrees of freedom (in *italic*). Effect size is calculated by $r = z/\sqrt{n}$ - Cohen's guidelines for r are .5 large effect, .3 medium effect, .1 small effect—when a t test is used, we calculated the effect size as Cohen's d

[3] The regression equation was marginally significant, $F(4, 29)$=2.2, p = .09, with an adjusted R^2=.13.

[4] M is the mean and SD the standard deviation of the sample

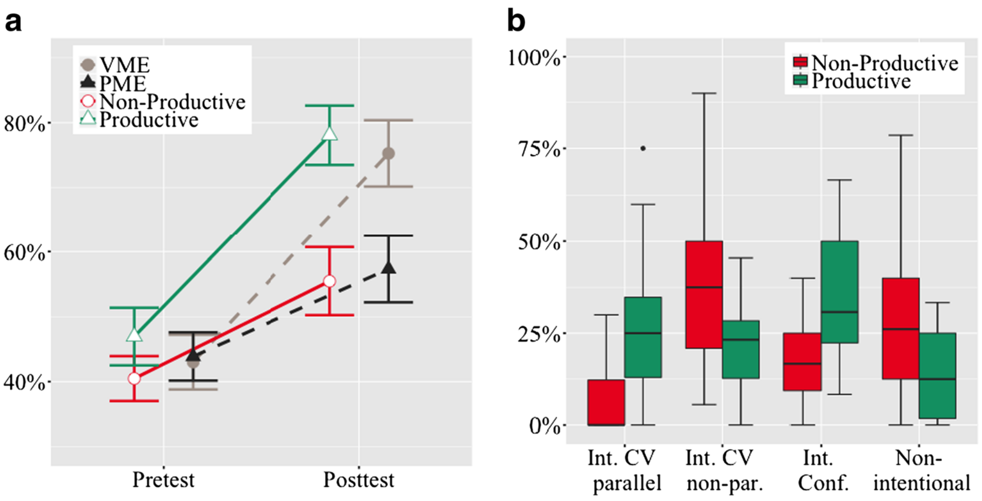

Fig. 6 Study 1 Electric Circuits Cluster Analysis: **a** Pre-test and post-test by cluster and condition of ME. Error bars indicate standard error. **b** Box plot of all manipulation types used to calculate the clusters: Intentional CV of the spring constant, or of the mass, intentional confounded, and non-intentional manipulations. Numbers show the proportions of all manipulations per participant that are of a manipulation type

Figure 5b shows that the Productive Cluster had significantly more ICVS, z=4.3, $p<.001$, r=.7, and less confounded manipulations, z=4.1, $p<.001$, r=.7. In particular, they had much less non-intentional manipulations, $z=-3.3$, $p<.001$, r=.6. Furthermore, the clusters differed in the proportion of experiments that targeted the spring constant ($M_{\mathrm{Prod.}}$=46.3%, SD=19.5%; $M_{\mathrm{Non\text{-}Prod.}}$ = 24.9%, SD=10.2%), $t(22.0)$=3.9, $p<.001$, d=1.4, and in experiments that targeted the mass constant ($\mathrm{Median}_{\mathrm{Prod.}}$= 8.4%, IQR=22.7%; $\mathrm{Median}_{\mathrm{Non\text{-}Prod.}}$=21.4%, IQR=16.8%), $z=-2.4$, p=.01, r=.4.

Finally, 82.3% of PME belonged to the Productive Cluster, compared to 11.8% of VME, $\chi^2(1, N=34)=17$, $p<.001$.

Electric Circuits Activity

We excluded three participants due to perfect pre-test scores, with 65 learners remaining for the analysis. We found no measurable effect of POE on conceptual understanding: A one-way ANCOVA of instruction and medium on post-test scores, controlling for pre-test scores, showed no overall effect of instruction, $F(1, 60)$=0.1, and no crossover effect, $F(1, 60)$=.2, p=.7. Similarly, we found no effect of POE on experimentation strategies (see Supplementary Materials S.2.1). Hence, we collapsed the data along the intervention dimension.

Conceptual Understanding by Medium VME performed significantly better than the PME, see Fig. 6a: Regressing post-test scores on pre-test scores, medium and their interaction revealed a significant main effect of medium, $\beta_{\mathrm{medium}}=-.18$, $t(61)=-2.8$, p=.006, and a marginally significant interaction of medium and pre-test scores, $\beta_{\mathrm{medium:pre}}$=.49, $t(61)$=1.7, p=.1.[5] This interaction suggests that participants with low prior knowledge benefited more from the VME. Participants overall had significantly lower prior knowledge on parallel circuits, M_{parallel}=29.6% (SD=33.3%) than on series circuits, M_{series}=70.6% (SD=34.6%), paired $t(64)=-7.5$, $d=-.4$, $p<.001$.

Conceptual Understanding and Experimentation Strategies Measures of productive experimentation strategies were significant factors for conceptual understanding: Adding the measures to the baseline regression model, and controlling for the number of circuits built, we find a significant effect for intentional CV manipulations with at least one parallel circuit (ICVP), β_{ICVP}=.61, $t(58)$=2.9, p=.005, but no effect for ICV with no parallel circuit (ICVNP), $\beta_{\mathrm{ICVNP}}=-.05$, $t(58)=-.3$, $p>.5$. In contrast, overall CV was not a significant predictor of post-test scores, β_{CV}=.2, $t(59)$=1.2, p=.2, similar to the mass and spring activity. For more details, see Supplementary Materials S.2.7.

Experimentation Strategies by Medium In line with the differences in conceptual understanding, VME and PME differed also in productive experimentation strategies. Table 3 shows that PME was significantly less intentional overall, which is particularly reflected in the amount of intentional CV. While there was no difference between conditions in the proportion of parallel circuits, virtual ME performed more ICVP.

Cluster Analysis of Experimentation Strategies Clustering participants based on the four different measures of experimentation strategies, see Fig. 6b, we found two well-defined clusters.[6] Regressing post-test scores on pre-test scores and cluster shows that participants in the *Productive*

[5] The regression equation was significant, $F(3, 61)$=8.4, $p<.001$, with an adjusted R^2=.26.

[6] The average silhouette score was .37. Nevertheless, the clusters are reasonably coherent in terms of experimentation strategies.

Table 3 Study 1 Electric Circuits: experiment manipulations and target concepts by ME

	VME		PME		z	Sig	Eff
	Mdn	IQR	Mdn	IQR			
Experiment manipulations [%]							
CV	65.0	28.6	50.0	25.2	2.3	.02*	.3
ICV	54.5	26.8	37.5	26.9	*3.9*	*.0003***	*1.*
ICVP	18.8	32.7	6.6	14.4	2.6	.001**	.3
ICVNP	33.3	20.4	25.0	27.8	*1.6*	*.12*	*.4*
Non-Int.	12.5	20.0	27.6	20.6	−2.8	.005**	.3
Circuit configurations [%]							
Parallel	33.3	26.3	18.8	46.8	.9	.4	.1
Series	61.9	24.8	74.6	38.8	−.8	.4	.1

Notes: †(p ≤.1), *(p ≤.05), **(p ≤.01); *Mdn* median, *IQR* interquartile range, z = z-score of the Mann-Whitney U tests between the samples- when the samples were normally distributed, we report results of a t-test instead, with 63 degrees of freedom (in *italic*). Effect size is calculated by $r = z/\sqrt{n}$ - when a t-test is used, we calculated the effect size as Cohen's d; for more information see Table 2

Table 4 Pairwise correlations of the key experimentation strategies between the two activities

M&S / EC	Int. CV Parallel	Int. CV Non-Par.
Int. CV Spring	−.48**	.27
Int. CV Mass	.29†	.07

Notes: †(p ≤.1), *(p ≤.05), **(p ≤.01); analysis based on 32 participants

Cluster (n=30) had on average a 19% higher score on the post-test than the *Non-Productive Cluster* (n=35), β_{cluster}=.19, $t(62)$=2.9, p=.004, controlling for pre-test scores, β_{pre}=.49, $t(62)$=3.3, p=.002; the regression model has an adjusted $R^2 = .24$.

We see in Fig. 6b that the Productive Cluster performed, proportionally, significantly more ICVP, z=4.7, p<.001, r=.6, less ICVNP, $t(58.9)$=−4.2, p<.001, d=1.0, however also more intentional confounded manipulations, z=4.0, p<.001, r=.5. The Non-Productive Cluster performed significantly more intentional manipulations overall, z=2.6, p=.01, r=.3.

Accordingly, 61.3% of VME belonged to the Productive Cluster, compared to only 32.4% of PME, $\chi^2(1, N=65)$=5.5, p=.02.

The Productive Cluster built significantly more parallel circuits (Median$_{\text{Prod.}}$=40.0%, IQR=21.4%) than the Non-Productive Cluster (Median$_{\text{Non-Prod.}}$=9.1%, IQR=21.3%), z=4.7, p<.001, r=.6. 6.7% of the Productive Cluster did not build any parallel circuit, compared to 45.7% of the Non-Productive Cluster. For the remaining participants in the latter cluster, the Median was 19.4% (IQR=27.5%).

As is to be expected, building parallel circuits improves assessed conceptual understanding: Regressing post-test scores on the proportion of parallel circuits, controlling for other baseline factors,[7] reveals a positive effect of parallel circuits, β_{parallel}=.25, $t(59)$=2.3, p=.03. But it cannot sufficiently explain the impact of ME on conceptual understanding, as there is no difference in the proportion of parallel circuits between media (see Table 3). Furthermore, the Productive Cluster built fewer circuits overall (Median$_{\text{Prod.}}$=11, IQR=6) than the Non-Productive Cluster (Median$_{\text{Non-Prod.}}$=16, IQR=10), z=2.3, p=.02, r=.3.

Overall, the cluster analysis indicates that the impact of medium on conceptual understanding relates to the impact of medium on experimentation strategies. Further evidence that experimentation strategies matter in understanding how ME impact inquiry-based learning comes from the mediation analysis in Supplementary Materials S.2.6, which revealed partial mediation by ICVP of 33.9% of the total effect of medium on post-test scores.

Experimentation Strategies Across Activities

We found that for both tasks, intentional control of variable manipulations that targeted the less familiar variables were significantly correlated with conceptual understanding. However, while users of the PME had a higher proportion of ICV in the first task, the opposite was true in the second task. Accordingly, Table 4 shows there was a significant negative correlation between the two relevant manipulation types, ICV involving parallel circuits (ICVP) and deliberate spring-only manipulations (ISO).[8]

This gives further support for the conclusion that the use of productive experimentation strategies in either task was less affected by domain-general characteristics of a participant than by the design features of the ME that might have been more or less conducive to the use of the productive strategies.

Discussion of Study 1

We set out with the hypothesis that the medium of ME affects the experimentation strategies, and that differences in strategy use can account for differences in conceptual understanding. The results of study 1 provide evidence in support of these hypotheses. As expected from the literature on inquiry strategies (Zimmerman 2000), higher use

[7] Specifically, we controlled for pre-test scores, medium, their interaction, and the number of circuits built.

[8] The correlations are based on a total of 32 participants, excluding participants from POE and participants with perfect pre-test scores.

of productive experimentation strategies led to higher conceptual understanding, in both inquiry tasks. However, the extent to which participants used productive experimentation strategies varied by medium of ME and by inquiry task: In the mass and spring activity, participants working with the PME were better in terms of strategy use than the ones using the VME, but this relation reversed in the electric circuits task. Furthermore, the cluster analysis for both activities confirm that strategy use could account for the impact of ME medium on conceptual understanding, at least partially.

Similar to Finkelstein et al. (2005) and Zacharia and de Jong (2014), we found that the physical equipment for electric circuits had detrimental effects on how participants went about the task. And similar to Renken and Nunez (2013), we found the opposite to be the true for the mass and spring activity. But as was the case with these studies, we cannot yet determine what affordances caused the differences in experimentation strategies, as the manipulative environments in each activity differed in multiple ways, see Table 5. We address this issue in study 2.

The results of study 1 hinge on the operationalization of experimentation strategies. In both activities, overall control of variable manipulation was neither significantly related to conceptual understanding nor did it explain differences in how participants went about the tasks. This is likely because certain manipulations in the VME, such as changing the spring constant simply by moving a slider, can be counted as CV even if the participant did not intentionally decide to control for variables.

By incorporating the dwell time between manipulations as a simple classifier of intentionality, and by distinguishing manipulations by the concepts they target, we found a reliable measure of productive experimentation strategies. Accordingly, intentional CV that target the less familiar concepts was a significant predictor of conceptual understanding, and a significant differentiator of participants across PME and the VME, in both activities.

However, it is not clear from Study 1 whether the time between manipulations represents the time for analyzing and reflecting on the experiment and inquiry, or whether it simply represents the time for setting up an experiment with the manipulative environment, or both. Study 2 sheds more light on the meaning of intentionality: With the measurement uncertainty as the only difference between the two virtual MEs, we expect the time to set up experiments to be the same for both. Thus, differences in time between manipulations are more likely due to factors other than the time for constructing and manipulating circuits. The Predict-Observe-Explain intervention did not help resolve this issue, as it did not have any measurable effect on participants.[9]

Table 5 Comparison of MEs in study 1 by types of affordance

Types of affordance	Mass and spring		Electric circuit	
	Comparison	Explanation	Comparison	Explanation
Ease of manipulation	VME>PME	Spring constant in VME is controlled via a simple slider.	VME<PME	The snap-fit design of PME is simpler than the drag-and-drop interface of VME.
Degrees of freedom of manipulation	VME>PME	The spring constant in the VME can take on continuous values.	VME = PME	
Confounding variables	VME = PME		VME < PME	PME has measurable internal resistance of wires.
Clarity of observation (Noise, etc.)	VME = PME		VME > PME	PME has noisy signal. Light bulb of VME is visually enhanced.

Notes: We analyzed the MEs by themselves, without considering user data. E.g. The first two cells mean that we considered the manipulation of variables to be easier in the VME than the PME for Mass & Spring as you could change the spring constant by moving a slider instead of manually replacing a spring

[9] We will discuss why POE did not show any effect in the section on limitations of the studies.

Study 2

In further analysis of the data for the electric circuits, we found that the number of circuits learners built was on average the same, but that participants working with the PME had a significantly lower percentage of unique circuits, because they did more repetitions and went back and forth between two circuit configurations (see Supplementary Materials S.2.4). We thought that one is more likely to engage in these types of behaviors when there is random noise in the data, because these behaviors allow learners to better discriminate the impact of an experimental manipulation (signal) from random noise. The noise existed only in the PME, and resulted in clearly detectable fluctuations in data readings and light bulb brightness. Thus, random noise decreased the clarity of observation and increased the complexity of information to be processed. We hypothesize that this accounted for a significant portion of the differences in experimentation strategies between MEs in Study 1.

System-inherent noise is generally seen as an important differentiator between virtual and physical ME that influences the clarity of observations (Chen 2010; Olympiou and Zacharia 2012; Renken and Nunez 2013), but the literature is ambiguous on its effect on learning outcomes. Some research found data free of random noise to be more conducive to conceptual understanding because they provide clearer observations that better expose cognitive conflicts between prior beliefs and observed evidence (Chinn and Malhotra 2002a; Toth et al. 2009). Other research found that VMEs free of noise might induce more "play"-like inquiry behaviors (Renken and Nunez 2013) and "may direct students to a naive thinking path that follows oversimplified logic or hypothetico-deductive reasoning" (p.1127, Chen 2010).

Either way, random noise is a representational affordance that changes the clarity of observation, which in turn likely influences how participants engage in the inquiry task. In study 2, we examine whether the presence or absence of random noise in the VME induces differences in experimentation strategies and conceptual understanding similar to the differences seen in the electric circuits activity in study 1.

General Study Design

The study was designed similarly to the electric circuits activity in study 1. Participants were assigned to one of the two conditions: Clear condition, in which participants used the unmodified version of the PhET Circuit Construction Kit (see "Manipulative Environments"); the noise condition, in which participants used a modified version of the same environment, where ammeter readings had some random noise. Similar to Chien et al. (2015), we programmed the random fluctuations in ammeter readings to match as closely as possible the noise observed in the physical toolkit used in Study 1.[10] The sample consisted of 60 students of the same community college as in study 1 (not the same students), that were randomly assigned to one of the two condition. The study was conducted individually.

Materials and Methods

Modifications in Procedure Compared to Study 1

Participants engaged in two activities about electric circuits. The first activity was the same as in study 1. We created a second activity in order to increase the potential impact of noise by requiring participants to focus more on small differences in ammeter readings. In that activity, participants had to use exactly two, three or six resistors to build circuits that matched either the highest or the lowest current they found when using only one resistor. The dependent variable in both activities was only the ammeter reading, and the circuits contained no light bulb or voltmeter.

Participants followed online instructions from a survey designed on Qualtrics.[11] This included videos that explained how to modify variables and build circuits using the VME, how to read the ammeter (without mentioning noise), and how to build four basic electric circuits using either a single resistor, two resistors in series, two resistors in parallel or a combination thereof. We showed these configurations again on the screen before starting the activity, as examples of the most elementary configurations they could build. Participants engaged in activity 1 for 12 min, followed by a re-set of the VME to initiate activity 2, which lasted for 10 min. They were allowed to stop anytime.

Modifications in Materials Compared to Study 1

Conceptual Knowledge Questions The questions were drawn from study 1, but extended to incorporate more decisions participants had to make in order to provide a more nuanced view of their conceptual understanding. We additionally asked participants to rate their confidence in their response. If participants indicated that they randomly guessed their answer, the score for that question was set to zero.

Noise-Related Question We included one question that targeted learners' interpretation of data in terms of range and uncertainty. It was a modified version of an item in the

[10] Random noise was designed as Gaussian noise centered on the theoretically correct current value, and spanning a maximum range of about 10% of the current value, up to a maximum of about 0.6Amp.

[11] Copyright 2017 Qualtrics. http://www.qualtrics.com.

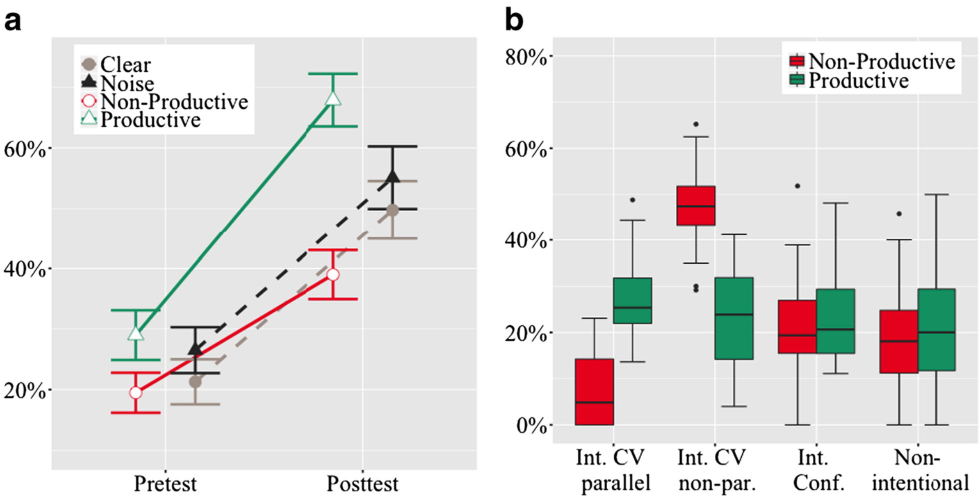

Fig. 7 Study 2 Cluster analysis: **a** Pre-test and post-test by cluster and condition of ME. Error bars indicate standard error. **b** Box plot of all manipulation types used to calculate the clusters: Intentional CV with at least one parallel circuit, or no parallel circuit, intentional confounded, and non-intentional manipulations. Numbers show the proportions of all manipulations per participant that are of a manipulation type

Physics Measurement Questionnaire (Buffler et al. 2001), adapted to electric circuits. The question presented participants with two noisy series of current readings of two circuits, and they had to evaluate based on that data whether the circuits were (1) the same, (2) probably the same, but more data was needed, (3) different, or (4) probably different, but more data was needed. The series consisted of six data points, and were designed such that their respective means were within each other's margins of error.

For more information on the assessment questions, see Supplementary Materials S.3.

Results

We excluded four participants in total: Three participants had perfect pre-test scores. One participant did not do the second activity. This left us with 56 participants in total, 28 in the Clear and 28 in the Noise condition.

Conceptual Understanding by Condition We found no differences in conceptual understanding between conditions: participants of both conditions started out at the same level of prior conceptual understanding, p=.4, and scored similarly on the post-test, p=.5, improving on average by 28.2%, $t(55)$=7.8, p<.01 (see Fig. 7a).

Table 6 Noise question: % participants per condition that evaluated the circuits being compared as either the same or different

Answers	Pre-test		Post-test	
	Clear	Noise	Clear	Noise
The same / Probably the same	39%	32%	25%	64%
Different / Probably different	61%	68%	75%	36%

However, participants' interpretation of noisy data evolved differently between conditions, see Table 6: Prior to the activity, there was no difference between conditions, p>.5; post activity, more participants in Noise thought the circuits were the same than in Clear, $\chi^2(1, N{=}56){=}8.7$, p=.003.

Conceptual Understanding and Experimentation Strategies Similar to the analysis of the electric circuits activity in study 1, we regressed post-test scores on intentional control of variable manipulations that include at least one parallel circuit (ICVP) or no parallel circuit at all (ICVNP), with the baseline covariates of pre-test scores, condition and number of circuits, $F(5, 50)$=5.7, p<.001, with an adjusted R^2 of .30. As in Study 1, ICVP was a significant predictor of post-test scores, β_{ICVP}=.68, $t(50)$=2.1, p=.04, but not ICVNP, β_{ICVNP}=−.28, $t(50)$=−1.1, p=.3.

Table 7 Study 2: evaluation of differences in experiment manipulations and target variables by ME

	Clear		Noise		t	Sig	Eff
	M	SD	M	SD			
Experiment manipulations [%]							
ICV	50.9	13.9	58.1	13.2	−2.3	.06†	.5
ICVP	13.5	12.9	19.9	12.3	*−2.0*	*.04**	*.3*
ICVNP	37.4	18.1	38.2	17.1	−.2	.9	.1
Non-Int.	23.1	12.3	17.2	10.9	1.9	.07†	.5
Circuit configurations [%]							
Parallel	24.3	23.2	29.1	19.4	*1.0*	*.3*	*.1*
Series	75.7	23.2	70.9	19.5	*1.1*	*.3*	*.1*

Notes: †(p ≤.1), *(p ≤.05) , **(p ≤.01); M Mean, SD Standard Deviation, t statistics from the two sample t test with 54 degrees of freedom—when the samples were not normally distributed, we report the z-score of the Mann-Whitney U tests (in *italic*). Effect size is calculated by Cohen's d (unless samples were not normally distributed, see Table 2)

Experimentation Strategies by Condition As shown in Table 7, the Clear condition was less intentional overall, with less intentional CV manipulations (ICV), and in particular less ICV with at least one parallel circuit (ICVP). There was no significant difference neither in overall CV manipulations, nor in the number of circuits per participant, p=.2, nor in the types of circuits they built.

Cluster Analysis of Experimentation Strategies Similar to the mass and spring activity in study 1, despite the differences in non-intentional manipulations and in ICVP, conceptual understanding was the same across conditions. Thus, we again expected cluster analysis to provide a more nuanced picture of how the different conditions went about the tasks.

The four key manipulation types shown in Fig. 7b covered on average 96.7% of manipulations for each participant. Clustering participants based on those manipulations, we found two well-distinguished clusters (avg silhouette=.48). As in the electric circuits activity in study 1, we found that *Productive Cluster* (n=26) scored better on the conceptual tests than the *Non-Productive Cluster* (n=30), see Fig. 7a. Regressing post-test scores on pre-test scores, β_{pre}=.31, $t(53)$=2.0, p=.05, and cluster gives a significant main effect for cluster, β_{cluster}=.26, $t(53)$=4.4, $p<$.001. There was a marginally significant difference in pre-test scores between the Productive Cluster (Median=34.6%, IQR=30.8%) and the Non-Productive Cluster (Median=13.5%, IQR=30.8%), z=1.6, r=.2, p=.1.

The difference in how the two conditions split across the two clusters was marginally significant, $\chi^2(1, N=56)$=2.6, p=.1: 57.1% of the Noise participants fell into the Productive Cluster compared to 35.7% of the Clear participants.

Figure 7b shows that the Productive Cluster performed more ICVP than the Non-Productive Cluster, $z=-6.0$, r=.8, $p<$.001, and less ICVNP, z=5.9, r=.8, $p<$.001; yet, there was no difference in the proportion of non-intentional manipulations, $t(54)$=.8, d=.1, p=.5.

The differences in experimentation strategies between clusters are related to differences in the proportion of parallel circuits the participants built: Participants in the Productive Cluster generated significantly more parallel circuits (Median=46.2%, IQR=25.0%) than in the Non-Productive Cluster (Median=9.1%, IQR=16.9%), z=5.9, r=.8, $p<$.001. 33.3% of the participants in the Non-Productive did not build any parallel circuits; for the remaining participants, the median of parallel circuits was 15.5% (IQR=10.2%). However, in contrast to the electric circuits activity in study 1, the Productive Cluster also built more circuits overall ($\text{Median}_{\text{Prod.}}$=36, IQR=16.8%) than the Non-Productive Cluster ($\text{Median}_{\text{Non-Prod.}}$=28.5, IQR=11.5%), z=2.2, p=.03, r=.3.

Discussion of Study 2

We have employed study 2 to evaluate how the presence or absence of system-inherent noise affects how learners use a virtual ME in inquiry-based learning.

While results confirmed that the amount of intentional control manipulations with at least one parallel circuit was the strongest factor for post-test scores (similar to study 1); but we were not able to re-create the differences in conceptual understanding and experimentation strategies found between the virtual and physical ME in study 1. On the contrary, it seems that the noisy data readings did not harm the learners as postulated by other studies (Olympiou et al. 2013; Renken and Nunez 2013). Results even suggest that participants in the Noise condition performed better than in the Clear condition when considering the quality of their experimentation strategies.

On first sight, it might seem counter-intuitive that the presence of noise had an impact on strategy use, because noise only changes the signal-to-noise ratio in the data but not the interactive possibilities of the ME. However, the fundamental difference in experimentation strategies between the two conditions was not in the quality of experiments, but in the intentionality of manipulations: Unlike Renken and Nunez (2013), both conditions had the same overall proportion of unconfounded experiments, but the proportion of intentional, unconfounded experiments was higher in the Noise condition. Systems in which variables can be manipulated easily reduce the time between manipulations by reducing the time to set up an experiment. This might explain why in the electric circuits activity in study 1, users of the VME were slower than users of the PME. However, in study 2, the time it took to build a circuit configuration was likely the same between the conditions, which suggests that any difference in dwell time can be attributed to different forms of engagement with the experiment. The presence of noise might have acted as a decelerator, slowing down participants because more time was needed to decide what effect a manipulation had on the current. The lack of noise in the clear VME however might have induced faster manipulations because the result of each experiment was immediately evident. Alternatively, the presence of noise might have increased participants' cognitive engagement with the data, requiring them to be cognitively more active in observing the experiments. In the present study, we cannot disentangle the different mechanisms of how random noise influenced participants' experimentation strategies; but we have clear indication that there was a cognitive effect, because of the differences between conditions in the test questions about random noise. Further studies are required to see whether participants developed different strategies to cope with fluctuating values during the activity, or whether they just became more tolerant in assessing such uncertainty.

Either way, the additional cognitive effort due to the noise did not harm learners' conceptual understanding, even if drawing inferences from noisy data is considered to be more difficult (Chinn and Malhotra 2002a; Toth et al. 2009).

General Discussion

In this paper, we presented a series of studies to address two issues related to the use of manipulative environments in inquiry-based learning. One issue involves a methodological argument about how to study the effectiveness of manipulative environments. The other issue concerns the relation between manipulative environments, their affordances, experimentation strategies, and conceptual learning. The results indicate that:

1. Manipulative environments affect experimentation strategies (see Fig. 8).
2. Experimentation strategies mediate the effect of manipulative environment on conceptual understanding. Participants using the manipulative environment that induced more productive experimentation strategies were more likely to belong to the cluster of participants with better conceptual understanding.
3. The effect of manipulative environments on experimentation strategies depends on their affordances; for example, the presence of noise in the data can induce different experimentation strategies (see Fig. 8c).

The Importance of Experimentation Strategies when Studying the Impact of Medium of ME on Conceptual Understanding

One aim of this paper was to examine to what extent accounting for students' experimentation strategies can help explain why they learn differently with different manipulative environments, and not to show that systematic experimentation facilitates learning in general, which has already been known (Zimmerman 2000). We demonstrated that the evaluation and comparison of manipulative environments can be more informative and consistent if experimentation strategies are measured together with conceptual understanding. There are multiple indications in support of this argument:

First, in line with prior work (Zimmerman 2000), measures of productive experimentation strategies were strongly and positively correlated factors of conceptual understanding irrespective of the task domain and manipulative environment. Differences in conceptual understanding were related to differences in experimentation strategies, either between conditions of MEs or between clusters of participants across conditions. Notably in the electric circuits task in study 1, experimentation strategies partially mediated the effect of medium of ME on conceptual understanding.

Second, we found little consistency in strategy use within learner across activities in study 1. This suggests that the way participants went about the activity was influenced not just by individual and task-related factors, but also by characteristics of the manipulative environments, which is supported by the results of study 2.

Finally, conditions that appeared to be the same based on conceptual understanding alone turned out to be different when considering participants' experimentation strategies. This was the case in both the mass and spring activity in study 1, and the electric circuits task in study 2. In both studies, participants developed on average a comparable conceptual understanding of mass and spring resp. electric circuits concepts. Based on that, we could have argued that there was no difference in benefits of MEs for learning, in line with previous research (Klahr et al. 2007; Olympiou and Zacharia 2012; Pyatt and Sims 2011; Triona and Klahr 2003).

However, cluster analysis of experimentation strategies revealed clear differences between the virtual and physical MEs in each activity: Participants in the productive strategies cluster consistently had higher learning outcomes in all activities; in each activity, the proportion of participants in the productive strategies cluster was different for each ME; and the ME with more participants in the productive strategies cluster tended to be better for learning outcomes. The difference in split of participants across clusters did not always translate into significant learning differences because of the following reasons: 1. In the mass & spring

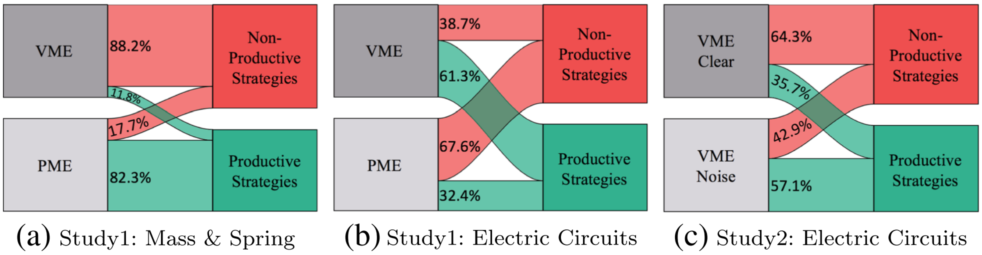

Fig. 8 Breakdown of conditions by cluster of experimentation strategies, for each inquiry activity in studies 1 and 2. In each case, the productive cluster had higher learning outcomes

activity in Study 1, significantly more participants using the PME were in the productive strategies cluster, yet the sample size was too small to reduce the inter-subject variance in experimentation strategies within each condition. Only very few participants in each condition deviated from the others (see Fig. 8). 2. In Study 2, the split of participants across the clusters was only marginally significant because the distinction in only one affordance between the two MEs was rather subtle. How a ME impacts the inquiry processes is also influenced by the nature of activity; future research is needed on the interaction of activity design and ME design on experimentation strategies.

The advantage of cluster analysis compared to multiple regression approaches is that it extracts regularities in learners' experimentation strategies across multiple dimensions simultaneously and thus increases the signal-to-noise ratio in the data.

One might argue based on our analyses that the difference in learning between the Productive and Non-Productive Clusters was more a matter of exposure to the right experiment configurations rather than a matter of systematicity of the experimentation strategies. After all, the proportion of parallel circuits participants built was a strong predictor of conceptual understanding and differentiating factor between the clusters, in both study 1 and study 2. It makes sense that if you have never built a parallel circuit, for example, you will likely not know what the impact of resistor configuration is on an electric circuit. However, building the "right" experiment configurations is not sufficient in explaining the impact of MEs on learning outcomes *in general*, as the findings from the electric circuits activities did not translate into the mass and spring activity; the proportion of manipulations targeting the spring constant or mass were neither predictive of conceptual understanding nor strongly differentiating factors between clusters. Systematic experimentation strategies are more consistently related to positive learning outcomes.

The Impact of Single Affordances on Students' Experimentation and Learning in Inquiry-Based Activities

The findings of study 2 about the beneficial effects of noise on participants' experimentation strategies and subsequent reasoning indicate two things: (1) individual affordances can influence how learners engage in and learn from inquiry activities in ways that are not necessarily anticipated by a priori analysis of MEs. Against our expectations, noisy data *increased* participants' use of intentional manipulations. (2) the aggregate effect of all affordances of a manipulative environment is not necessarily summative, but rather the result of complex interactions between the affordances. This could explain why in study 2 we found that participants used more productive experimentation strategies in the presence of noise, yet in study 1, where the MEs differed along multiple dimensions of affordance (see Table 5), participants had worse strategy use when working with the physical ME, which inherently had noisy data.

Therefore, approaches to determine post-hoc what affordances caused differences between learners in comparisons of physical and virtual MEs can be misleading, in particular in light of the many differences in affordances between environments (de Jong et al. 2013). Rather, we propose to consistently study the role of affordances in inquiry-based learning by targeting single affordance, and compare manipulative environments that differ only in the target affordance. Further studies to examine what caused the learning differences found in study 1 might need to employ such a factorial design for every single affordance, such as the ease of manipulation, *and/or* complexity of the models.

The Relevance of Intentionality in Finding Productive Experimentation Strategies

We focused on the control of variable strategy because it provides a simple yet robust enough measure of generally productive experimentation strategies to address our research questions. However, operationalizing CVS as mere frequency counts of CV manipulations proved insufficient to find relevant patterns in experimentation strategies that were productive for learning. Rather, *intentionality*, as defined by the time between manipulations, was a particularly important characteristic of productive experimentation strategies.

The relevance of intentional manipulations, i.e. of having "enough" time between manipulations stands in contrast to previous work that considered fast manipulations to be more beneficial for developing conceptual understanding by exposing learners to more experiments, and by reducing time to run experiments to free time for the conceptually relevant aspect (Zacharia et al. 2008). One potential explanation for this contradiction is that the number of *possible* experimental configurations was significantly lower in earlier work than in the current study; in these studies, learners could cover the entire experimental space in a short amount of time with even simple trial-and-error.

We think that intentionality is relevant in activities that give learners more choices and more flexibility in their exploration. When learners are free to choose how to go about an activity, both in terms of what kind of the actions to take and how to take them, time between manipulations seems to capture cognitive aspects of the inquiry process that are relevant for learning. A recent study by Perez et al. (2017) confirms our findings. They analyzed action sequences of students in an inquiry task with electric circuits, and found that strategic use of pauses between building and testing circuits was strongly associated with successful learning. In their study, pauses are indicative of active cognitive processes.

Intentionality also seems to capture how interactive and representational affordances of MEs affect how learners engage with them. In both activities of study 1, participants in the Non-Productive Clusters were less intentional than participants in the Productive Clusters. And in both activities, the MEs we considered as being easier to manipulate were more conducive to non-intentional manipulations—the VME for mass and spring allowed to change the spring constant simply by using a slider, and the PME for electric circuits allowed to quickly change circuits by snapping together magnetically connected pieces.

However, in study 2, where the ease of manipulation was the same for both MEs, the clusters did not differ in terms of overall intentional manipulations. Instead, the difference in clarity of observation mainly impacted the intentional control of variable manipulations, but not the overall control of variable manipulations. This indicates that the cognitive engagement of participants was different with these more *informative* experiments. As elaborated in section "Discussion of Study 2", further research is needed to understand the mechanism of how clarity of observation impacted the learning experiences.

Limitations and Future Research

A major limitation of our studies was the length of each activity. Ten to 22 min is a short time span for inquiry activities, especially when there exist inaccurate prior beliefs. The process of developing accurate conceptual understanding ideally requires longer time spans with multiple iterations of the discovery activities (Renken and Nunez 2013). Furthermore, such short durations punish behaviors that might be desired under other circumstances. For example, we observed in study 1 that some participants using the physical toolkit for electric circuits explored the impact of wire length on current and voltage; while this was irrelevant to the learning objective of the task, it could be interpreted as a sign of attentiveness and curiosity that eventually could have led to a deeper understanding about electric circuits. But in short tasks, any time spent on aspects irrelevant to the learning goal leaves less time for goal-relevant ones. At the same time, these short durations bear some ecological validity, because active learning tasks in introductory college classes and recitation sessions are typically not much longer than the activities in our study.

Another limitation was the implementation of the Predict-Observe-Explain intervention in study 1. There are many possible reasons for why the POE intervention did not have any effect on learning outcomes, such as the difficulty of transfer from the mass and spring to the electric circuit activity. We think that a longer intervention might have been more informative.

In both studies, the within-condition variabilities in participants' use of experimentation strategies were quite large. The corresponding reduction in signal-to-noise ratio could have been compensated for by larger sample sizes.

We think that for future research on how affordances of MEs influence inquiry strategies, additional measures of inquiry strategies than the ones used in this paper could provide a more fine-grained picture. Our measures of experimentation strategies represent average strategies that were aggregated from all manipulations per participant, ignoring the temporal dynamics of the inquiry behaviors. Recent work used action sequence mining to reveal patterns of actions that were associated with productive learning (Perez et al. 2017), or sequence labeling of experimental configurations to extract different styles of exploration in the experiment space (Levy and Wilensky 2011). Additionally, advanced sensing and artificial intelligence technologies can be used to explore more complex dynamics and properties of inquiry behaviors (Worsley and Blikstein 2015).

Conclusions

"Are interactive simulations or physical manipulative environments better for inquiry-based learning?" Although research studies have been used to support both sides of this debate, we believe that this is the *wrong* question to ask.

We have shown that whether or not a ME is effective is likely a question of affordances rather than medium of environment and that a ME is effective if it is conducive to productive inquiry strategies. Using cluster analysis of a multi-dimensional representation of experimentation strategies, we found that differences in affordances can cause participants to cluster differently based on strategy use, and that the clusters correlate strongly with that the degree of conceptual understanding. Even when ME seemed to be equally conducive to learning in terms of learning outcomes, we found differences in experimentation strategies between the conditions; these differences were just not big enough to show up in differences in learning outcomes.

To date, many design decisions for manipulative environments are made on a case-by-case basis, guided by heuristics that stem from research contrasting virtual and physical MEs. However, comparing manipulative environments at the level of medium is not sufficient for informing productive designs because MEs of the same medium can still vary significantly in terms of their affordances (e.g., Electricity Exploration Tool (Jaakkola and Nurmi 2008) versus PhET Circuit Construction Kit (Perkins et al. 2005)). Affordances are also not "hard-wired," domain-general characteristics of the medium of a ME, as new technologies allow us to implement in one medium affordances that were considered as unique to the other. For instance, the affordance for fast manipulations of variables, traditionally seen as inherent to virtual manipulative environments, can be altered by turning a simple drop-down menu into a more complex interaction element.

We believe that this paper presents a type of analysis that can contribute to developing a general framework of productive design principles for MEs that focuses on affordances independent of medium. Such analysis will clear up many of the apparent ambiguities and inconsistencies in the existing literature, which did not look in detail at how manipulative environments influence inquiry strategies, and what role affordances play independent from the medium.

References

Ainsworth, S., & VanLabeke, N. (2004). Multiple forms of dynamic representation. *Learning and Instruction*, *14*(3), 241–255.

Buffler, A., Allie, S., Lubben, F. (2001). The development of first year physics students' ideas about measurement in terms of point and set paradigms. *International Journal of Science Education*, *23*(11), 1137–1156.

Chan, J., Pondicherry, T., Blikstein, P. (2013). LightUp: an augmented, learning platform for electronics. In *Proceedings of the 12th international conference on interaction design and children, IDC '13* (pp. 491–494). New York: ACM.

Chen, S. (2010). The view of scientific inquiry conveyed by simulation-based virtual laboratories. *Computers & Education*, *55*(3), 1123–1130.

Chen, Z., & Klahr, D. (1999). All other things being equal: acquisition and transfer of the control of variables strategy. *Child development*, *70*(5), 1098–1120.

Chen, S., Chang, W.-H., Lai, C.-H., Tsai, C.-Y. (2014). A comparison of students' approaches to inquiry, conceptual learning, and attitudes in simulation-based and microcomputer-based laboratories. *Science Education*, *98*(5), 905–935.

Chien, K.-P., Tsai, C.-Y., Chen, H.-L., Chang, W.-H., Chen, S. (2015). Learning differences and eye fixation patterns in virtual and physical science laboratories. *Computers & Education*, *82*, 191–201.

Chini, J.J., Madsen, A., Gire, E., Rebello, N.S., Puntambekar, S. (2012). Exploration of factors that affect the comparative effectiveness of physical and virtual manipulatives in an undergraduate laboratory. *Physics Review ST Physics Education Research*, *8*(1), 010113.

Chinn, C.A., & Malhotra, B.A. (2002a). Children's responses to anomalous scientific data: how is conceptual change impeded? *Journal of Educational Psychology*, *94*(2), 327–343.

Chinn, C.A., & Malhotra, B.A. (2002b). Epistemologically authentic inquiry in schools: a theoretical framework for evaluating inquiry tasks. *Science Education*, *86*(2), 175–218.

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

Conlin, L. (2014). Supporting middle schoolers' use of inquiry strategies for discovering multivariate relations. In Polman, J.L., Kyza, E.A., O'Neill, D.K., Tabak, I., Penuel, W.R., Jurow, A.S., O'Connor, K., Lee, T., D'Amico, L. (Eds.) *Interactive physics simulations*. Boulder.

de Jong, T., Linn, M.C., Zacharia, Z.C. (2013). Physical and virtual laboratories, in science and engineering education. *Science*, *340*(6130), 305–308.

Duschl, R.A., & Grandy, R.E. (2008). *Teaching scientific inquiry: recommendations for research and implementation*. Sense Publishers.

Finkelstein, N.D., Adams, W.K., Keller, C.J., Kohl, P.B., Perkins, K.K., Podolefsky, N.S., Reid, S., LeMaster, R. (2005). When learning about the real world is better done virtually: a study of substituting computer simulations for laboratory equipment. *Physical Review Special Topics - Physics Education Research*, *1*, 1.

Fischer, K.W. (1980). A theory of cognitive development: the control and construction of hierarchies of skills. *Psychological Review*, *87*(6), 477–531.

Gire, E., Carmichael, A., Chini, J.J., Rouinfar, A., Rebello, S., Smith, G., Puntambekar, S. (2010). The effects of physical and virtual manipulatives on students' conceptual learning about pulleys. In *Proceedings of the 9th international conference of the learning sciences - volume 1, ICLS '10* (pp. 937–943). Chicago: International Society of the Learning Sciences.

Gunstone, R.F., & Mitchell, I.J. (2005). Metacognition and conceptual change. In *Teaching science for understanding* (pp. 133–163). Elsevier.

Heradio, R., de la Torre, L., Galan, D., Cabrerizo, F.J., Herrera-Viedma, E., Dormido, S. (2016). Virtual and remote labs in education: a bibliometric analysis. *Computers & Education*, *98*, 14–38.

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

Hsu, Y.-S., & Thomas, R.A. (2002). The impacts of a web-aided instructional simulation on science learning. *International Journal of Science Education*, *24*(9), 955–979.

Huppert, J., Lomask, S.M., Lazarowitz, R. (2002). Computer simulations in the high school: students' cognitive stages, science process skills and academic achievement in microbiology. *International Journal of Science Education*, *24*(8), 803–821.

Jaakkola, T. (2012). Thinking outside the box: enhancing science teaching by combining (instead of contrasting) laboratory and simulation activities.

Jaakkola, T., & Nurmi, S. (2008). Fostering elementary school students' understanding of simple electricity by combining simulation and laboratory activities. *Journal of Computer Assisted Learning*, *24*(4), 271–283.

Kanari, Z., & Millar, R. (2004). Reasoning from data: how students collect and interpret data in science investigations. *Journal of Research in Science Teaching*, *41*(7), 748–769.

Kearney, M., Treagust, D.F., Yeo, S., Zadnik, M.G. (2001). Student and teacher perceptions of the use of multimedia supported predict–observe-explain tasks to probe understanding. *Research in Science Education*, *31*(4), 589–615.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*(1), 1–48.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction effects of direct instruction and discovery learning. *Psychological Science*, *15*(10), 661–667.

Klahr, D., Triona, L.M., Williams, C. (2007). Hands on what? The relative effectiveness of physical versus virtual materials in an engineering design project by middle school children. *Journal of Research in Science teaching*, *44*(1), 183–203.

Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science*, *16*(11), 866–870.

Kuhn, D., Iordanou, K., Pease, M., Wirkala, C. (2008). Beyond control of variables: what needs to develop to achieve skilled scientific thinking? *Cognitive Development*, *23*(4), 435–451.

Lazonder, A., & Ehrenhard, S. (2014). Relative effectiveness of physical and virtual manipulatives for conceptual change in science: how falling objects fall. *Journal of Computer Assisted Learning*, *30*(2), 110–120.

Levy, S.T., & Wilensky, U. (2006). Gas laws and beyond: strategies in exploring models of the dynamics of change in the gaseous state. In Buckley, B., Gobert, J., Kraicik, J. (Eds.) *Supporting science learning and science education reform with information technologies*. San Francisco: Paper presented at the National Association for Research in Science Teaching.

Levy, S.T., & Wilensky, U. (2011). Mining students' inquiry actions for understanding of complex systems. *Computers & Education*, *56*(3), 556–573.

Marshall, J.A., & Young, E.S. (2006). Preservice teachers' theory development in physical and simulated environments. *Journal of Research in Science Teaching*, *43*(9), 907–937.

McDermott, L.C., & Shaffer, P.S. (1992). Research as a guide for curriculum development: an example from introductory electricity. Part I: investigation of student understanding. *American Journal of Physics*, *60*(11), 994–1003.

McElhaney, K. (2010). Making controlled experimentation more informative in inquiry investigations.

McElhaney, K.W., Chang, H.-Y., Chiu, J.L., Linn, M.C. (2015). Evidence for effective uses of dynamic visualisations in science curriculum materials. *Studies in Science Education*, *51*(1), 49–85.

Olympiou, G., & Zacharia, Z.C. (2012). Blending physical and virtual manipulatives: an effort to improve students' conceptual understanding through science laboratory experimentation. *Science Education*, *96*(1), 21–47.

Olympiou, G., Zacharias, Z., de Jong, T. (2013). Making the invisible visible: enhancing students' conceptual understanding by introducing representations of abstract objects in a simulation. *Instructional Science*, *41*(3), 575–596.

Parnafes, O. (2006). Developing conceptual understanding through the use of computer-based representations. In *The learning man in the technological era, proceedings of the 1 St Chais conference on instructional technology research*. Israel: Open University Press.

Paulson, A., Perkins, K.K., Adams, W.K. (2009). How does the type of guidance affect student use of an interactive simulation. *Physical Review Special Topics-Physics Education Research*, In review, 141–158.

Pedaste, M., Mäeots, M., Siiman, L.A., de Jong, T., van Riesen, S.A.N., Kamp, E.T., Manoli, C.C., Zacharia, Z.C., Tsourlidaki, E. (2015). Phases of inquiry-based learning: definitions and the inquiry cycle. *Educational Research Review*, *14*, 47–61.

Perez, S., Massey-Allard, J., Butler, D., Ives, J., Bonn, D.A., Yee, N., Roll, I. (2017). Identifying productive inquiry in an exploratory virtual labs using sequence mining. In: *Proceedings of the 18th international conference on artificial intelligence in education*. Wuhan.

Perkins, K., Adams, W., Dubson, M., Finkelstein, N., Reid, S., Wieman, C., LeMaster, R. (2005). PhET: interactive simulations for teaching and learning physics. *The Physics Teacher*, *44*(1), 18–23.

Pyatt, K., & Sims, R. (2011). Virtual and physical experimentation in inquiry-based science labs: attitudes, performance and access. *Journal of Science Education and Technology*, *21*(1), 133–147.

Quinn, H., Schweingruber, H., Keller, T. (2012). *A framework for K-12 science education, practices, crosscutting concepts, and core ideas*. National Academies Press.

Renken, M.D., & Nunez, N. (2013). Computer simulations and clear observations do not guarantee conceptual understanding. *Learning and Instruction*, *23*, 10–23.

Reynolds, A.P., Richards, G., de la Iglesia, B., Rayward-Smith, V.J. (2006). Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematics Modelling Algorithm*, *5*(4), 475–504.

Rickey, D., & Stacy, A.M. (2000). The role of metacognition in learning chemistry. *Journal of Chemical Education*, *77*(7), 915.

Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Salehi, S., Keil, M., Kuo, E., Wieman, C.E. (2015). How to structure an unstructured activity: generating physics rules from simulation or contrasting cases. In Churukian, A.D., Jones, D.L., Ding, L. (Eds.) *2015 physics education research conference* (pp. 291–294). College Park.

Sandoval, W.A., & Reiser, B.J. (2004). Explanation-driven inquiry: integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education*, *88*(3), 345–372.

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology*, *32*(1), 102.

Schauble, L., Glaser, R., Raghavan, K., Reiner, M. (1992). The integration of knowledge and experimentation strategies in understanding a physical system. *Applied Cognitive Psychology*, *6*(4), 321–343.

Siegler, R.S., & Crowley, K. (1991). The microgenetic method. A direct means for studying cognitive development. *American Psychologist*, *46*(6), 606–620.

Toth, E.E., Morrow, B.L., Ludvico, L.R. (2009). Designing blended inquiry learning in a laboratory context: a study of incorporating hands-on and virtual laboratories. *Innovative Higher Education*, *33*(5), 333–344.

Triona, L.M., & Klahr, D. (2003). Point and click or grab and heft: comparing the influence of physical and virtual instructional materials on elementary school students' ability to design experiments. *Cognition and Instruction*, *21*(2), 149–173.

van der Meij, J., & de Jong, T. (2006). Supporting students' learning with multiple representations in a dynamic simulation-based learning environment. *Learning and Instruction*, *16*(3), 199–212.

Windschitl, M. (2000). Supporting the development of science inquiry skills with special classes of software. *ETR&D*, *48*(2), 81–95.

Winn, W., Stahr, F., Sarason, C., Fruland, R., Oppenheimer, P., Lee, Y.-L. (2006). Learning oceanography from a computer simulation compared with direct experience at sea. *Journal of Research in Science Teaching*, *43*(1), 25–42.

Worsley, M., & Blikstein, P. (2015). Leveraging multimodal learning analytics to differentiate student learning strategies. In *Proceedings of the fifth international conference on learning analytics and knowledge, LAK '15* (pp. 360–367). New York: ACM.

Zacharia, Z.C. (2007). Comparing and combining real and virtual experimentation: an effort to enhance students' conceptual understanding of electric circuits. *Journal of Computer Assisted Learning*, *23*(2), 120–132.

Zacharia, Z.C., & de Jong, T. (2014). The effects on students' conceptual understanding of electric circuits of introducing virtual manipulatives within a physical manipulatives-oriented curriculum. *Cognition and Instruction*, *32*(2), 101–158.

Zacharia, Z.C., & Michael, M. (2016). Using physical and virtual manipulatives to improve primary school students' understanding of concepts of electric circuits. In Riopel, M., & Smyrnaiou,

Z. (Eds.) *New developments in science and technology education, number 23 in innovations in science education and technology* (pp. 125–140): Springer International Publishing.

Zacharia, Z.C., Loizou, E., Papaevripidou, M. (2012). Is physicality an important aspect of learning through science experimentation among kindergarten students? *Early Childhood Research Quarterly*, *27*(3), 447–457.

Zacharia, Z.C., Manoli, C., Xenofontos, N., de Jong, T., Pedaste, M., van Riesen, S.A.N., Kamp, E.T., Mäeots, M., Siiman, L., Tsourlidaki, E. (2015). Identifying potential types of guidance for supporting student inquiry when using virtual and remote labs in science: a literature review. *Educational Technology Research and Development*, *63*(2), 257–302.

Zacharia, Z.C., Olympiou, G., Papaevripidou, M. (2008). Effects of experimenting with physical and virtual manipulatives on students' conceptual understanding in heat and temperature. *Journal of Research in Science Teaching*, *45*(9), 1021–1035.

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review*, *20*(1), 99–149.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*(2), 172–223.