# Learning Environments and Inquiry Behaviors in Science Inquiry Learning: How their Interplay Affects the Development of Conceptual Understanding in Physics

Engin Bumbacher* buben@stanford.edu
Shima Salehi* salehi@stanford.edu
Miriam Wierzchula miriamw1989@gmail.com
Paulo Blikstein* paulob@stanford.edu

* Stanford University, CERAS 102, 520 Galvez Mall, Stanford, CA, 94305

## ABSTRACT

Studies comparing virtual and physical manipulative environments (VME and PME) in inquiry-based science learning have mostly focused on students' learning outcomes but not on the actual processes they engage in during the learning activities. In this paper, we examined experimentation strategies in an inquiry activity and their relation to conceptual learning outcomes. We assigned college students to either use VME or PME for a goal-directed physics inquiry task on mass-spring systems. Our analysis showed that the best predictors of learning outcomes were experimental manipulations that followed a control of variable (CV) strategy, with a delay between manipulations ("systematic inquiry"). Cluster analysis of the prevalence of these manipulations per participant revealed two distinct clusters of participants, systematic inquiry or not. The systematic inquiry cluster had significantly higher learning outcomes than the less systematic one. Furthermore, the majority of the participants using the PME belonged to the more systematic cluster, while most of the participants using the VME fell into the non-systematic cluster, likely because of the specific affordances of the real and virtual equipment they were using. However, beyond this impact on inquiry process, condition had little effect. In light of these results, we argue that investigating processes displayed during learning activities, in addition to outcomes, enables us to properly evaluate the strengths and weaknesses of different learning environments for inquiry-based learning.

## Keywords

Science Discovery Learning, Computer Simulations, Real Laboratories, Inquiry Learning, Cluster Analysis, Virtual and Physical Science Laboratories

## 1. Introduction

Over the past decades, the science teaching community has adopted the view that "students cannot fully understand scientific and engineering ideas without engaging in the practices of inquiry and the discourses by which such ideas are developed and refined" (NRC, 2012, p.218). Inquiry-based instruction requires students to model the practices of scientific inquiry to actively develop their conceptual understanding [1,2]. While physical laboratories were the traditional environments for such inquiry-based learning, there is accumulating evidence that virtual laboratories are similarly well suited to meet the goals of science investigation [3,4]. In particular, they are considered to be at least equally conducive to active manipulations for experimentation [2,3], which is seen as the crucial aspect of inquiry learning [5,6,7].

A major limitation of the research comparing physical and virtual manipulative environments (PME and VME) for science learning was the predominant focus on the learning *outcomes* rather than the learning *processes* when students engage in inquiry activities. This has not changed with recent work that shifted from treating the environments as two competing entities to examining how to best combine them for increased learning benefits [4]. We argue that research on how learners engage with these manipulative environments could provide a more comprehensive understanding of how the interaction of a learner with an environment impacts the learners' construction of knowledge, and in turn what design features of these environments foster desired manipulative behaviors in the context of science inquiry learning.

The present study lies at the intersection of research on learning environments and research on inquiry behaviors in order to study the characteristics of productive experimentation strategies in open-ended science investigation tasks, and how such strategy use might be influenced by the different affordances of the learning environments. For this purpose we encoded the actual experiments students ran, which allows us to basically replay their processes. This allows us to explore customized operationalizations of inquiry behaviors of interest. This approach integrates data-driven methods with relevant theoretical concepts. As a result, we found a robust characterization of experimentation strategies that meaningfully predicts learning outcomes, and show how participants' strategy use differs between the learning environments. This study is part of a larger research project with the goal of developing automated detectors of systematic inquiry in open-ended science investigation activities for formative assessment and for the design of productive learning environments.

## 2. Inquiry Behaviors

### 2.1. Control of Variable Strategy

Scientific learning through self-directed inquiry activities depends on the actual inquiry behaviors employed [8,9]. In particular, adequate experimentation strategies are required that result in interpretable observations, i.e. evidence that facilitates drawing valid inferences. Research has particularly focused on the abilities to systematically combine variables and to design unconfounded experiments, i.e. experiments that modify variables such that competing hypothesis can be ruled out. The design of unconfounded experiments requires the ability to employ the *control of variables strategy* (CVS), that is, to create experiments with a single contrast between experimental conditions [10]. This is in contrast to inadequate strategies such as changing multiple variables at the same time, which hampers valid inferences and subsequent knowledge [11].

Previous research has examined a host of individual and contextual factors of strategy use [8]. However, only a very small number of studies have explicitly examined the impact of affordances of learning environments on strategy use in experimentation activities [2,12]. While Triona & Klahr [2] focused on the impact of physicality of manipulatives alone on

learning outcomes, Renken & Nunez [12] had students engage in an inquiry activity on pendulum motion using either a PME or a VME that differed in both ease of manipulation and freedom of choice: while the PME provided participants with only three different levels for either pendulum length or mass, the VME allowed participants to modify the variables smoothly by means of continuous valued sliders. Even if there was no difference in conceptual understanding between the VME and PME conditions, participants using the computer simulation ran more trials and were less likely to control variables. Renken and Nunez [12] argued that the additional flexibility and breadth of choice in experimentation in VME was detrimental to participants' use of adequate experimentation strategy.

While this study suggests that indeed strategy use in inquiry-based learning activities is influenced by affordances of the learning environments, it is difficult to generalize these results to less structured and scaffolded inquiry activities.

## 2.2. Operationalization of Inquiry Strategies

As most studies cited mainly focused on CV strategy, they used highly structured tasks where either variables were dichotomous, or there was only one outcome variable, or the activity was restricted. In order to develop a more nuanced characterization of inquiry strategies, we need more complex inquiry tasks. Data mining techniques employed in such contexts have been successful at discovering groups of similar users [13,14,15]. Most of these data-mined systems are based on the user interaction logs [16]. While they achieve good predictive power, such machine-learned detectors of interaction behaviors often come at the cost of interpretability [17]. However, it is crucial to develop data-mined models of inquiry strategies that are interpretable in order to advance our understanding of learning processes through inquiry activities. We apply a different approach, where we do not use labelled action logs but code the actual experiment configurations of each participant. Based on video data, we extract each configuration a participant tried and feed it into a database of experiments of all participants. This allows us to quickly extract and explore relevant variables of inquiry, such as the number of spring-only or mass-only changes, the number of unique configurations, repetitions, etc. That way, we can integrate relevant theoretical concepts into the operationalization of inquiry behaviors.

In the context of this study, we focused on experimentation strategies only. We collected data on the number of experiment trials, the experiment configurations, and the time between manipulations, and coded the type of manipulation per experiment. Particular focus is given to "*control of variable" manipulations, "deliberate" manipulations, and "deliberate control" of variable manipulations*. Deliberate manipulations (DM) are manipulations into which a participant has put some thought, as measured by *dwell time* between two consecutive manipulations. We assume that participants who are cognitively engaged – reflecting on evidence from a preceding manipulation, trying to make sense of it in the context of previous observations, or taking notes or planning the next manipulation(s) – will spend more time before executing the next change than those who are cognitively less engaged.

For this reason, we include the third category of manipulations that lies at the intersection of the prior two categories, *deliberate control of variable manipulations* (DCVM). As prior research on experimentation strategies in inquiry-based activities characterized them as solely CVS or not, activities were designed such that controlling variables in an experiment had to be a deliberate choice of participants [19,20,21]. However, in less structured, open-ended inquiry like those used in this study, it is possible in some cases to manipulate variables according to a CV strategy without the deliberate intention to do so. For example in the computer simulation for our mass & spring activity, one could change the value of the spring constant continuously by means of a slider, without having to interrupt an ongoing experiment. Inherently, this corresponds to a control of variable manipulation (CVM) but not necessarily to a *deliberate* control of variable manipulation (DCVM).

# 3. Present Study

The study reported here was part of a larger study examining participants' inquiry behaviors in different scientific domains using either PME or VME as learning environments. Participants engaged in two activities; the first one was on mass and spring oscillation (see Figure 1), and the second one on basic electric circuits. The current paper presents analysis of the first inquiry activity. During the first activity, participants were either asked to simply think-aloud while engaging in the inquiry or were trained to implement the Predict-Observe-Explain framework (POE) [18]. The training session of the POE framework was highly structured and guided: During the entire activity, before each intended manipulation, participants were asked to predict its result, then observe the actual results of the manipulation, and finally explain their observation in light of the initial prediction. On the other hand, the think-aloud group did not receive any scaffolds or guidance by the experimenter. Therefore, for the purposes of this paper, we report only data for the participants in the think-aloud condition, as the difference in guidance might have altered the nature of the activity, and masked the effect of medium on inquiry processes of the participants.

The main research questions that guided the present study were:

- How can we operationalize inquiry strategies in less well-structured and more complex activities?
- What inquiry strategies are related to better learning outcomes?
- How does strategy use differ between participants using either the physical or the simulation environment?

## 3.1. Sample

For Mass and spring activity in think-aloud condition, we had 36 community college students (24 female, 12 male, average age=20.5, SD=3.6).

## 3.2. Design

The study reported here is a between subject design with two levels. We randomly assigned participants to use either *physical* (PHY) or *computer simulation* (SIM) to engage in an inquiry-based activity on mass and spring oscillation ($n_{PHY}$=18, $n_{SIM}$=18). The task was to discover how the mass and the spring constant affect both the amplitude and the frequency of oscillation of a mass-spring system. We administrated a conceptual test before and after the activity. The post-test scores were the dependent measures of the experiment, while the pre-test scores were used as covariates in the corresponding statistical analyses. The relevant

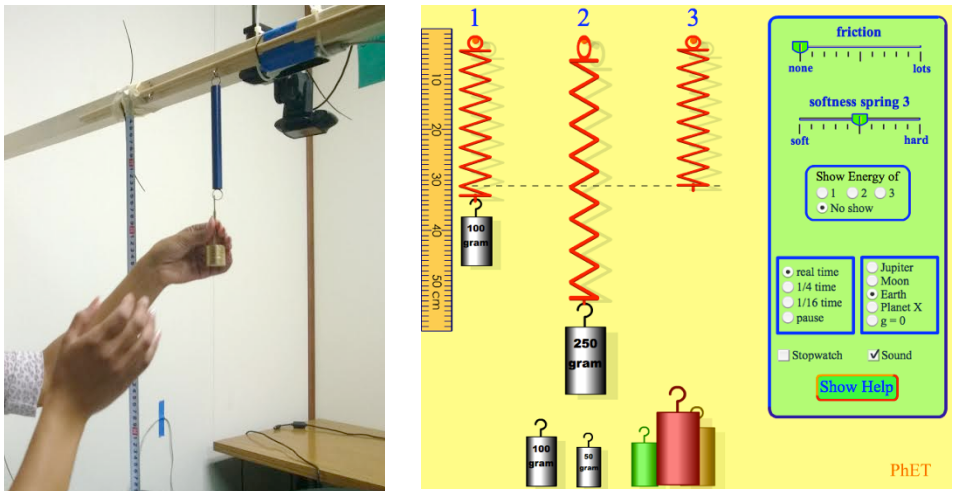


Figure 1. Experimental Setup: Left: Physical toolkit in action: The first hook is just next to the measure tape. Right: Computer Simulation: Participants were only allowed to change the "softness spring 3".

behavioral measures were treated as independent within-subject variables since they were expected to predict learning outcomes.

## 3.3. Materials

### *3.3.1. Learning Environment*

**Physical Learning Environment**. The physical toolkit consisted of the PASCO[1] Demonstration Spring Set and Mass and Hanger Set. There are four pairs of springs, each with a spring constant between 4 N/m and 14 N/m. The masses consist of hangers to which slices of weights can be attached, ranging from 5 to 20 g. The environment consisted of two hooks, each being able to hold one spring, see Figure 1. For measuring extensions and duration, we provided a measuring tape and a stopwatch.

**Simulation Learning Environment.** The computer simulation we used was created by PhET [22], see Figure 1. It consists of three springs, two of which have a fixed and equal spring constant. The spring constant of the third spring can be changed continuously by means of a slider. It further entails seven weights, four of which are 50g, 100g, 100g and 250g respectively. The other three have no indication of their actual weight. The weights can be attached to and removed from the springs by simple drag-and-drop. The simulation comes with a displaceable measuring tape as well as a stopwatch.

**Differences in Learning Environment.** Instead of designing the learning environments ourselves, we selected the ones that we considered as state of the art of their respective domains. This prevented us from setting up the necessary control of the differences in affordances of the environments for making causal claims about the relation of learning environment and experimentation strategies. However, we can reason about the potentially relevant differences based on the specific user interfaces and interaction designs. The main differences are the following ones: 1. The VME allows participants to use up to three springs, compared to two in the PME; 2. In the PME, participants could change the spring constant of both springs if needed, while the VME allowed to change the spring constant of only the third spring; 3. In the VME, manipulating the spring constant is easier as it requires only changing the value of a continuous valued slider. Participants could change its value on the fly, without interrupting an ongoing experiment. In order to change the spring constants in the PME, participants had to stop an experiment, and physically replace a spring with another one.

[1] PASCO scientific, 10101 Foothills Boulevard, P O Box 619011, Roseville, Ca 95678-9011, USA.Web: http://www.pasco.com. E-mail: sales@pasco.com. National representatives of PASCO can be reached through the USA office.

### *3.3.2. Subject Knowledge Assessment Questionnaire*

The pre-test and the post-test consisted of four qualitative questions, each with two sub-questions. 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 & Nunez [12].

### *3.3.3. Procedure.*

Students participated individually in the study. They were assigned randomly to either the PHY or the SIM condition. Prior to taking the pre-test, each participant was introduced to the nature and goal of the activity, and to definitions of relevant variables. Possible experiments were restricted only by the given set of weights and springs. The definition sheet contained basic definitions, both verbal and visual, of relevant concepts of harmonic oscillation of mass-spring systems. After the pre-test, the experimenter explained how to manipulate the variables and how to perform measurements, depending on condition using either the physical toolkit or the computer simulation. Participants were instructed to adjust only the settings related to the two variables of interest. They were further asked to think-aloud during the activity. The maximal duration of the inquiry task was 10 minutes. Participants then completed the post-test. Both pre-test and post-test took 5 minutes each.

## 3.4. Coding

### *3.4.1. Conceptual Tests*

Pre-test and post-test items received a score of 1 if they were correctly answered, and 0 otherwise. Questions that required participants to explain their reasoning were given 0.5 for partially correct answers. The maximum score was 8. Besides the overall aggregate score, we calculated also sub-scores for the two conceptual categories, spring constant (two items) and mass dependence (two items).

### *3.4.2. Inquiry Behaviors*

In a first pass, we extracted every experiment a participant ran from the corresponding video records of the experiment. This was done manually. Once the database was established, we could code every experiment computationally based on customized rules for

extracting relevant variables such as number of manipulated objects, etc. Even if the initial step was done by hand, the extraction procedure was operationalized such that we can automatize this process for future iterations: An experiment was characterized by the state of each relevant variable. A new experiment started when either one or more variables of the system were manipulated, or when a current experimental setup was re-initiated, either by touching a mass-spring system with the hand or with the mouse. The type of performed manipulation was then extracted from the contrast between two experiments. All variables representing inquiry behaviors are coded proportionally, relative to the total number of experiments run per activity.

An experiment consisted of the number of springs used, their spring constants, and the weights attached to the springs. The possible manipulations were (1) change of the spring constant, (2) change of the weight, (3) change both, (4) repeat an experiment, and (5) start a new experiment by changing the number of springs used. Changing either the mass only or the spring only corresponded to a *control of variables manipulation* (CVM), while a *confounded manipulation* referred to changing both variables at the same time. In cases participants used only one mass-spring configuration, we defined an experimental comparison through the contrast set up by the configurations in two consecutive runs. When two configurations were used simultaneously, the experimental comparison was defined by the contrast of those two sets of masses and springs. When participants in the SIM condition used all three springs, we defined the experimental comparison by the *most optimal* contrast out of the three possible pairwise combinations (optimal being the mass-spring configurations that differ only in one independent variable).

**Table 1. Regression Models of Post-Test Scores**

| ***Variables / Models*** | ***1*** | ***2*** | ***3*** | ***4*** | ***5*** |
|---|---|---|---|---|---|
| (Intercept) | 3.79*** | 3.12** | 1.09 | 2.07* | 2.01* |
| | (0.68) | (0.24) | (1.38) | (0.16) | (0.96) |
| Pre-test Scores | 0.32$^{\dagger}$ | 0.32$^{\dagger}$ | 0.29$^{\dagger}$ | 0.32$^{\dagger}$ | 0.34* |
| | (0.17) | (0.17) | (0.16) | (0.16) | (0.16) |
| Condition | 0.33 | 0.49 | -0.05 | 0.38 | 0.36 |
| | (0.33) | (0.44) | (0.35) | (0.37) | (0.37) |
| % Control of Variable | | 0.89 | | | |
| | | (1.60) | | | |
| % Confounded | | 1.28 | | | |
| | | (2.17) | | | |
| % Delib. Manip. | | | 3.33* | | |
| | | | (1.50) | | |
| % Delib. CV | | | | 3.17* | |
| | | | | (1.44) | |
| % Delib. Confounded | | | | 3.21 | 3.29 |
| | | | | (2.09) | (2.06) |
| % Delib. Spring-Only | | | | | 3.95** |
| | | | | | (1.52) |
| % Delib. Mass-Only | | | | | 1.46 |
| | | | | | (1.86) |
| $R^2$ | 0.113 | 0.127 | 0.238 | 0.254 | 0.304 |
| *adj.* $R^2$ | 0.056 | 0.007 | 0.162 | 0.151 | 0.179 |
| *N* | 34 | 34 | 34 | 34 | 34 |

*Note*: Standard error are in parentheses; † ($p \leq 0.1$), * ($p \leq 0.05$), ** ($p \leq 0.01$) ), *** ($p \leq 0.001$); each model regresses post-test scores on the given independent variables.

As explained before, just looking at whether an experiment was unconfounded or not misses out on other relevant aspects. In particular, such a perspective does not provide any insights into how deliberately or considered participants executed and reflected on an experiment. Therefore, we additionally captured the duration of each experiment as the *dwell time* between two succeeding experimental manipulations. Based on the dwell time, we developed a *measure of deliberateness*; any manipulation that had a dwell time bigger than first quartile of all dwell times of all participants was coded as a *deliberate manipulation*.

## 3.5. Data Analysis.

### *3.5.1. Analysis of Learning Outcomes*

In order to analyze the relation between inquiry behaviors and learning outcomes, we ran multiple linear regressions on post-test scores, with condition as independent factor, pre-test scores as covariate, and the corresponding measures of inquiry behavior as independent variables. For pairwise comparisons between variables within the same category that violated the normality assumptions, we report results from the nonparametric Mann-Whitney-Wilcoxon test.

### *3.5.2. Analysis of Inquiry Behaviors*

We applied a cluster method on all experimental manipulation variables to group participants by their inquiry behaviors. We used portioning around medoids (PAM) as the clustering algorithm, which is a more robust version of the standard k-means clustering algorithm, as it minimizes a sum of dissimilarities instead of a sum of squared Euclidian distances [23]. The quality of the clustering result was evaluated based on the silhouette score [24], a measure of similarity between points and the clusters they are assigned to. The larger the silhouette value, the better the clustering. However, instead of selecting the clusters that maximize the silhouette score, we have to make a trade-off between silhouette score and number of clusters in order to have theoretically relevant results. Ideally, we could set the number of clusters to 2, as we were interested in analysis of behaviors with respect to condition.

# 4. Results

## 4.1. Baseline Knowledge

Participants in the two conditions did not differ significantly in pre-test scores, $t(32) = 1.49$, $p = 0.15$ (PHY: M = 3.53, SD = 1.59; SIM: M = 4.23, SD = 1.15). However, the high overall pre-test score average of about 52.5% of the maximal possible score indicates that participants had relevant prior knowledge with regards to the subject. We excluded two participants who scored perfectly on the pre-test. In terms of prior knowledge related to impact of the spring constant versus the mass on harmonic oscillations, there were no significant differences in pre-test scores on the corresponding subcategories (Spring constant: M = 41.2%, SD = 31.3%; Mass: M = 52.9%, SD = 30.0%), paired $t(33) = -1.54$, $d = 0.38$, $p = 0.13$. However, as the trend in data nevertheless points in the expected direction, we classify experiments that involve spring manipulations as less familiar than those involving mass manipulations.

## 4.2. Effect of Condition on Learning Gain

The two conditions were not significantly different in terms of average learning outcomes as condition was not a significant

factor for post-test scores, controlling for pre-test scores, $\beta = 0.33$, $t(32) = 1.01$, $p = 0.32$, $\eta_p^2 = 0.03$ (see Figure 2.B.).

## 4.3. Learning Outcome by Inquiry Behaviors

We examined how various measures of inquiry behaviors related to learning outcomes by multiple linear regression analysis. The baseline variables of each regression model were condition as independent factor, and pre-test score as covariate. All the corresponding regression models are shown in Table 1.

### *4.3.1. Time on Task and Number of Experiments*

While time on task was the same across conditions, $t(32) = 0.28$, $p > 0.5$, the total number of experiments per participant was higher for the SIM condition (M = 18.7, SD = 8.3) than for the PHY condition (M = 13.7, SD = 7.3), $d = 0.64$, $t(32) = 1.87$, $p = 0.07$. Additionally, pre-test scores were not correlated with number of experiments, $r(32) = -0.05$, $p > 0.5$. An ANCOVA suggests that the number of experiments was not a significant factor for post-test scores, controlling for pre-test scores, $F(1, 30) = 0.02$, $p > 0.5$, $\eta_P^2 < 0.01$. Overall, participants performed 533 different experiments, based on which we built the database.

### *4.3.2. Control of Variables Manipulations*

We did not find a significant effect for overall CVM on post-test scores, $\beta = 0.89$, $t(31) = 0.33$, $p > 0.5$ (see model 2 in Table 1). Even when looking at mass-only or spring-only manipulations, the respective regression coefficients are not significantly different from zero. These results indicate that performing control of variable manipulations of either the spring or the mass does not necessarily lead to better learning outcomes per se, which is in contrast to the prior literature [8]. We find that control of variable manipulations alone cannot explain the variability in learning outcomes both within and across conditions.

### *4.3.3. Deliberate Manipulations*

We coded the deliberateness of an experimental manipulation by means of the time spend on an experiment. We extracted the duration between manipulations across all participants, and defined the cut-off value between a *rapid* and a *deliberate* manipulation as the 25th percentile of the duration histogram (Mdn = 20 seconds). This was at 11 seconds.

Overall deliberate manipulations was a relevant positive predictor of post-test scores, $\beta = 3.33$, $t(31) = 2.21$, $p = 0.03$, $\eta_P^2 = 0.14$ (model 3 in Table 1). While CVM was not relevant for learning outcomes, deliberate control of variable manipulations (DCVM) was a significant factor in the regression model 4 in Table 1, $\beta = 3.17$, $t(31) = 2.19$, $p = 0.04$, $\eta_P^2 = 0.15$. This effect was mainly driven by deliberate spring-only manipulations (see model 5 in Table 1). On the other hand, deliberate confounded manipulations had a comparably high coefficient value, even if it was not significant. With an adjusted $R^2 = 0.18$, $F(5,28) = 2.44$, $p = 0.06$, model 5 did not explain a higher proportion of variance than model 3, $F(1,2) = 1.32$, $p = 0.28$.

None of the manipulation types correlated with pre-test scores (all correlation coefficients were lower than 0.1 in absolute value). The lack of correlation supports the claim that the manipulations were context-dependent variables of inquiry behavior.

## 4.4. Inquiry Behavior by Condition

### *4.4.1. Control of Variables Manipulations and Deliberate Manipulations*

The physical and the simulation condition did not differ in terms of control of variables manipulations, d = 0.14, $t(32) = -0.08$, $p = 0.94$ (SIM: M = 0.51, SD = 0.13; PHY: M = 0.53, SD = 0.16). In contrast to that, the two conditions differed significantly in the amount of deliberate control variable manipulations (DCV), $d = 0.77$, $t(32) = 2.23$, $p = 0.033$ (SIM: M = 0.35, SD = 0.15; PHY: M = 0.47, SD = 0.18). There is a significant drop in CV when considering the deliberate manipulations for the SIM condition only. In line with the hypothesis that the simulation environment was easier to manipulate, there were significantly more rapid manipula-tions in the SIM condition (Mdn = 17.6%, $CI_{.95} = \pm 24.5\%$) than in the PHY condition (Mdn = 0% , $CI_{.95} = \pm 12.9\%$), $U = 219.5$, $r = 0.46$, $p = 0.007$.

### *4.4.2. Cluster Analysis of Inquiry Behaviors*

Overall, DCV manipulations were a significant predictor for learning outcomes, in particular the deliberate spring-only manipulations. However, even if there was a significant difference in the amount of these manipulations between the PHY and SIM conditions, learning outcomes did not differ significantly by

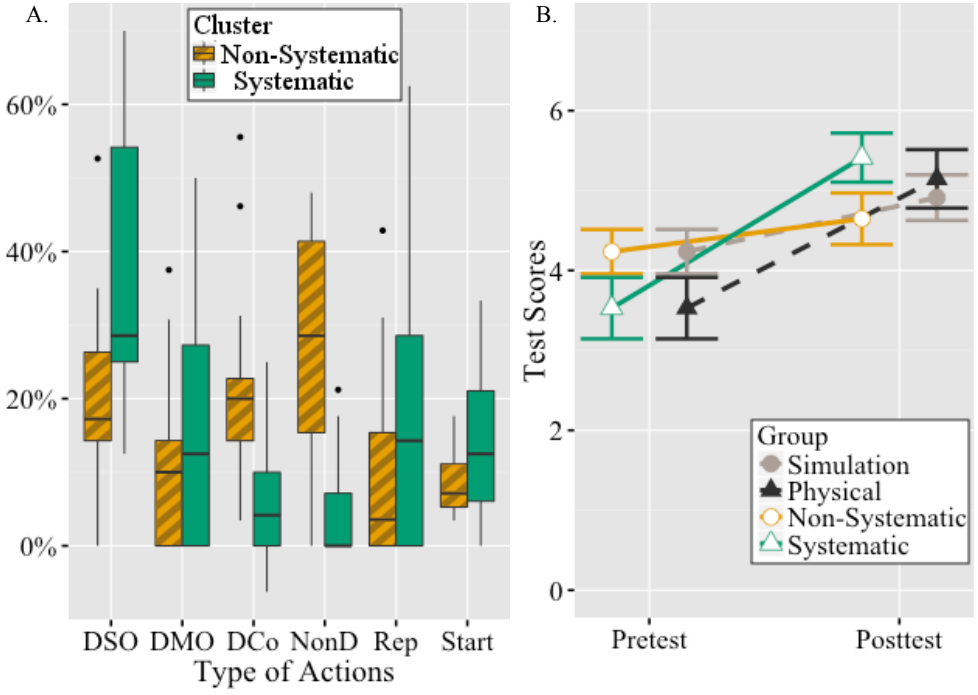


Figure 2. A. Boxplot of proportions of deliberate spring-only (DSO), deliberate mass-only (DMO), deliberate confounded (DCo), and non-deliberate (NonD) manipulations, repetitions (REP) and start of new experiments (Start). B. Comparison of pre-test and post-test scores by cluster as well as condition. Bars indicate standard errors.

condition. It appears that individual differences in inquiry strategies of participants within each condition washed out the actual impact of the learning environment on average post-test scores. There might be people in the physical and the simulation condition that deviated from the average inquiry behaviors for the condition towards the other condition's characteristics. We address this question by grouping all participants by considering all inquiry variables simultaneously instead of grouping them by condition, and then see how the groups distribute across the conditions. This can be done by means of cluster analysis.

Clustering was performed on the 6 possible manipulation types (see Figure 2.A) of the entire sample, which resulted in 2 clusters with 17 participants in each cluster. The average silhouette score was 0.30. While this score is not high enough to exclude the possibility of artificial data structures, an examination of the clusters in terms of variables confirms the clusters reasonably distinguish people by the level of systematicity of their inquiry behaviors: Generally, the participants of *Cluster 1 ("non-systematic")* were less strategic and less deliberate in their manipulations than *Cluster 2 ("systematic")* (see Figure 2.A). Cluster 2 had a higher proportion of deliberate spring-only manipulations than Cluster 1, $U = 58$, $r = 0.51$, $p = 0.002$, a lower proportion of non-deliberate manipulations than Cluster 1, $U = 262.5$, $r = 0.72$, $p < 0.001$, and a lower proportion of confounded manipulations, $U = 240.0$, $r = 0.57$, $p < 0.001$. There was no significant difference in the other variables. Additionally, even if the clustering was not performed on overall DCV, there is a large difference between the clusters; participants in the systematic cluster proportionally performed significantly more DCV (Mdn = 49.8%, $CI_{.95} = \pm 16.3\%$) than in the non-systematic cluster (Mdn = 30.7%, $CI_{.95} = \pm 12.1\%$), $d = 1.33$, $t(32) = 3.89$, $p < 0.001$.

The two clusters meaningfully differ in learning outcomes, as indicated by a regression of post-test scores on the cluster variable, with pre-test scores as covariates, which revealed a significant main effect of cluster, $\beta = 1.03$, $t(31) = 2.39$, $p = 0.015$, $\eta_P^2 = 0.16$. As expected, participants in the systematic scored higher than those in the non-systematic cluster (see Figure 2.B). The regression model explained a significant proportion of variance, adjusted $R^2 = 0.18$, $F(2,31) = 4.55$, $p = 0.02$.

**Table 2. Conditions distributed across clusters**

| **Condition** | Non-Systematic (n = 17) | Systematic (n = 17) |
|---|---|---|
| Physical (n = 17) | *3 (17.6%)* | *14 (82.4%)* |
| Simulation (n = 17) | *14 (82.4%)* | *3 (17.6%)* |

Finally, Table 2 shows that the majority of participants in the systematic cluster used the physical toolkit, while the majority of participants that belonged to the non-systematic cluster were in the simulation condition, as confirmed by a Fisher's exact test, $p < 0.0001$.

# 5. DISCUSSION

Considerable attention has been given separately to research on the impact of virtual and physical learning environment [4] and of inquiry behaviors on the learning outcomes in science discovery activities [8,9]. The aim of the present study was to link these two realms by (1) studying the relation of strategy use and learning outcomes, and (2) comparing strategy use between learning environments in order to shed light on how different affordances of the learning environments might influence strategy use.

## 5.1. Nuanced View of Experimentation Strategies in Open-Ended Inquiry Tasks

One main finding from this study was that one of the strongest predictors for learning outcomes when controlling for prior knowledge was the manipulation type that (a) created a single contrast in experiment conditions, (b) targeted the problem type that participants generally were less familiar with, and (c) was deliberate. In the context of the mass and spring activity, these were deliberate manipulations that changed only the spring constant from one mass-spring system to the other.

Importantly, this further implies that the control of variables (CV) in experiment design was a necessary but not sufficient condition for developing conceptual understanding through experimentation. This is in contrast to prior research that has predominantly focused on the ability to design unconfounded experiments as the main factor of knowledge acquisition in inquiry learning [2,10,12]. Using control of variable strategy as an important factor for characterizing experimentation strategies works when the student has to make a conscious decision to actually apply this strategy. It fails if the affordances of the user interface do not require that. In the computer simulation, one could change the spring constant continuously using a slider, even during an ongoing experiment. In the physical condition however, an experiment had to be interrupted in order to change either the mass or the spring, which required the participant to deliberately decide what to manipulate, but both changes are coded as CV manipulations. As a consequence, we not only found that there was no difference in CV manipulations between conditions, but also that these manipulations did not have predictive value for learning outcomes.

This picture changed when accounting for the *deliberateness* of experimental manipulations. It turned out to that in contrast to CV manipulations, the percentage of deliberate CV manipulations significantly predicted learning outcomes, as well as differed between conditions. The drop from CV to deliberate CV manipulations was significant only for the SIM condition. This is in line with our reasoning that the user interface for the computer simulation did not make the control of variables a deliberate choice. Even by itself, deliberate manipulations were among the strongest predictor for post-test scores. We suggest that time between manipulations as a measure of deliberateness is not just reflective of the ease of manipulation in a learning environment, but also of the level of cognitive engagement of a participant with an experiment.

Finally, only manipulations targeting the less familiar concept (spring) contributed to conceptual learning, while those targeting the more familiar one (mass) did not seem to impact the learning outcomes, which seems reasonable given that the participants tended to know less about the springs' role in the harmonic oscillation. However, contrary to previous studies [12] that consider confounded manipulations as detrimental to developing conceptual understanding, we found a relatively large though insignificant positive regression coefficient for confounded manipulations on post-test scores. At this point, we can only speculate as to why this is the case; for example, it could be that people with low prior knowledge ran preliminary experiments to get a sense of the physical phenomenon. Further investigation is needed to understand this process.

## 5.2. Differences in Inquiry Behaviors by Learning Environment

We found that conditions did not differ in terms of learning outcomes. In line with previous research that showed equal knowledge gains for virtual and physical manipulative environments [2, 3, 5, 7], we could have argued that there is no difference in benefits of learning environments for developing conceptual understanding in inquiry tasks on mass-spring systems. However, as indicated by the results of the cluster analysis of inquiry behaviors, this would have been the wrong conclusion. The cluster analysis revealed that participants across both conditions could be grouped into two clusters according to how systematic their inquiry behavior was, and that the more systematic cluster had significantly higher learning outcomes than the less systematic cluster. Importantly, almost all of the participants in the physical condition belonged to the more systematic cluster, while most of the participants in the simulation condition fell into the less systematic cluster. This suggests that the learning environments did differ in terms of benefits for developing conceptual understanding. It is important to note that this is not in contradiction to the multiple regression models that show no significant effect for condition. Both analyses show that inquiry strategies had a strong influence on learning outcomes. However, enough participants deviated from their peers in the same condition in terms of inquiry behaviors such that the overall differences in learning outcomes between conditions were canceled. By using more than one variable of inquiry behavior for grouping participants, cluster analysis better accounts for between subject differences in overall inquiry behaviour in each condition. Thus, at least for activities that span a short period of time, we think that measures of experimentation strategies have to be incorporated in studies of the impact of learning environments on learning outcomes in open-ended science inquiry learning.

A possible explanation for these differences in experimental manipulations between conditions is that the ability to employ systematic experimentation strategies is not necessarily a stable domain-general skill but a context-dependent behavior. It is likely that specific affordances of the two learning environments are related to these differences in experimentation strategies, such as the need to pause the experiment to change the spring constant in the real but not virtual environment. While there is consensus on the impact of different affordances of virtual and physical environments on learning outcomes [4], we argue in light of these results that we also need to study the impact of these affordances on the experimentation processes during science inquiry activities. However, as we did not manipulate the specific affordances in the learning environments, we can currently only make educated guesses.

For example, the fact that participants in the SIM condition ran more experiments than in PHY, while spending the same amount of time at the task, supports the claim that it was easier to manipulate variables in the computer simulation than in the physical setup. As argued by Renken and Nunez [12], it might be that systems that enable quick changes with various options prompt participants to get into "play" mode, in which they revert to simple heuristic methods such as trial-and-error and spend less effort on setting up valid experiments. This could explain why proportion of deliberate manipulations was higher for participants using the physical systems.

Another difference in affordances is that in the computer simulation, participants could change the spring constant even as experiments were running, which led to short perturbations in the oscillations that were due to the change, and not necessarily due to the actual spring-mass configurations. Especially in cases "non-deliberate" manipulations that were too short for the perturbations to vanish, participants might have wrongly interpreted these fluctuations.

## 5.3. Limitations and Future Directions

While the study provided evidence that an investigation of inquiry strategies is more informative than merely looking at outcomes, it only offered hints as to what determines the use of those strategies. These appear to be influenced by the different affordances of a learning environment, but studies with longer interaction times, and a greater range and control of environments is needed to understand the characteristics of these relationships in more detail. Future studies should better control and match the virtual and physical environments in order to focus on one or two specific affordances. Studies that manipulate design features *within* a learning environment to assess its impact on inquiry processes are also needed.

Further studies should incorporate the assessment of hypothesis generation and inference processes to examine the impact of affordances of learning environments not just on experimentation strategies, but on these other critical inquiry behaviors as well.

We found that time between manipulations was an important correlate of learning outcomes; however, wit the current study, we can make only educated guesses as to what cognitive processes longer dwell times correspond to. Dwell time could signify the time spent on comparing the current with the prior experiment configuration, on reflecting on existing confusions, on planning the next steps to be taken, or it could just represent the time it takes to perform a manipulation in the learning environment.

Additionally, the lack of difference on learning outcomes between media seems to contradict prior research on virtual versus physical learning environments in comparable inquiry tasks [12]. However, as the tendency of the data goes into the expected direction, we believe that a larger sample size would provide the required power to detect the learning outcome differences.

We currently did not employ automated tracking of participants' behaviors to extract their experiment configurations. However, novel computer vision algorithms, as well as logging systems would address this limitation. Our data organization scheme can be easily integrated with automatized tracking systems.

# 6. CONCLUSION

Drawing on work on scientific reasoning and inquiry, we developed a novel operationalization of systematic experimentation strategies that predict learning outcomes in open-ended inquiry-based learning activities. We further showed that strategy use is context-dependent, in that participants using the physical system went about the inquiry activity differently than participants using the computer simulation.

These findings suggest that we have to broaden the notion of what counts as "systematic experimentation" from mainly consisting of the design of unconfounded experiments and the performance of optimal heuristic search to a more comprehensive views that integrates contextual and cognitive factors (e.g. deliberateness). Data mining algorithms are particularly well suited for exploring such behaviors. However, it is crucial to develop data-mined models of inquiry strategies that are interpretable in order to advance our understanding of learning processes in more complex

inquiry activities. We suggest that any machine-learned model of inquiry behaviors should incorporate semantic representations of what participants' actually explore in inquiry activities, in order to meaningfully extend the data from interaction logs of users engaging in the learning environment.

A further implication of our results is that research on learning environments for science inquiry learning should focus on developing a broader framework that focuses on the affordances as relevant dimensions, irrespective of medium and examines how under what circumstances they benefit learning.

## 7. ACKNOWLEDGEMENT

We would like to thank Prof. Carl Wieman and Eric Kuo, PhD, for their guidance and strong support in this research, as well as members of the AAALab at the Stanford University for their insightful feedback.

## 8. REFERENCES

[1] van Joolingen, W., & Zacharia, Z. (2009). Developments in inquiry learning. In *Technology-enhanced learning.* Netherlands: Springer.

[2] Triona, L., & 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*, 149-173.

[3] Zacharia, Z., & Olympiou, G. (2011). Physical versus virtual manipulative experimentation in physics learning. *Learning and Instruction , 21*, 317-331.

[4] de Jong, T., Linn, M., & Zacharia, Z. (2013). Physical and virtual laboratories in science and engineering education. *Science, 340*, 305-308.

[5] Klahr, D., Triona, L., & 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*, 183-203.

[6] Zacharia, Z., & Constantinou, C. (2008). Comparing the influence of physical and virtual manipulatives in the context of the physics by inquiry curriculum: The case of undergraduate students' conceptual understanding of heat and temperature. *American Journal of Physics, 76*, 425-430.

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

[8] Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review, 20*, 99-149.

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

[10] Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*, 1098-1120.

[11] Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes.* Cambridge: MIT Press.

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

[13] Shih, B., Koedinger, K., & Scheines, R. (2010). Unsupervised Discovery of Student Strategies. *Proceedings of the 3rd Intl. Conf. on Educational Data Mining*, (pp. 201-210).

[14] Kardan, S., & Conati, C. (2011). A Framework for Capturing Distinguishing User Interaction Behaviours in Novel Interfaces. *Proceedings of the 4th Intl. Conf. on Educational Data Mining*, (pp. 159-168). Eindhoven, the Netherlands.

[15] Kardan, S., Roll, I., & Conati, C. (2014). The usefulness of log based clustering in a complex simulation environment. *Intelligent Tutoring Systems* (pp. 168-177). Springer International Publishing.

[16] Sao Pedro, M., Baker, R., Gobert, J., Montalvo, O., & Nakama, A. (2013). Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. *User Modeling and User-Adapted Interaction, 23* (1), 1-39.

[17] Sao Pedro, M., Baker, R., & Gobert, J. (2012). Improving construct validity yields better models of systematic inquiry, even with less information. *User Modeling, Adaptation, and Personalization* (pp. 249-260). Berlin Heidelberg: Springer.

[18] Palmer, D. (1995). *The POE in the primary school: An evaluation. Research in Science Education*, 25 (3), 323-332.

[19] Penner, D., & Klahr, D. (1996). The interaction of domain-specific knowledge and domain-general discovery strategies: A study with sinking objects. *Child Development, 67*, 2709-2727.

[20] Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology , 32*, 102-119.

[21] Garcia-Mila, M., & Andersen, C. (2007). Developmental change in notetaking during scientific inquiry. *International Journal of Science Education, 29* (8), 1035-1058.

[22] Perkins, K., Adams, W., Dubson, M., Finkelstein, N., Reid, S., Wieman, C., & LeMaster, R. (2006). PhET: Interactive simulations for teaching and learning physics. The Physics Teacher, *44*(1), 18-23.

[23] Reynolds, A., Richards, G., de la Iglesia, B., & Rayward-Smith, V. (1992, 5). Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, 475-504.

[24] 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.