An Assessment Framework for Complex Systems Understanding

Hala Khodr, Aditi Kothiyal, Barbara Bruno, Pierre Dillenbourg
Hala.khodr@epfl.ch, aditi.kothiyal@epfl.ch, barbara.bruno@epfl.ch, pierre.dillenbourg@epfl.ch
Ecole Polytechnique Fédérale de Lausanne

Abstract: Complex systems understanding is crucial for understanding the sciences of the twenty-first century. Helping students with the principles of complex systems is thus both a challenge and an opportunity. In order to investigate the effectiveness of complex systems learning activities, we would like to have an assessment instrument which we can ultimately use as pre and post-test to assess the learning gain. In this paper, we developed an assessment instrument as a questionnaire containing five scenarios related to key concepts of complex systems understanding. After validating the understandability of the questions, we collected answers from 37 participants including experts as well as non-experts. We finally propose a scoring scheme based on fuzzy logic. This approach results in a general procedure which is compact, automatic and human legible to quantify expertise in complex systems concepts.

Introduction

Order, regularities, and patterns are ubiquitous around us. A flock of birds winging in the sky, the self-organization of social insects, or the synchronized blinking of fireflies are examples of natural complex systems where "macro" patterns arise from the interaction of "micro" agents. Complex systems can be defined as "an interdisciplinary field of research that seeks to explain how large numbers of relatively simple entities organize themselves, without the benefit of any central controller, into a collective whole that creates patterns, uses information, and, in some cases, evolves and learns" (Mitchell, 2009, p. 4).

Not only is this an important domain to learn about on its own, but it actually consists of a "powerful idea" (Papert, 1980) that cuts across disciplines and can lead to the understanding of a large class of physical and social phenomena taught as conceptually different subjects. Researchers argue that different complex systems in nature, such as insect colonies, brain activity, immune systems, economies, the atmosphere, and human social interactions, have much in common. In fact, understanding complex systems is fundamental to understanding science, and it is becoming one of the essential cognitive skills needed for the twenty-first century (Reed, 2020). It is therefore not surprising that the Nobel Prize in physics in 2021 was awarded for groundbreaking contributions to our understanding of complex physical systems.

An active research area within the teaching and learning of complex systems is that of developing computational tools and resources for facilitating the learning of complex systems concepts. These are mostly based on agent-based modeling environments such as StartLogo or NetLogo that help students construct a connection between the interactions of micro-level particles/agents and the emergent complex macro behaviors. These learning environments are based on the assumption that whenever students have the opportunity to take the point of view of an atom, an electron, a sheep, a fish, an ant, or a trader, they are more ready to conceive complex systems (Wilensky & Rand, 2015). Repenning et al. (2010) developed "Collective Simulations" that help students to learn about the intricacies of interdependent complex systems by engaging them through social learning techniques in the classroom along with networked computers. Similarly, participatory simulations, such as Hubnet (Wilensky et al., 2006), were developed as a type of role-playing activity used in classrooms to explore how complex dynamic systems evolve over time through individual elements acting. In order to evaluate the effectiveness of these resources, it is necessary to assess the learners' complex systems knowledge, which is challenging given its cross-disciplinary nature and vastness of its underlying concepts. Therefore, an important area of research consists of developing instruments to assess complex systems understanding.

In this paper, we focus on this second research goal. We develop a general assessment instrument, freely publicly available, assessing the five essential complex systems ontological concepts, discuss its validation and provide an automatic, human legible general scheme to classify expertise in complex systems concepts.

Theoretical Background

Complex systems are present everywhere in nature, and many are commonly observed or experienced by all of us during our daily life. However, experts and novices think and build their knowledge about complexity in significantly different ways. Hmelo-Silver (2007) found that novices focus on the visible and static structures of the subsystems involved while experts incorporate structures, causal behaviors, and functions. Similarly, Jacobson (2001) documented that the differences between experts and novices lie in the cognitive and ontological ways of thinking. Novices were committed to believe that systems work in a centralized, predictable and reductive ways, while experts tend more to describe systems as decentralized, nonlinear, and nonreductive.

Learning about complex systems is challenging (Hmelo-Silver, 2006), and many studies investigate the difficulty that humans have in understanding and grasping complex systems, alongside common misconceptions about them. Resnick (1994) discusses students' bias to generally assume that the behavior of a system is controlled by lead - a leader orchestrating the system's patterns, or by seed - a single pre-existing entity in the environment, rather than resulting from decentralized interactions. This inclination toward centralization is referred to as Deterministic-Centralized Mindset. Moreover, probabilistic non-deterministic behavior in systems is also found to be a challenging concept to comprehend (Wilkerson-Jerde, 2015). Similarly, people tend to assume immediate effects instead of indirect cascading effects (Grotzer et al., 2015). Chi et al. (2005) have also shown that a process is generally thought of as a direct static phenomenon rather than emergent. An assessment instrument for complex systems understanding should therefore be designed to detect the presence of these misconceptions.
Yoon et al. conducted a study where they analyzed students' responses to one open-ended question in an ecology context concerning the effects of the arrival of geese on a park ecosystem. Results showed that the most difficult ideas to grasp are those related to the decentralized organization of the system and the unpredictable or non-deterministic nature of effects. They then proposed a learning progression to systematize the pathways students undertake to enhance their conceptual competence in complex systems. In another context, Wilensky and Abrahamson (2006) studied the student learning of the spread of diseases in a participatory simulation activity. Their findings identified several reasons for students' incorrect interpretation when explaining the agent-to-aggregate (i.e. micro to macro) relationship. These include proportional and linear reasoning, randomness-determinism confusions, disregarding feedback loop effects, and anticipating emergence from agents' rule-based interactions. Their assessment was composed of questionnaires and interviews where the situations, modeling, and representation tools relate thematically to the disease spread context. The students were provided with a modeling kit (stationery, Lego, strings, etc.) and data-charting tools to give them a choice between either modeling the process itself with concluding their response based on this, or directly charting quantitative data without thinking through the process. The key items in the questions were the spread of a mold on a moldy piece of bread, the scattering of people in a gym class, the spread of a rumor in a class, and the disease spread.

Hmelo-Silver et al. (2007) suggested a framework based on structure-behavior-function theory to represent how people think about complex systems, in particular, in a biological context. Structures refer to the elements, functions to the role of an element, behaviors to mechanisms of how the structures achieve their function. The case study included two biological complex systems: the human respiratory system and an aquarium ecosystem. To assess complex systems understanding, interviews were conducted, including drawing and open-ended questions. These were then coded by the presence and absence of the target concepts (structures, behaviors, functions). Jacobson (2001) developed another framework depicting multiple categories of complex systems concepts based on ontological and epistemological "component beliefs". Multiple studies then followed and adapted this framework (Jacobson et al., 2011; Yoon, 2011, 2019). Jacobson et al. (2011) refined this framework to focus on five main concepts that we also rely on in this work (more details later). They developed a hypermedia learning environment consisting of five NetLogo agent-based models: foraging of ants, traffic jams, self-organization of slime molds, social segregation, and wolf-sheep predation. Interactive scaffolding was introduced related to declarative knowledge of the complex ontologies as well as problem-solving exercises to compare and contrast the different aspects. Ontological shifts were observed, which further confirms and validates earlier results concerning the main concepts comprising complex systems.

In the literature discussed above, many studies explored the importance of as well as the difficulties encountered when learning about complex systems, and defined and validated its main concepts. However, there is no instrument available to objectively score a learner's understanding across different concepts of complex systems, in addition to different contexts. The contribution of our paper is motivated by this limitation. Based on and inspired by the findings above, in this paper, we compose a general assessment instrument that can be used to assess expertise within complex systems. The instrument requires participants to answer questions related to five different scenarios which are then scored to extract their competence in five ontological concepts required for complex systems understanding. More specifically, our research question is: can we quantify the expertise in complex systems based on answering questions targeting related ontological concepts?

Method

Materials

The instrument is designed as a questionnaire consisting of five scenarios related to key concepts of complex systems understanding. These concepts, identified as core complex systems characteristics, used and validated in prior research (Jacobson, 2001, Jacobson et al. 2011, Yoon et al, 2019) are the five ontological categories of complex systems understanding — Order (Centralized, Decentralized), Causes (Single, Multiple), Actions effect (Linear, Nonlinear), Agents effect (Predictable, Random), and Processes (Static, Emergent). The scenarios draw inspiration from literature and are: Traffic Jam (Resnick, 1994), Scattering (Wilensky and Abrahamson, 2006), Flock of Birds (Jacobson, 2001), Butterfly Effect (Jacobson, 2001), and Robots and Gold (Resnick, 1994). For each scenario, the questions were formulated with a mix of Multiple Choice Questions (MCQs) and open-ended questions. They were written in an appropriate way to be answered by both experts and novices alike. No formal declarative concepts were used. The detailed questions can be found here. Whereas the last two scenarios (butterfly effect and robots and gold) are thought experiments, the first three scenarios include familiar (scattering, traffic jam) or naturally observed behaviors (flocks of birds). Moreover, we distinguish two types of questions: whether it involves explaining the behavior of a system (e.g. *How is it that the birds fly in a flock?* in the third scenario) or defining rules to enact it (e.g. *How should the robots move in order to find the gold?* in the last scenario). Finally, the questions, even in the same scenario, can tackle different perspectives: local (agent perspective, e.g. *What does each person need to do in order to scatter?* in the second scenario) and global (external observers, e.g. *If you were up really high, and could watch your class scatter, what would you see from above when they're scattering?* in the second scenario).

Procedure
An iterative approach was followed to refine the questions and validate the instrument. A first version of the questionnaire was sent out to 11 participants of varying ages and educational backgrounds. The goal of this first validation was to evaluate the understandability of the scenarios and questions and to obtain a range of the possible answers from our target population in order to develop a coding scheme. Based on this first feedback, the questions were refined, and confusing statements were clarified. The first draft of the coding scheme was also developed. Our goal in the second iteration was to validate that the chosen questions assessed the set of complex systems concepts, further validate the understandability of the questions and refine the coding scheme. The second validation phase included two steps. First, we discussed and validated the questions and coding scheme with an expert in complex systems education to ensure that the five scenarios and accompanying questions were representative of all the complex systems concepts. Second, we followed a verbal cognitive methodology in which three novices (undergraduate students) and three experts (graduates/post graduates in swarm robotics) were asked to read the questions and describe verbally what they understood from the question and how they would answer the questions. Then they wrote the answers in the questionnaire. Based on this second validation, final refinements were done to the questions and coding scheme.

Subjects
The final version of the questionnaire was administered to two groups of participants:
- An ***expert*** group including eight subjects who responded to a personal request for participation sent by email. They were purposively sampled to include PhD students or holders in a related domain (mainly in swarm robotics, physics, and computer science). They self-rated their expertise in complex systems with an average of 5/7 and confidence in answers with an average of 5.625/7. The mean age of the group was 33.88 years (SD = 3.59), with six males and two females. Classified by ISCED Majors, 3 are in Engineering and 5 are in Science.
- A ***non-expert*** group including 29 subjects. They were recruited from requests diffused online and among students in university and high school. The mean age of the group was 24.21 years old (SD = 6.31) with 16 females (55%) and 13 males (45%). These included five students from high school, 25 in Bachelor, 7 in Masters, and 2 PhD students. Classified by ISCED Majors, 15 are in Science, 3 in Health, 4 in Engineering, 1 in Education, 2 in Social Science, and 4 in General Knowledge (high school students).

Scoring and Coding Scheme
The responses were scored based on the five key concepts mentioned previously. We consider 3 levels of understanding: a clockwork (corresponding to a centralized mindset) mental model, a complex systems mental model, and an intermediate or in-between level. Table 1 provides the concepts, focus, and the difference between

clockwork and complex responses (Jacobson et al. 2011, Yoon et al, 2019). Based on these concepts a detailed coding manual was developed for each of the questions (can be found here).

Table 1

General coding scheme: concepts, focus, and the difference between clockwork and complex responses.

Concept	Focus	Clockwork Mental Model	Complex Systems Mental Model
Control	The control is centralized or decentralized	Response indicates that the system is controlled by a leader (order is top-down)	Response indicates that the system is decentralized (order is bottom-up)
Causes	The number of causes that contribute to the outcome	Response attributes the outcome to one primary cause/factor	Response attributes the outcome to multiple causes/factors
Action Effect	Interaction between agents are linear/proportional or not	Small action→ small effect Large action→ large effect	Small action →large effects Effects of actions may not be repeatable
Agents Effect	Predictability of actions of agents	Actions of agents may be predictable	Actions are not predictable, random
Process	How the system works	System is an event defined by its global behavior	System is a dynamic process, self-organizes through agent interactions

The inter-rater reliability of the coding scheme was assessed with three raters coding 25% of the responses. The interrater agreement was measured using Cohen's Kappa and was found to be 0.66 which denotes a substantial agreement. The discrepancies were discussed and related details were added to the coding scheme to make it clearer. The remainder of the responses were subsequently split among the three raters to code.

Below we describe the traffic jam scenario and its coding scheme as an example. The instruction is:

> *On the road, each car followed only three rules: 1) If there is a car close ahead of you, slow down*
> *2) If there aren't any cars close ahead of you, speed up (unless you are already moving at the speed limit)*
> *3) Comply with driving regulations concerning all other elements on the road, such as: slow down if you detect a radar trap, traffic lights, accidents, entry ramps… (we attach a gif)*
> Question 1: What is causing this traffic jam? Select all applicable choices (radar trap, broken bridge, entry ramp with merging traffic, accident, nothing, other)
> Question 2: Explain in your own words how your choice(s) is (are) causing the traffic jam.
> Question 3: The radar trap is removed, and there are no accidents on the road, no broken bridge, no entry ramp or any other external events. Can a traffic jam still form? (Yes, No)
> Question 4: If yes, why do you think a traffic jam can still form? If No, explain in your own words, how the cars will behave on the road.

We target three main complexity concepts with this scenario: causes, control, and action effects. Examples of responses and scores are given in Table 2.

Table 2

Scores of 3 participants for the first scenario: Traffic Jam

ID	**Q.1**	**Q.2**	**Q.3**	**Q.4**	**Scores**
P1	Radar trap, Broken Bridge, Entry Ramp with merging traffic, Accident, Nothing	Slow down compounds resulting in a complete halt. Netlogo has this model and explains it well too.	Yes	As one car's behavior is dependent on others, the speeds and behaviors become nondeterministic. Even a slight slowdown to avoid bumping into the car in front can compound as the reaction time may have a slight delay as well.	Control (1) Causes (1) Action (1)
P2	Radar trap	Cars are driving slower after they stopped suddenly	No	They will behave freely with no stops or jams	Control (0) Causes (0) Action (0)

P3	Broken Bridge, Accident	They basically cause a blockage to traffic, so that vehicles 'pile up'	Yes	It also depends on the amount of cars on the street	Control (1) Causes (0.5) Action (0)

Results

For each participant, we summed up all the scores per concept, as well as per scenario. Each raw score was divided by the total possible score, to have standardized scores for the different concepts and scenarios. The total score was calculated as the average over all the concepts. A significant difference was found between the scores of experts and the non-expert over all concepts, except the concept “Causes” where the non-expert also scored high (see Figure 1). Considering scenarios, significant differences were found between the experts and non-expert participants in all scenarios except in the second (scatter) scenario (see Figure 2). In the robots and gold scenario, the difference was larger with a huge effect size (Cohen’s d = 2.2).

Figure 1

*Comparison of experts and non-experts over the concepts. * (**) indicates a significant difference with p-value<0.05 (<0.005) with a Mann-Whitney Test.*

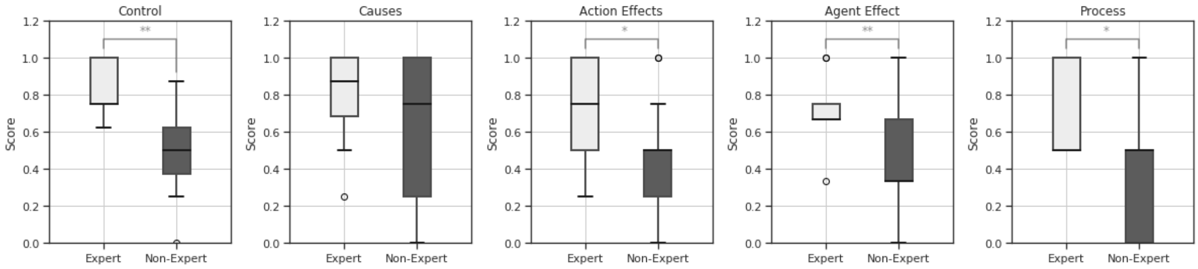
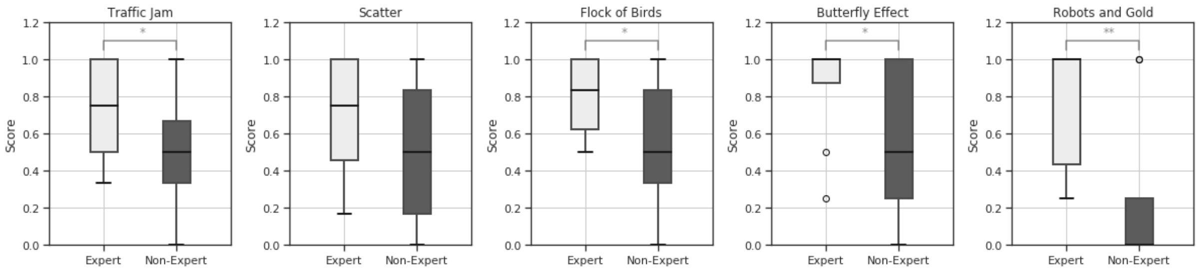


Figure 2

Differences per scenario among the experts versus the non-experts.

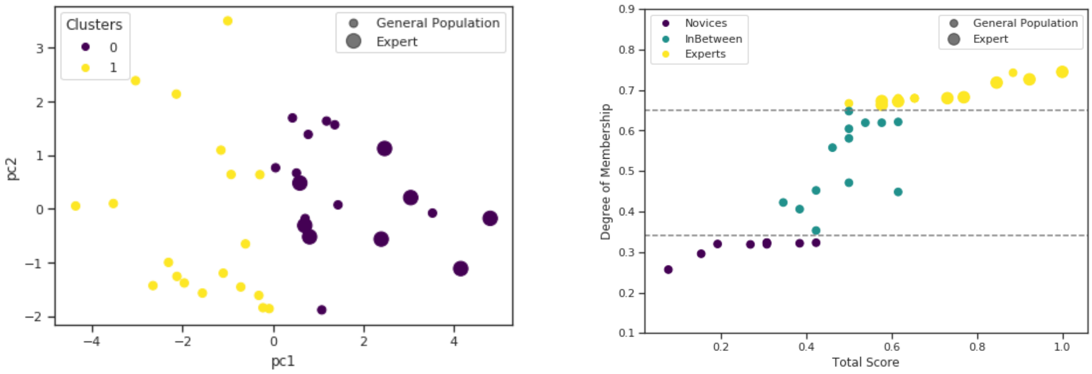


To verify that our mapping of concepts to questions was complete, we investigate the correlation between concepts. A strong significant correlation exists between total score and each of the respective concepts (spearman correlation with control = 0.75, action effects = 0.72, agent effects = 0.68, process = 0.66) and a moderate correlation with the causes concept (0.46). A moderate significant correlation also exists between the control concept and each of action effects (0.44) and agent effects (0.62) concepts and between action effects and process (0.48) concepts. A weak significant correlation (0.34) also exists between the concept action effects and agent effects. Furthermore, we investigate the correlation between scenarios. We only found weak significant correlation between scenarios 1 and 3 (0.35), as well as scenarios 2 and 4 (0.37).

While experts are expected to be knowledgeable on the concepts (and this assumption is supported by their high scores), participants belonging to the non-expert group can have varying degrees of understanding of complex systems, lying at different points on the continuum from novices to experts. Therefore, to verify whether the population can be divided into two groups, namely the novices and the experts, we conduct a clustering on the whole dataset involving the 37 participants on the basis of 10 features taken into account: the scores of each concept, as well as the scores of each scenario. A principal component analysis (PCA) is conducted to extract the 6 most significant principal components, explaining 95% of the variance. The purpose is to make groups which are hardly separable in the space by the original features, but better separable in the new space spanned by the Principal Components (PCs). To verify the separability of our population, we perform a C-means clustering for the association of each participant to a cluster. The number of clusters is fixed to 2 to verify whether the expert participants indeed cluster together when all features are considered together. The results, shown in Figure 3a, revealed that all the experts belong to the same cluster, as well as some of the participants from the non-experts. An extended discussion of this finding is reported in the Discussion section.

Figure 3

(a) Clustering results (b) fuzzy system output vs total score and division into 3 classes



(a) (b)

Figure 4

Differences per concept among the three groups.

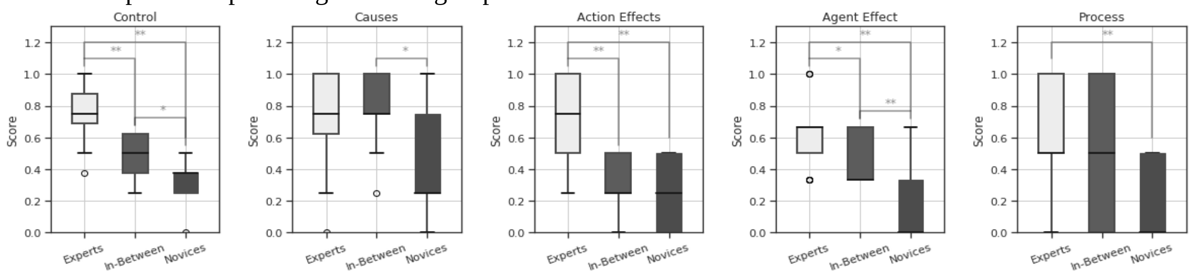
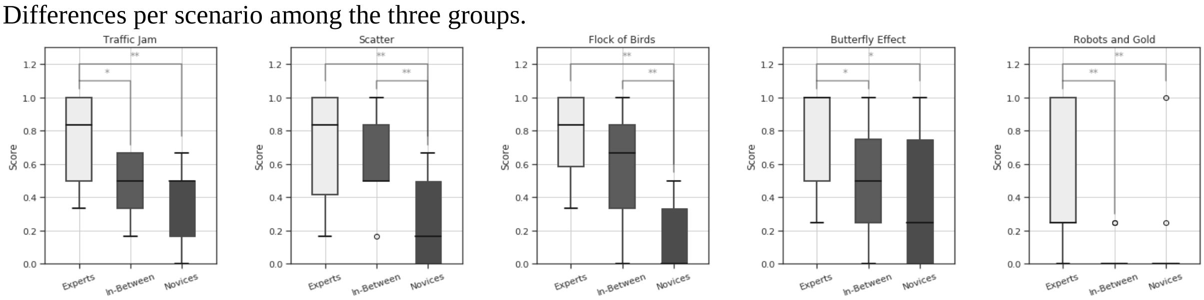


Figure 5

Differences per scenario among the three groups.



So far, we have analyzed our participants' responses either over single concepts/scenarios, or over the dimensions returned by PCA, which are a combination of these scores. What would be interesting is to have a global complex systems expertise score, which makes it immediate to evaluate a person's complex systems understanding. More concretely, our goal is to define a mapping between the concepts/scenarios and a global score of expertise. One approach is to sum and average over all the concepts; however, with this method, the distribution over all concepts is blended. We address the above challenges by using techniques from fuzzy logic systems. Fuzzy logic allows us to encode qualitative relations among multiple variables (fuzzy rules) in a unifying formalism (Driankov, 2001). In our case, the relation to describe is the one between a given vector of values for the complex systems' scores (concepts and scenarios), and the corresponding global level of expertise. The fuzzy rules are encoded as if-then rules underlying the qualitative fact that an expert level should have a high level in all the input variables; conversely, a novice level should have a low level in all the input variables. Each input membership function is modeled as a logistic sigmoidal function. The parameters of a sigmoid are fitted after applying a 1D clustering on each of the dimensions. The output membership function is modeled as a trapezoidal function with cut-off limits set between 0.35 and 0.65. Figure 3.b shows the relation between the total score and the output fuzzy logic score. We notice that the relation is sigmoidal rather than linear. Again, we verify that the group that we already know are experts actually score high on the degree of membership of the expert level. We also can identify a separation of three classes: experts scoring higher than 0.65, novices scoring lower than 0.35 and an average in-between group. We recomputed the differences per concepts and scenarios between the populations of three identified classes. Results are shown in Figures 4 and 5.

Discussions and Implications

In this paper, we propose a tool that can be used to assess a person's expertise in complex systems which requires solving five different scenarios and relies on a coding scheme to extract their competence in five ontological concepts required for complex systems understanding. A study with 37 participants including experts and non-experts was conducted to validate our instrument. Significant differences were found in the scores of the two groups. Finally, an automatic approach based on fuzzy logic for getting a general score is proposed and analyzed.

In our first analysis, we found a significant difference between the scores of experts and the non-experts over different concepts and scenarios, which is in line with previous literature (Jacobson, 2001). One exception is the concept of "Causes" where the non-experts scored high too (see Figure 1). While it may be possible that this concept is easier to grasp than others, an alternative explanation is that in our instrument the questions measuring causes were mainly MCQs, which possibly led respondents to select more than one option. Conversely, in the last scenario (robots and gold), the difference in scores was much larger: a possible explanation is that in this scenario, participants were asked to enact/design rules in a novel scenario, rather than explaining an observed phenomenon. Given the ubiquitousness of centrally controlled systems in the world around us and novices' inclination towards the centralized mindset (Resnick, 1994), it is likely that novices believe in the superiority of centralized over decentralized systems and so chose to enact centralized rules.

Subsequently, we designed a Fuzzy System to compute a global score of expertise in complex systems understanding. The score computed by the Fuzzy Inference System (FIS) is not always equal to the average score, since the FIS takes into account the combination of scoring high on the different features, rather than having more skewed scores by one high value. For example, a participant with very high scores on the *causes* concept and scenario 4 (both 1) but very low scores on the *control* concept, and scenario 5 (0.25, and 0 respectively), had the final score pushed down with comparison to the average score (0.44 instead of 0.61). Similarly, a participant with very low scores on the *causes* concepts, and scenario 2 (0.25 and 0.33) but high scores on the *control* concept, and scenario 5 (0.75, and 1 respectively), had the final score pushed up with comparison to the average score (0.67 instead of 0.58). Moreover, in our total score analysis, the results indicate a separation between 3 groups of expertise. The subjects already known to be experts all belong to the expert cluster. Few participants from the non-experts also belonged to the same expert group. By looking at their major, they were a group of students doing their masters in biology, so they perhaps have a background knowledge in biological complex systems, which could explain their performance. The other two groups had populations of mixed educational backgrounds. Although not significant, looking at the mean age of the clusters, the novice group had the lowest mean (22.0), then the in-between group (26.5), then the experts group had the highest (28.7).

To better understand the differences between the three groups, we look again at the difference between the concepts. Significant differences were found for the concepts: control, causes, and agent effects between the novices and in-between groups; and for the concepts: control, action effects and agent effects between the expert and in-between groups. For the scenarios, it was scatter and flock of birds between the novices and the in-between groups, and traffic jam, butterfly effect and robots and gold between the expert and the in-between groups. Looking back to the theoretical background, we observe similar trends to link difficulties related complex ideas. Yoon et al (2019), reported that order and deterministic effects (which link to control and agent and action effects in our definitions) constitute the hardest concept to grasp. Similarly, Resnick (1994) suggested that the commitment to centralized mindset constitutes a strong misconception for learners. Lastly, Grotzer et al. (2015) focused on the difficulty of understanding non-linear effects in a system.

Finally, although through our questionnaire we only target the implicit knowledge of complex systems, the use of declarative knowledge (emphasis in italics) was observed in the experts' responses (Jacobson et al., 2011). An example is the explanation of the butterfly effect given by an expert:

> In a system with *nonlinear dynamics* such as with *positive feedback* loops, small changes in certain periods or locations can have large effects. Collective decision-making for a new nest in ant groups is one example, where the number of insects recruited to go to a specific new nest depends on how many are recruiting […]

On the other hand, an answer to the same question, given by a participant from the non-experts and which also display implicit but lacks declarative knowledge is:

> I think those effects are the norm rather than an exception. For example, teachers influence the future life of students in small ways every day, and therefore also influence the lives of those they interact with.

In conclusion, the assessment, coding scheme and scoring approach proposed in this paper advance our understanding of people's understanding of complex systems. Moreover, our instrument contributes to the design of learning activities to engage students about complex systems and helps investigate the effectiveness of such activities. The differences per concepts and scenarios between three different groups observed in our study suggest to follow a learning progression moving from less to more advanced concepts, as well as following a learning process based on the conceptual change approach to trap existing conceptions from a clockwork mental model to a complex systems mental model, which will be part of our future work (Khodr, et al, 2022).

References

Brown, A. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of Learning Sciences*, *2*(2), 141–178

Charles, E. S., & d'Apollonia, S. (2004). Developing a conceptual framework to explain emergent causality: Overcoming ontological beliefs to achieve conceptual change. In K. Forbus, D. Gentner, & T. Reiger (Eds.), Proceedings of the 26th annual cognitive science society. Mahwah, NJ: Lawrence

Chi, M. T. H. (2005). Commonsense conceptions of emergent processes: Why some misconceptions are robust. The Journal of the Learning Sciences, 14(2), 161–199.

Driankov D. (2001) A Reminder on Fuzzy Logic. In: Driankov D., Saffiotti A. Fuzzy Logic Techniques for Autonomous Vehicle Navigation. Studies in Fuzziness and Soft Computing, vol 61, Heidelberg.

Grotzer, T. A., Powell, M. M., Derbiszewska, K. M., Courter, C. J., Kamarainen, A. M., Metcalf, S. J., & Dede, C. J. (2015). Turning transfer inside out: The affordances of virtual worlds and mobile devices in real world contexts for teaching about causality across time and distance in ecosystems. Technology, Knowledge and Learning, 20, 43–69

Hmelo-Silver, C. E., & Azevedo, R. (2006). Understanding complex systems: Some core challenges. Journal of the Learning Sciences, 15, 53–61.

Hmelo-Silver, C. E., Marathe, S., & Liu, L. (2007). Fish swim, rocks sit, and lungs breathe: Expert-novice understanding of complex systems. The Journal of the Learning Sciences, 16(3), 307-331.

Jacobson, M. J. (2001). Problem solving, cognition, and complex systems: Differences between experts and novices. Complexity, 6(3), 41-49.

Jacobson, M. J., Kapur, M., So, H. J., & Lee, J. (2011). The ontologies of complexity and learning about complex systems. Instructional Science, 39(5), 763-783.

Khodr, H., Kothiyal A., Bruno B. & Dillenbourg, P. (2022, May). Cellulan World: Interactive platform to learn swarm behaviors. In *Proceedings of the 21st International Conference on Autonomous Agents and MultiAgent Systems*.

Mitchell, M. (2009). *Complexity: A guided tour*. Oxford University Press.

Papert, S. (1980) Mindstorms: Children, Computers, and Powerful Ideas. Basic Books, Inc., New York, USA

Reed, S. K. (2020). Cognitive Skills You Need for the 21st Century. Oxford University Press, USA.

Repenning, A., Ioannidou, A., Luhn, L., Daetwyler, C. & Repenning, N. (2010). Mr. Vetro: Assessing a Collective Simulation Framework. Journal of Interactive Learning Research, 21(4), 515-537. Waynesville, NC: Association for the Advancement of Computing in Education (AACE)

Resnick, M (1994). Turtles, termites, and traffic jams: explorations in massively parallel microworlds; MIT Press, Cambridge, MA.

Wilensky U. & Rand W. (2015) An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo. Mit Press

Wilensky, U., & Abrahamson, D. (2006). Is a disease like a lottery?: Classroom networked technology that enables student reasoning about complexity. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA, April 9-13

Wilkerson-Jerde, M. H., & Wilensky, U. J. (2015). Patterns, probabilities, and people: Making sense of quantitative change in complex systems. Journal of the Learning Sciences, 24(2), 204–25

Yoon, S., Goh, S. E., & Yang, Z. (2019). Toward a learning progression of complex systems understanding. Complicity: An International Journal of Complexity and Education, 16(1).

Acknowledgments

This work was supported as a part of NCCR Robotics, a National Centre of Competence in Research, funded by the Swiss National Science Foundation (grant number 51NF40_185543). We thank all the participants and experts who participated in our study.