Measuring systematicity of students' experimentation in an open-ended simulation environment from logging data

Gey-Hong Gweon

Physics Front LLC, Sam@physicsfront.com

Hee-Sun Lee

The Concord Consortium, hlee@concord.org

William Finzer

The Concord Consortium, wfinzer@concord.org

Acknowledgments
This work is in part supported by the National Science Foundation under grants IIS-1147621 and DRL-1435470. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Reference Citation
Gweon, G. -H., Lee, H. -S., & Finzer, W. (2016). Measuring systematcity of students' experimentation in an open-ended simulation environment from logging data. Paper submitted for the Annual Meeting of the American Educational Research Association. Washington, D.C.

Abstract

In this paper, we propose a metric based on the sample entropy concept for measuring the systematicity of students' experimentation patterns in an open-ended simulation environment where a number of parameters are at students' disposal to explore. Unlike other indicators of systematicity proposed in the literature, the sample entropy metric provides a continuous scale and draws upon the up-to-date computational algorithm applied to dynamic processes involved in physical and biological systems. This sample entropy-based metric correlates significantly with student learning outcomes related to (1) how well students described the nature of relationship explored during their experimentation and (2) whether students coordinated between claim and data collected from their experimentation. Our analysis indicates that (1) the sample entropy metric captures the aspect of students' experimentations that is not captured by several conventional measured and (2) it has potential for general application to a variety of simulation-based activities when assessing students.

Keywords: Sample entropy, systematicity, experimentation, simulation, analytics

Introduction

Inquiry-based science learning has been explicitly emphasized in many recent science education reform documents in the last twenty years. While the definition for inquiry-based learning has been continually updated through several revisions, experimentation along with explanation, argumentation, and modeling, has shown prominence throughout (NRC, 1996, 2000, 2007, & 2012). Simulations equipped with user controls provide an excellent opportunity for students to experiment with a scientific system in order to gain understanding of its multi-variate relationships (Honey & Hilton, 2011). In a science learning environment where simulations are used for students to carry out experimentation, systematic investigation is considered important to measure (Kanari & Millar, 2004). Though there is no common definition for what it means to be systematic in experimentation and how to measure it, the control of variables strategy is often regarded as an experimentation skill that can indicate students' systematic investigation (Gobert et al., 2013).

In the literature on performance based assessments (Ruiz-Primo & Shavelson, 1996) or inquiry assessments (Gobert et al., 2012), the control of variables strategy (Kuhn & Dean, 2004) is determined as the presence or the absence of evidence that students changed only one variable at a time in a sustained manner, giving a binary scale, rather than a continuous scale. Such a binary scale is helpful but it is a crude measure to take into account variations in students' changes in variables. For example, is the systematicity of experimentation conducted by student A who varies one variable five times in a row is different from that of experimentation by student B who varies one variable four times in a row followed by another variable for once? If the experimentation systematicity is defined on a continuous scale, then the confusion arising from where to draw a line between unsystematic and systematic investigations can be cleared and thus would make it possible to examine the correlation of the degree of the experimentation systematicity to other variables captured during and after learning with simulations such as students' explanations and test scores. Moreover, a continuous scale may make it possible to investigate the following possibility: too systematic a behavior in controlling variables may be reflective of a rote learning rather than a true and creative inquiry. A question is how can the systematic degree of experimentation be defined in such a way that captures variations in students' variable changes in open-ended investigations?

The objective of this study is to (1) develop a metric of the systematic degree of a student's experimentation, (2) examine how the metric is correlated with student learning outcome variables, and (3) compare the metric with other variables used in the literature to describe students' experimentation patterns such as the number of variables students changed, the total number of trials students carried out, the average number of trials per variable, and the average range of values explored per variable during experimentation.

Review of Literature

Experimentation can occur with physical apparatus or simulations. With physical experimentation, research has focused on whether, how, or to what extent students can design and conduct experiments (Hackling & Garnett, 1992; Kanari & Millar, 2004). Students' experimentation skills include recognizing multi-covariate relationships (Amsel & Brock, 1996; Kuhn & Dean, 2004), dealing with experimental errors (Allchin, 2012), addressing variability in the data (Masnick, Klahr, & Morris, 2007; Petrisino, Lehrer, & Schauble, 2003), applying statistical reasoning (Lubben et al., 2001), treating anomalous data (Chinn & Brewer, 1993), and revising hypotheses, experiments, and questions after reflecting on evidence (Schauble, 1996). Studies have found that students have difficulties in recognizing, identifying, and controlling variables (Toth et al., 2000).

Masnick and Klahr (2003) identified five phases of experimentation with physical apparatus such as design, physical setup, execution, outcome measurement, and analysis. With simulation-based experimentation, the three middle phases of physical setup, execution, and outcome measurement can be done rather easily without errors by students. Therefore, simulation-based experimentation highlights the design and analysis phases of experimentation. When an experiment is designed to find relationships between all salient variables of a scientific system and an outcome variable, students' decisions and subsequent executions related to how to change values on which variables in what sequence becomes most prominent. As a result, students' patterns of variable changes can effectively summarize students' experimentation patterns.

When studying students who were engaged in simulation-based experiments, McElhaney and Linn (2011) found that students' experimentation patterns could be characterized as intentional, unsystematic, and exhaustive based on the number of trials attempted by students, trial variability defined by the range of each of the tried variables, and experimentation validity hand-scored to represent the extent to which control of variable was followed and whether the variable change matched a question selected for an investigation. Even though McElhaney and Linn (2011) did not develop a measure for systematicity of experimentation, they used three student experimentation examples to illustrate the difference in ways that the variables were changed. Gobert et al. (2013) developed a detector that assesses the designing controlled experiments skill using log data recorded from students' simulation-based investigations in two steps. First, human coders hand-scored presence of the controlled experimentation skill from log data where human coders examined the tag assigned to each clip with a certain description, and then examined multiple clips to determine presence of the control of variable strategy. Second, data mining was conducted to model a set of features from students' interactions with simulations could predict presence or absence of the controlled experimentation skill. In these two studies, students' systematic experimentation was nonetheless a binary measure and always relied on human judgments at least initially before automatization. Furthermore, these indicators were validated based on human judgments but failed to show that they were independently and significantly correlated with student learning as hypothesized in the inquiry-based learning literature. In this study, we conceptualize the degree of systematicity in student experimentation on a continuous scale that does not rely on human judgments. Nor does it require considerations of students' interactions with simulations other than students' value changes in variables, making the new systematicity indicator unequivocally definable across all types of simulation-based experimentations.

Methods

Learning Context

We developed a computer-supported learning environment called InquirySpace (IS) where typical high school students in diverse science classrooms can undertake scientific inquiry of their own designs. To illustrate, we use an activity where students explore a simple computational model of how the Earth's average global temperature is influenced by several factors influencing climate change. In this experiment, four parameters can be controlled:

- CO_2 Level: button clicks for adding and removing CO_2 in the atmosphere ranging from 0 to 1000 ppm.
- Sun Brightness: a sliding bar with one percent increment ranging from 75 percent to 125 percent

- Albedo: a sliding bar with .01 increment ranging from 0.00 to 1.00.
- Number of Clouds: button clicks for adding and removing clouds ranging from 0 to 25.

The simulation generates global temperatures that are continuously plotted over time on a time-series graph. The simulation speed can be controlled on the sliding bar at the top of the simulation. After one simulation run is finished, students can export the values of the four parameters and the final global temperature. Figure 1 shows a screenshot of the climate change simulation combined with the data analysis involving automatically-generated tables and student-generated graphs. In Figure 1, two graphs are generated by students. The time-series graph on the left shows how temperatures change over time when albedo values are changed five times. The graph on the right shows the relationship between albedo values students chose and the temperature change between beginning and ending of a simulation run.

Global atmospheric temperatures result from the energy balance between Sun's radiation coming to Earth and Earth's radiation emitted to space. An increase in Sun's brightness, thus Sun's radiation energy, makes global temperatures rise. Albedo is the average fraction of incident radiation that is reradiated without being absorbed. Therefore, an increase in albedo decreases the amount of Sun's radiation energy absorbed by Earth, leading to a decrease in global temperatures. Since CO_2 is a greenhouse gas that absorbs the Earth's emitted infrared radiation, an increase in the CO_2 level in the atmosphere increases global temperatures. On the other hand, the effect of clouds is more complex. Since clouds reflect the incoming Sun's radiation energy back to space, an increase in clouds can decrease overall global temperatures. However, clouds can also trap infrared energy radiated from the Earth's surface, leading to warming.

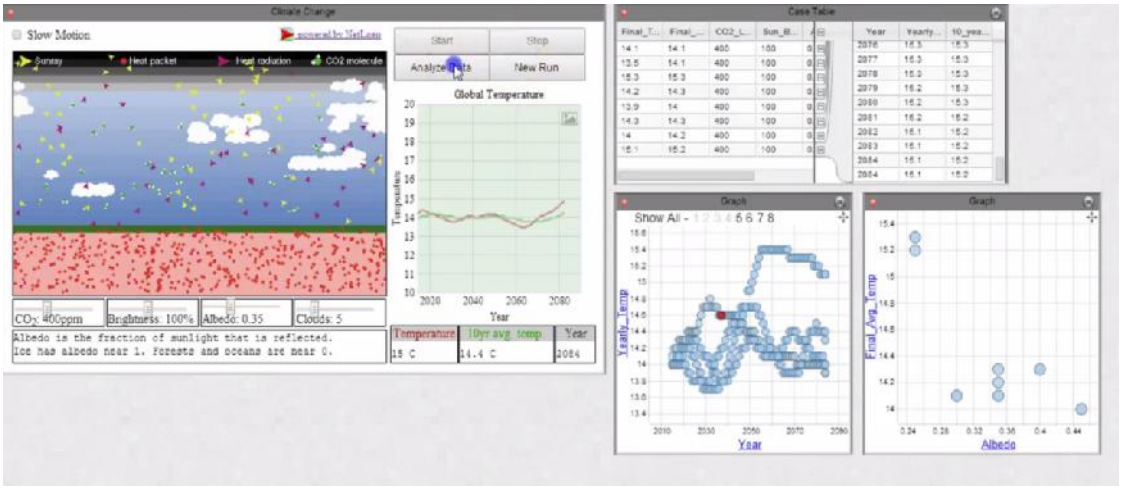

Figure 1. Students can manipulate CO_2 amount, sun brightness, albedo, and cloud amount in the climate model shown above. Students can observe global temperature changes and export data sets to draw graphs. In this figure, students drew two graphs: one between time and global temperature is shown for five albedo values, and the other between albedo and average terminal global temperatures.

Data Collection

The simulation-based climate investigation took place as performance assessment of students' experimentation abilities over one class period for high school classrooms taught by

three physics teachers in three high schools. 221 students worked in 100 small groups of one to three students. Each group chose its own question. Among the students, 92% spoke English as a first language, 52% were female; 52% self-reported to have used computers regularly for school learning. Prior to the climate model experimentation, students were asked to select one out of the four climate variables to investigate how the selected climate factor would influence global temperatures 50 years later. Students then were asked to carry out simulation runs until satisfied. Then, students were asked to make a claim about how the selected climate factor affected global temperatures in the future and explain their claim based on evidence from the simulation runs. Both claim and explanation prompts elicited open-ended responses. All of students' transactions made with the climate model were logged. The log file included 14,259 lines of text and 3.6 MB in size. Among the 100 student groups, 89 groups wrote claims and explanations while 94 groups' log data were available; 83 groups had both claims/explanations and log data.

Data Analysis

Parameter value space vs. parameter change space.

The experimentation systematicity measure is drawn from the concept of entropy, a measure of disorderliness originating from physics and used in many science disciplines such as computer science, chemistry, and biology. The experimentation systematicy measure is defined on a hyper-dimensional space consisting of the number of variables students are allowed to vary. We define each variable in the climate model students can vary as a parameter. Since students can vary four variables, the climate system depicted in the model consists of four parameters, creating a four dimensional parameter value space. Each dimension of the parameter value space represents the values allowed to explore by students. Each simulation run thus can map onto this **Parameter Value Space** using four coordinates, i.e., (CO_2 Level, Sun Brightness, Albedo, Cloud Amount). For our work, however, the **Parameter Change Space**, in which we record whether a parameter has been changed or not, turns out to be more important. Lastly, for the visualization purpose, the **Cumulative Change Space**, whose coordinate value represents the sum of the corresponding coordinate value of the Parameter Change Space, is convenient. These spaces are illustrated in Figure 2 and are explained in more detail in the caption .

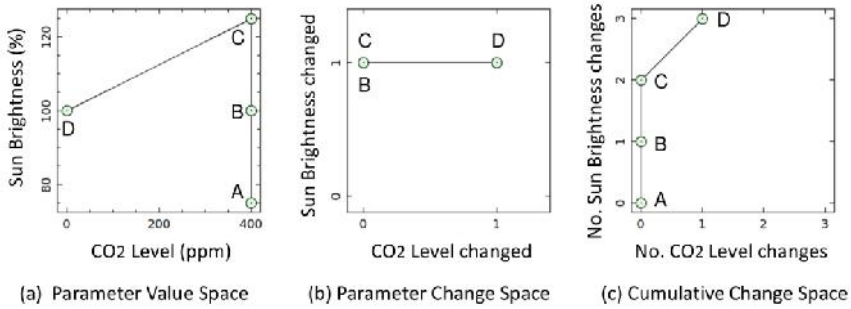

Figure 2. **Parameter Value Space, Parameter Change Space, and Cumulative Change Space.** Here we show an example where a student group runs the model with the following subsequent values of parameters for (CO_2 Level, Sun Brightness): (400,75), (400, 100), (400, 125), (0, 100). These values are plotted in (a) **Parameter Value Space**, and marked with simulation run labels A, B, C, D. The other two parameter values (Albedo, Cloud Amount) are held fixed in this example, and thus are not be shown. In (b) **Parameter Change Space**, we

indicate only whether parameter changed (1) or not (0) from the previous run. At run B, the CO_2 level did not change, while the Sun Brightness changed, and so B is represented as (0, 1). And, so is C. At run D, both parameters changed, and so D is represented as (1, 1). In (c) **Cumulative Change Space**, the coordinate values (0 or 1) in (b) are summed over up to the current run, starting from the origin, which by definition corresponds to the first run A. Thus, the coordinate value represents the total number of changes up to that run. Note that for analyzing the actual data, these spaces must be taken as **four-dimensional**, not two-dimensional (as chosen for convenient visualization here), corresponding to four parameters.

Experimentation systematicity metric based on the sample entropy in the Parameter Change Space

In order to develop a metric for experimentation systematicity, we investigated various entropy concepts associated with the thermodynamic entropy used in the domain of physics. Since students' value changes were made in succession, we used several entropy conceptualizations defined for time-dependent (generally non-linear) processes. The most promising is the concept of "sample entropy" (Richman & Moorman, 2000), which approximates the Kolmogorov-Sinai entropy (Eckmann & Ruelle, 1985) and has been applied to quantify the disorderly nature of a time-dependent physical process, for example to estimate the degree of irregularity in patients' heart beats.

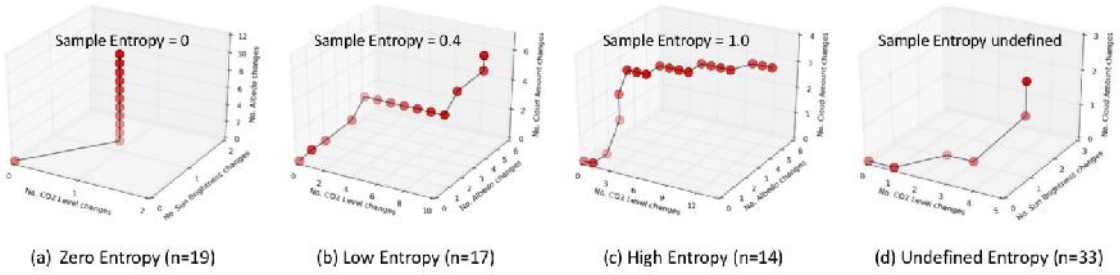

Figure 3. Some examples of students' experimentation sequences illustrate the sample entropy. Note that the sample entropy values are calculated from the four dimensional data, while these plots are three-dimensional. For an obvious reason, in each plot we omit one axis, the axis with the least number of changes. The number of changes for the omitted axis is zero for (a) and (d), but is non-zero, while small, for (b, c): 2 for (b) and 1 for (c).

In the literature, the sample entropy has been defined for a stream of numbers, i.e., for a dynamical process in one dimension. While it can be generalized to a dynamical process in any dimensions, such as four dimensions in the case of our Parameter Value Space, we found that the small size of our sample gives results with poor statistics if we do so. The sample entropy can be defined meaningfully in the Parameter Change Space also, and it, thus defined as we discuss now, was found to be statistically meaningful for our sample. As the coordinate value in the Parameter Change Space is 0 or 1, the coordinate values in this space, such as (0, 1, 0, 0) or (1, 1, 0, 0) can be considered as four-bit binary numbers, 0100 (=4) or 1100 (=12). The following simplest possible form of the sample entropy was found to be the most effective one as well. Let us consider a stream of four bit integers: $u_1, u_2, u_3, \dots, u_N$. Let E_i, where $i = 1, \dots, N-1$, be defined as the number of other integers in the stream that are equal to u_i. Then, let us consider a derived sequence, composed of adjacent pairs of integers:

$(u_1, u_2), (u_2, u_3), \ldots, (u_{N-1}, u_N)$. Let E_i' be defined as the number of other pairs that are equal to (u_i, u_{i+1}). Then the sample entropy is defined as

$$\text{Sample Entropy} = \log\left(\frac{\sum_i E_i}{\sum_i E_i'}\right),$$

where the symbol log represents the natural logarithm and both summations run from $i = 1$ through $N - 1$. Note that (1) $\sum_i E_i \geq \sum_i E_i' \geq 0$, and (2) if $\sum_i E_i'$=0 then the sample entropy is undefined.

Other metrics related to student experimentation

In order to examine how our entropy-based students' experimentation systematicity indicator compares with other indicators of student experimentation, we calculated four additional indicators from the log data for each student group's experimentation as follows:

- the total number of simulation trials run by students
- the number of climate factors students varied by students
- the number of simulation trials per climate factor
- the relative value range explored by students as compared to the range allowed by the simulation model per climate factor (e.g. if students changed albedo from 0.2 to 0.8 in four trials, then it was calculated as 0.6/1.0 = 0.6; if students covered the value range of 0.6 on albedo and 0.4 on Sun's brightness, then it was calculated as 0.5)

Scoring learning outcomes: nature of relationship declared in claim and evidence-based claim

Students' open-ended claims about the relationship between the climate factor they chose and future global temperatures were scored in four levels as follows:

- Score 0: Off-task or irrelevant responses
- Score 1: The declared relationship was scientifically incorrect (e.g. the higher the albedo, the higher the future global temperature)
- Score 2: The declared relationship was not specific but acknowledged the presence of association (e.g., Sun's brightness has an effect on future global temperatures)
- Score 3: The direction of the declared relationship was correctly stated (e.g., the lower the albedo, the higher the future temperatures)

We also scored for the coordination between data and claim by looking at both claim and evidence cited in their explanation. We assigned "1" when students' data in their explanations supported their claims and "0" when did not. We used these two student learning outcome variables as dependent variables to investigate how the five indicators of students' experimentation were correlated with them. We hypothesized that quality indicators of students' experimentation should be significantly correlated with the two student learning outcome variables.

Results

Five indicators for student experimentation were calculated, their means and standard deviations are listed in Table 1 and their distributions are shown in Figure 4 as histograms. Overall, students made an average of 16.61 simulation trials by varying an average number of 2.45 variables. Students changed values an average of 6.21 times per climate factor while an

average of 81% of the value range was covered by students per climate factor. Note that the distribution of the total number of simulation trials is not similar to that of the number of trials per climate factor. The total number of simulation trials could not differentiate cases where students explored one climate factor thoroughly from those whether students haphazardly explored multiple climate factors. This was not the case for the number of trials per climate factor. In addition, more than 47% of the student groups explored the full allowable range per climate factor as shown in Figure 4(e). Among the 83 student groups we analyzed, we were able to calculate sample entropy values for 50 student groups. Student groups with missing sample entropy values occurred mainly because (1) they conducted less than 3 runs—too small a number to enable the calculation of entropy and (2) their parameter space pathways were too chaotic to lead a convergent value as exemplified in Figure 3(d). Among the 50 groups for which sample entropy values were calculated, 19 student groups had a zero sample entropy value, indicating that, basically, they only changed one variable during their entire experimentation. See Figure 4(a).

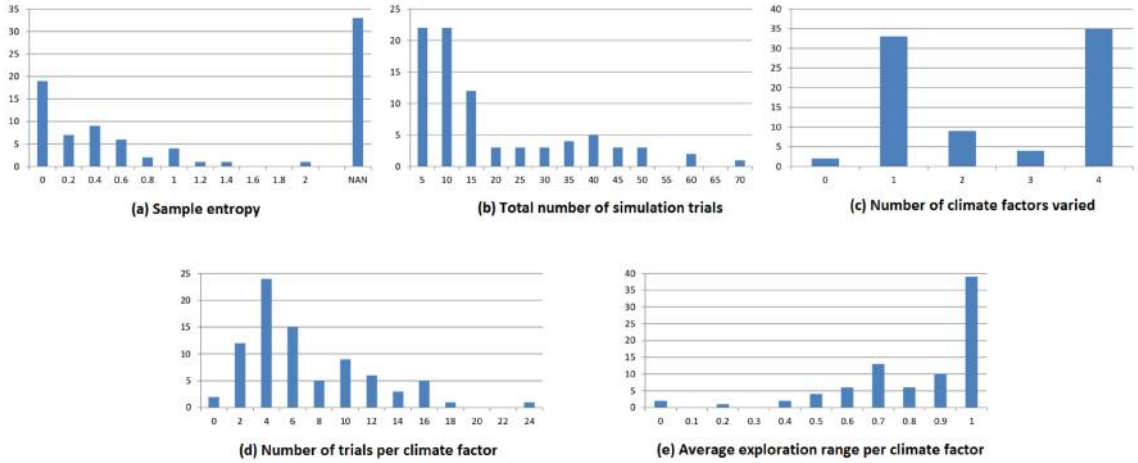

Figure 4. Distributions of five indicators related to student experimentation

Table 1. Descriptive statistics on four indicators related to student experimentation

Experimentation indicators	Sample entropy	Total number of simulation trials	Number of climate factors varied	Number of trials per climate factor	Exploration range per climate factor
No. of trials	50	83	83	83	83
Mean	0.31	16.61	2.45	6.21	0.81
Standard Deviation	0.40	15.62	1.42	4.52	0.24
Correlation with nature of the relationship in the claim	**-0.35***	0.15	-0.056	**0.27***	**0.27***
Correlation with coordination between evidence and claim	**-0.38****	0.27	0.04	**0.39*****	**0.27***

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

According to Table 1, sample entropy was significantly negatively correlated with the nature of relationship declared in claim (r = - 0.35, p < .05) and with coordination between evidence and claim (r = - 0.38, p < .01). The significant negative relationships were predicted because the higher the students' sample entropy, the more chaotic the variable change during the students' entire experimentation. On the other hand, the number of climate factors varied and the total number of simulation trials were significantly correlated neither with the claim nor with the coordination between evidence and claim, which agrees with McElhaney and Linn (2011). The number of trials per climate factor and the exploration range per climate factor variables were significantly positively correlated with the claim with the lesser degrees than the sample entropy. The number of trials per climate factor variable was significantly positively correlated with the evidence-claim coordination with the similar magnitude. So was the exploration range per climate factor variable but with a lesser degree. Since sample entropy, number of trials per climate factor, and exploration range per climate factor variables showed similar correlation patterns with the two learning outcome variables, we investigated correlations among these three variables to see whether sample entropy was different from the other two. According to Table 2, sample entropy was neither significantly nor highly correlated with the two variables, indicating that sample entropy is capturing different aspects of students' experimentation patterns from the two.

Table 2. Correlations among indicators of student experimentation

Experimentation indicators	Sample entropy	Total number of simulation trials	Number of climate factors varied	Number of trials per climate factor	Exploration range per climate factor
Sample entropy	-	0.19	0.54***	0.16	0.02
Total number of simulation trials	-	-	0.45***	0.83***	0.30**
Number of climate factors varied	-	-	-	0.19	0.07
Number of trials per climate factor	-	-	-	-	0.44***
Exploration range per climate factor	-	-	-	-	-

Note. * p < .05, ** p < .01, *** p < .001

In order to examine the correlation between sample entropy and the two learning outcomes, we divided students into four groups with undefined entropy (n = 33), zero entropy, low entropy (less than 0.5 sample entropy values), and high entropy (more than 0.5 sample entropy values). The cutoff entropy value of 0.5 to determine high vs. low entropy groups was taken from the mean value for students groups with non-zero definable sample entropy values. Figure 5 shows that all student groups whose entropy values ranged from zero to 0.5 identified scientifically correct and detailed relationships between the climate factor they chose and the

global temperatures. They also coordinated their claim with evidence very well. The performance levels fell for students in the high entropy group and even further for those in the undefined entropy group. The mean differences among the four groups were significant based on the one-way ANOVA result, $F(3,79) = 3.97$, $p < .05$, for the nature of relationship in claim and $F(3,79) = 8.89$, $p < .001$ for the claim-evidence coordination.

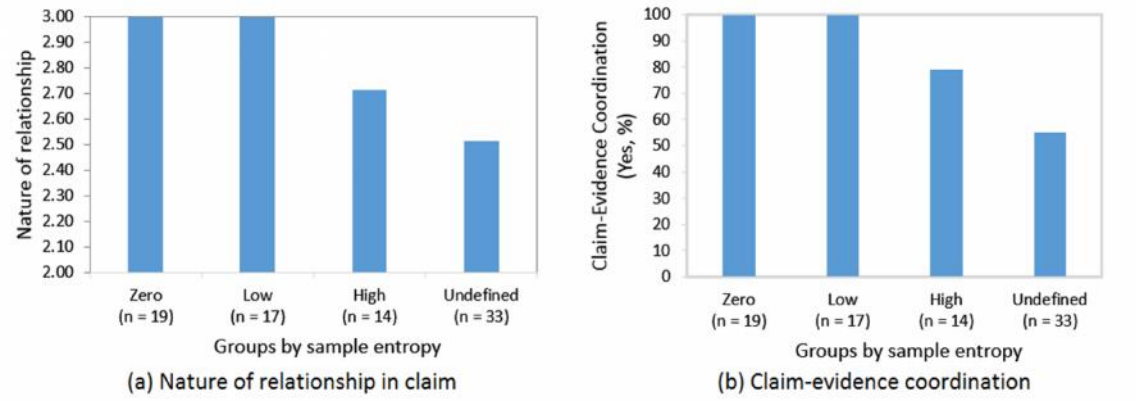

Figure 5. Student learning outcomes across sample entropy groups

Conclusions and Implications

The educational research community had attempted to extract information related to instructional dynamic at the microgenetic level in order to describe the process of student learning during an intervention. In this study, we developed an indicator that can represent the systematicity of students' experimentation based on the sample entropy concept defined in physical and biological sciences. Our results indicate that sample entropy delivers a new piece of information that is not readily captured by the number of simulation trials per changed factor or the explored range per changed factor. We propose that the sample entropy is a practical way to measure the disorderly nature of a time dependent process of our current interest, i.e., a sequence of events where a student changes the states of multiple variables. Then, the negative of the sample entropy corresponds to the metric for the systematic degree of student experimentation. The sample entropy on a continuous scale is significantly correlated with the student learning outcomes related to experimentation, in agreement with the general idea regarding the systematicity in the literature. The sample entropy algorithm could potentially be useful to automatically diagnose students' experimentation patterns in real time to provide just-in-time scaffolding when needed. For example, for students whose experimentation patterns show high entropy or non-calculable entropy after sufficient number of trials, feedback can be provided to guide them to focus their trials around a single variable. Our current findings, while promising, are limited because they were based on one simulation with a small number of student groups. With a much larger sample size, our future plans include expressing the sample entropy on the parameter value space as well as testing the sample entropy on a variety of simulations to understand the systematicity metric based on the sample entropy in multiple learning situations.

References

Allchin, D. (2012). Teaching the nature of science through scientific errors. *Science Education, 96*(5), 904-926.

Amsel, E., & Brock, S. (1996). The development of evidence evaluation skills. *Cognitive Development, 11*(523-550).

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research, 63*(1), 1-49.

Eckmann, J. P. & Ruelle, D. (1985). Ergodic theory of chaos and strange attractors. *Reviews of Modern Physics, 57*(3), 617–656.

Gobert, J., Sao Pedro, M., Baker, R., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining, 4*(1), 111-143.

Gobert, J., Sao Pedro, M., Raziuddin, J., & Baker, R. (2013). From log files to assessment metrics for science inquiry using educational data mining. *Journal of the Learning Sciences, 22*(4), 521-563.

Hackling, M. W., & Garnett, P. J. (1992). Expert-novice differences in science investigation skills. *Research in Science Education, 22*, 170-177.

Honey, M. A., & Hilton, M. (2011). *Learning science through computer games and simulations*. Washington D.C.: The National Academies Press.

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.

Kuhn, D., & Dean, D. J. (2004). Connecting scientific reasoning and causal inference. *Journal of Cognition and Development, 5*, 261-288.

Masnick, A. M., Klahr, D., & Morris, B. J. (2007). Separating signal from noise: Children's understanding of error and variability in experimental outcomes. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 3-26). New York: Lawrence Erlbaum Associates.

Lubben, F., Campbell, B., Buffler, A., & Allie, S. (2001). Point and set reasoning in practical science measurement by entering university freshman. *Science Education, 85*(4), 311-327.

McElhaney, K. W., & Linn, M. C. (2011). Investigations of a complex, realistic task: Intentional, unsystematic, and exhaustive experimenters. *Journal of Research in Science Teaching, 48*(7), 745-770.

National Research Council (1996). *National science education standards*. Washington, DC: National Academy Press.

National Research Council (2000). *Inquiry and the National Science Education Standards*. Washington, DC: National Academy Press.

National Research Council (2007). *Taking science to school*. Washington, DC: National Academic Press.

National Research Council (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.

Petrisino, A., Lehrer, R., & Schauble, L. (2003). Structuring error and experimental variation as distribution in the fourth grade. *Mathematical Thinking and Learning, 5*, 131-156.

Richman, J. S., & Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology - Heart and Circulatory Physiology, 278*(6).

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: An update. *Journal of Research in Science Teaching, 33*(10), 1045-1063.

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

Toth, E. E., Klahr, D., & Chen, Z. (2000). Bridging research and practice: A cognitively based classroom intervention for teaching experimentation skills to elementary school children. *Cognition and Instruction, 18*(4), 423-459