Schema abstraction with productive failure and analogical comparison: Learning designs for far across domain transfer

Michael J. Jacobson\textsuperscript{a,*}, Micah Goldwater\textsuperscript{a}, Lina Markauskaite\textsuperscript{e}, Polly K. Lai\textsuperscript{b}, Manu Kapur\textsuperscript{c}, Gareth Roberts\textsuperscript{a}, Courtney Hilton\textsuperscript{a}

\textsuperscript{a} The University of Sydney, Australia
\textsuperscript{b} Queensland University of Technology, Australia
\textsuperscript{c} ETH Zurich, Switzerland

A B S T R A C T

Although there has been considerable research into knowledge transfer for over a century, there remains a need for specific, validated techniques for teaching for transfer. This article reports on classroom-based research in which students learned about complex systems and climate change with agent-based computer models using two different instructional approaches based on productive failure (PF). In both PF approaches, students initially explored a problem space on their own and then received teacher-led instruction. One treatment group used climate computer models whereas the other group engaged in analogical comparisons between the same climate computer models and complexity computer models in different domains. The study found both groups demonstrated significant learning gains by posttest on assessments of declarative and explanatory knowledge and on within domain near transfer. However, students in the two models treatment group performed at a significantly higher level on an across domain far transfer problem solving task. Theoretical and practical implications are considered.

1. Introduction

“Transfer of learning is universally accepted as the ultimate aim of teaching. However, achieving this goal is one of teaching’s most formidable problems”

While not everyone might accept the assertion of McKeough, Lupart, and Marini (1995) that the ultimate goal of teaching is the transfer of learning, certainly there is high agreement it is an important and challenging goal to achieve. Since the early 20th century there has been considerable research and theorizing about cognitive and situative facets of transfer, what is transferred (e.g., abstract concepts, procedures), types of transfer (e.g., near within domain, far across domain), teaching approaches (e.g., analogical problem solving, metacognitive strategies), learner dispositions, and so on (for an early review, see McKeough, Lupart, and Marini (1995)).

Interest in transfer continues in the learning and cognitive sciences, but with shifts in areas of research focus. In the introduction to a special issue on transfer, Engle (2012) observed that researchers have been investigating expanded views of what learners transfer to include representations, perceptually-grounded principles, learning strategies, embodied episodic feelings, discourse practices, preparing for future learning, and so on. Engle further argued that because “of these expansions of what counts as transfer, unproductive arguments about whether or not transfer occurs have been replaced by careful empirical investigations of transfer that specify what kinds of transfer occurred and under what conditions” (p. 348).

Goldstone and Day (2012), in another special issue on transfer, note that although “schools often measure the efficiency of learning in terms of speed and retention of knowledge, a relatively neglected and subtler component of efficiency is the generality and applicability of the acquired knowledge” (p. 149). They identify three main research themes in their special issue: (a) the learner’s perspective, (b) motivation in determining transfer, and (c) validated teaching techniques for facilitating transfer. The research reported in this paper aligns the Engle’s theme of investigating the conditions under which certain kinds of transfer occur and with the Goldstone’s and Day’s theme to study and validate teaching techniques for transfer, in our case, by using computer models and visualization systems intended to foster deep learning of difficult scientific knowledge and the ability to apply this knowledge to new problems.

The paper discusses a study of learning for transfer that employed computer models to help high school students understand the science of climate change. The theories to inform the learning designs we employed in this study were analogical comparison (Gentner, Loewenstein, & Thompson, 2003) and productive failure (Kapur & Bielaczyc, 2012), both of which have had significant prior research in which learning for transfer has been demonstrated. This content area and the use of computer models was selected for four main reasons.
First, current scientific perspectives about atmospheric systems and climate change are examples of complex systems (Bar-Yam, 2003; Donner, Barbosa, Kurths, & Marwan, 2009; Rind, 2011). Second, these domains are now an important part of the school curriculum. For example, in the United States Next Generation Science Standards, climate change has been identified as a Disciplinary Core Idea and complex systems concepts are included as Crosscutting Concepts (National Research, 2013). Third, Next Generation Science Standards also recommend the use of computer models as part of Scientific and Engineering Practices. Fourth, helping students to learn and transfer these concepts and practices represents a significant instructional challenge that relates to the Goldstone and Day (2012) issue of developing and validating teaching approaches for fostering transfer in the “real world” of regular classroom teaching.

There are five main sections in this paper. We next provide an overview of the relevant literature that have informed this classroom-based research study and the methods we employed, followed by the results and discussion sections. The paper concludes with considerations of the educational and scientific importance of this research in terms of theorizing related to learning and to applying—i.e., transfering—challenging scientific knowledge and skills and of practical teaching practices.

2. Literature review

In addition to general transfer research we discussed above, there are three main literature that inform the research reported in this paper: (a) transfer based on analogical comparison, (b) transfer based on productive failure, and (c) learning complex systems with computational models. We provide overviews of these areas in turn.

2.1. Transfer based on analogical comparison

It is quite rare to find transfer of learning to new problems that do not also share surface commonalities. When this occurs, it is because of shared deep commonalities in the relations among key elements of the conceptual phenomena (Gick & Holyoak, 1980). For example, if one initially learns about the cause and effect relations of a positive feedback loop underlying the development of economic pricing bubbles, ideally the learner could then apply that understanding to generate an explanation and form predictions for how the same positive feedback structure is causing the polar ice caps to melt. However, learners often fail to see these kinds of connections in the shared causal structure across domains when the problems have very different surface features (e.g., economics and climate change) (Rottman, Gentner, & Goldwater, 2012).

There has been a considerable amount of research over the past thirty-five years into various facets of learning for transfer (e.g., Catrambone & Holyoak, 1989; Gadgil, Nokes-Malach, & Chi, 2012; Gentner et al., 2003; Gick & Holyoak, 1983; Goldwater & Gentner, 2015; Kurtz & Loewenstein, 2007). Alfieri, Nokes-Malach, and Schunn (2013) published a meta-analysis of this literature in which a consistent finding was that learning by comparing two cases results in greater learning for transfer than learning from a single case. Another key finding was that learning from comparing two cases resulted in greater transfer if there were explicit scaffolds to compare than if the two cases were considered without this scaffolding. In analogical encoding theory, Gentner et al. (2003) propose that comparison fosters the representation of the relations across both cases as a more abstract schema (i.e., schema abstraction), which in turn enables the spontaneous recognition of key structural relations in new problems and situations. In other research, it was further found that presenting a key underlying principle after learners compared cases elicited greater learning and transfer benefits than only scaffolding the comparison of two cases (e.g., Schwartz, Chase, Oppezzo, & Chin, 2011). This last finding aligns with the work on productive failure that we next discuss.

2.2. Transfer based on productive failure

Research into productive failure (PF) designs for learning has been ongoing for over a decade (Kapur, 2008, 2010, 2014, 2016; Kapur & Bielaczyc, 2012; Kapur, Voiklis, & Kinzer, 2008). Briefly, PF involves two main phases. First, students are engaged in an Exploration and Generation phase as they work on a problem where they might struggle since they are novices who have incomplete or inaccurate knowledge necessary for an appropriate solution. Next is the Consolidation and Knowledge Assembly phase in which students compare their representations and solution methods (RSMs) to those provided by the teacher. Kapur and Bielaczyc (2012) have theorized that the generation and Exploration activity helps students activate their prior knowledge during the initial problem activity, and that the comparison of the RSM generated by a student to the teacher’s RSM sets up a cognitive process of schema abstraction and, in turn, the assembly of new knowledge by the student. There have been consistent findings that students in PF treatment conditions demonstrate higher results on assessments of learning and/or transfer than students in control conditions who were directly taught correct concepts and procedures without having the initial Exploration and Generation phase (Kapur, 2016; Loibl, Roll, & Rummel, 2017). Also, of specific relevance to the research reported in this paper, PF approaches have been successfully used with agent–based models for learning scientific knowledge and skills related to electricity (Jacobson, Kim, Pathak, & Zhang, 2015) and to climate change and complex systems (Jacobson et al., 2017).

There are clear instructional similarities between PF phases—with students comparing their RSMs from generated solutions to those of the teacher—and analogical comparison activities—with scaffolding the comparison of cases. Further, both of these approaches propose that the learning and transfer findings are due to cognitive theories of schema abstraction and generalized schema construction. However, there has not been a study that explicitly compares learning designs for transfer based only on PF to a hybrid design the combines PF with analogical comparison, which is the focus of the research we report in this paper.

2.3. Learning about complex systems with computational models

Computational modeling and scientific visualization tools are being increasingly used in modern scientific practice (e.g., Clement, 2000; National Research Council, 2012; Latour, 1987; National Research; Wilensky & Rand, 2015). In terms of science education, it is now possible to use computer modeling and visualization technologies that are similar to, or even the same as, those used by scientists (Jacobson & Wilensky, 2006). Of relevance to the research in this paper, there has been extensive use of computer models, especially agent-based models (Wilensky & Rand, 2015), for research into learning about complex systems (see Jacobson et al. (2017) for a recent overview of this research). Researchers in this area have demonstrated significant learning of scientific ideas both from students creating their own agent-based models (sometimes referred to as “modeling”) (e.g., Blikstein & Wilensky, 2010; Wilensky, 1996; Wilensky & Reisman, 2006) as well as from students using pre-constructed agent-based models (e.g., Jacobson, Richards, Taylor, Wong, & Newstead, 2011; Levy & Wilensky, 2009; Sengupta & Wilensky, 2009).

Goldstone and Wilensky (2008) have proposed that agent-based models not only help students learn complex systems principles, but in conjunction with understanding these principles, also can help foster transfer. Learning of complex systems principles can be fostered in part because computer models can be designed to be spatially-temporally grounded to align with mental models and perceptual capabilities learners have (Goldstone & Wilensky, 2008; Markauskaite, Kelly, & Jacobson, 2017). Further, designing the models to be idealized or schematic (not overly detailed) can focus on key conceptual aspects of complex systems related to the phenomenon being modeled, which can both reduce cognitive load and enhance transfer.
Goldstone and Wilensky (2008) propose that far across domain transfer could be fostered “not by applying a rule from one domain to a new domain but rather by allowing two scenarios to be seen as embodying the same principle” (Goldstone & Wilensky, 2008, p. 507). While this is intriguing and even provocative, it must be noted much of the research cited by Goldstone and Wilensky (2008) was conducted in controlled experimental settings. A critical question thus remains: Can instructional approaches allow students to learn complex systems principles using agent-based models in ways that foster far across domain transfer in “real world” classroom settings? The study reported in this paper was, in part, an attempt to answer this question.

2.4. Research overview

In this study, students used computer models as part of two different learning designs based on PF (Kapur & Bielaczyc, 2012) to learn about complex systems and climate change. The experimental treatment Two Models Group worked on weather/climate and complex systems problems over the four days of the study and used a hybrid PF and analogical comparison design with two different agent-based computer models each day: one model to work on a weather or climate problem and one model to work on a complex systems problem in a domain other than climate (and each day, this second model was a different non-climate domain, i.e., the Two Model Group learned using models of four distinct non-climate domains). The control treatment One Model Group also used a PF design over four days and worked on the same weather/climate and complex systems problems as the Two Models Group. However, the One Model Group used the climate model for both activities. Importantly, weather/climate models and non-climate models given each day to the students were analogs in that they implemented the same underlying principles of complexity. Finally, both the Two Models Group and One Model Group worked on the same problems and had the same amount of time for completing all activities. In sum, both conditions were instructed on the same climate and complex system principles in the same amount of time, but the One Model group learned using climate models only, whereas the Two Model group learned using the climate models and four analog models from distinct non-climate domains.

This study explored four specific research questions:

- Research question 1 (RQ1): Does having one computer model increase development of declarative and explanatory knowledge about the climate versus using two computer models?
- Research question 2 (RQ2): Does having two computer models to compare increase development of declarative and explanatory knowledge about complex system principles versus using one computer model?
- Research question 3 (RQ3): Does having two computer models to compare increase near within domain transfer of complex system principles versus using one computer model?
- Research question 4 (RQ4): Does having two computer models to compare increase far across domain transfer of complex system principles versus using one computer model?

For RQ1, it was hypothesized that there would be higher learning outcomes related to declarative and explanatory knowledge about the targeted weather and climate ideas by the One Model Group as they would be using the climate models for a longer period of time than the Two Models group. Another reason the Two Models Group might perform at a lower level on these items is they might be confused or experience greater cognitive load than the One Model Group. Regarding RQ2, it was expected that all students would have little understanding of the targeted complexity ideas such as dynamic equilibrium, dynamic systems feedback, emergence, and tipping points. It was hypothesized that the students in the Two Models Group would demonstrate greater gains on the declarative and explanatory knowledge items than the One Model Group as they would have used complex systems computer models as well as the weather or climate models. A counter expectation would be lower performance on the explanatory knowledge because of student confusion or increased cognitive load related to using two computer models rather than one.

RQ3 and RQ4 both concern transfer assessments for near within domain and far across domain transfer of complex systems ideas. Students in the Two Models Group were expected to have higher scores on these items for two reasons. First, they would have greater explanatory knowledge of complex systems ideas (i.e., RQ2’s expected outcome). This would set up an important condition for transfer, which is initially understanding an idea before being able to use the idea on a transfer task (Chi & VanLehn, 2012). Second, the activity of comparing the two different models in the Exploration and Generation phase was expected to have the students activate more relevant prior knowledge than the students in the One Model Group, which in turn would help the students benefit more from the Consolidation and Knowledge Assembly phase, leading to higher performance on the near and far transfer items. Although not expected, as with RQ2, lower performance on both of the types of transfer might result due to student confusion or enhanced cognitive load associated using two computer models rather than one. We next discuss the methods we employed in this study to investigate these research questions.

3. Methods

3.1. Participants

The study was conducted in an all-girls school with high academic achieving students in Australia. There were 114 ninth grade students in four different classes, with some eliminated from the final analysis because of missing pretest or posttest data. The final analyses included 96 students across the four classes: The One Model Group classes were 27 (one eliminated) and 30 (zero eliminated), and the Two Models Group classes were 29 (one eliminated) and 19 (eight eliminated). 1 (See the online supplementary materials link provided in Appendix A to access a table showing more information about their demographics (such as the languages other than English spoken at home, and their recent science marks)). Students worked in pairs during the Exploration and Generation phase, but they were assessed on the pretest and posttest individually. In both conditions, participants were randomly paired based on their student number.

3.2. Materials

The main learning materials consisted of four weather and climate computer models—Carbon Cycle, Climate with Feedback, Wind and Storm, and Climate with Water Feedback (see Fig. 1)—developed in NetLogo (Wilensky, 1999) by one of the participating science teachers. These models were used by both treatment groups. The Two Models Group also used four models from the NetLogo Model Library: Wolf-Sheep Predation, Ants Foraging, Birds Flocking, and Forest Fire (see Fig. 2).

An online Student Guidebook was developed that provided the students with daily instructions for using the computer models, the “challenge problems” they worked to answer using the computer model or models, and fields to type up and submit their answers to the problems. Also, a Teacher’s Guide was developed for each treatment condition that provided written instructional scripts for the climate and

1 We discussed the higher number of eliminated students in the Two Models Group with the main classroom teacher we coordinated the study with. He indicated this was simply scheduling activities and individual sick days for students in this treatment group and there was no indication these eliminated students had issues with the intervention.
complex systems concepts and were virtually identical for teachers in the two conditions for each of the four days of the interventions. The main difference in the instructional scripts used by the teachers was that for the Two Models Group, the teacher discussed climate concepts referring to the climate models and the complex systems concepts referring to the complexity models, whereas in the One Model Group, the teacher discussed the same climate and complexity concepts but referred only to the climate models.

Table 1 shows the daily topics and content knowledge for both treatment groups for the six days of the study. The pretest and posttest were on days one and six respectively, and the learning activities were on days two to five. A set of “challenge problems” were authored for each day with the intent that a scientifically correct answer would require using weather and/or climate specific ideas in conjunction with “crosscutting” concepts about complex systems behaviors, such equilibrium, feedback, micro/macro system levels, emergence, and so on. The learning trajectory was designed to systematically build on core complex systems concepts aligned with the weather and climate ideas, leading to linearity and non-linearity on day five, which earlier research has demonstrated are particularly difficult complexity concepts for students to learn (Jacobson et al., 2011). The pretest and posttest, along with the student guidebooks, are included in the online supplemental materials (see Appendix A).

3.3. Experimental design

This experiment was conducted over six class periods (one per day) of 80 min each. The first presented the pretest; the next four consisted

<table>
<thead>
<tr>
<th>Day</th>
<th>Complex Systems Daily Topics</th>
<th>Climate Model</th>
<th>Complex Systems Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pre-test</td>
<td>Carbon Cycle</td>
<td>Wolf-Sheep Predation</td>
</tr>
<tr>
<td>2</td>
<td>Dynamic equilibrium</td>
<td>Climate topics: Equivalent carbon, carbon cycle, fossil fuel, carbon sinks</td>
<td>Climate topics: None</td>
</tr>
<tr>
<td></td>
<td>Complexity topics (both models): Equilibrium, dynamic equilibrium, closed system</td>
<td>Climate topics: None</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Feedback in dynamic systems</td>
<td>Climate with Feedback</td>
<td>Ants Foraging</td>
</tr>
<tr>
<td></td>
<td>Climate topics: Global temperature, greenhouse effect, cloud cover</td>
<td>Climate topics: None</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Complexity topics (both models): Input, output, positive feedback, negative feedback, self-organization</td>
<td>Climate topics: None</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Emergence</td>
<td>Climate topics: Convection, wind, greenhouse effect, enhanced greenhouse effect</td>
<td>Birds Flocking</td>
</tr>
<tr>
<td></td>
<td>Complexity topics (both models): Emergence, micro level of systems, macro level of systems</td>
<td>Climate topics: None</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Tipping points and positive feedback</td>
<td>Climate with Water Feedback</td>
<td>Forest Fire</td>
</tr>
<tr>
<td></td>
<td>Climate topics: Atmospheric water feedback, albedo</td>
<td>Climate topics: None</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Complexity topics (both models): Positive feedback, tipping points, linear versus non-linear effects</td>
<td>Climate topics: None</td>
<td></td>
</tr>
</tbody>
</table>
of the learning activities; the sixth presented the posttest.

This study employed a quasi-experimental design (for the components of the study, see Table 2).2 In the PF Exploration and Generation phase, the two groups each worked on identical challenge problems, however, the One Model Group used a single weather or climate model whereas the Two Models Group students used a complex systems model (from a different non-climate domain each day) and then the same weather or climate models as the other group. The Two Models Group may be regarded as a hybrid learning design with the same PF phases as the One Model Group, but with two models to compare in the Exploration and Generation phase rather than one model. Thus, this hybrid PF learning design had highly contrasting surface elements because each complexity model was from a non-climate domain, such as the Day 3 Climate Model with Feedback versus the Ants Foraging Model, while they shared the deep structure conceptual similarity of feedback in dynamic systems. Besides using one or two models, another difference between the groups was that the Two Models students were instructed (i.e., scaffolded) to write out similarities and differences of the models. This was done as research such as Gentner et al. (2003) has shown that without explicit instruction to write out similarities, learners often will not deeply compare pairs of analog cases. In contrast, the One Model group was instructed to write “the key ideas about this model.”

The duration of each phase of the 80 min class periods were the same for both conditions: Exploration and Idea Generation Phase was 20 min; Consolidation Phase was 20 min; Knowledge Assembly Phase was 40 min. Exploration and Generation phase was the same for both groups (see Table 2). The student dyads collaboratively worked on the challenge problems. During the Consolidation and Knowledge Assembly phase, the teachers used the same instructional scripts that discussed the relevant weather or climate ideas and the targeted complex systems concepts, and the students in both treatment groups worked on the same challenges after the instruction provided by the teachers.

There were three teachers who taught the four classes. One teacher had a Ph.D. in physics and helped design the climate change agent-based models. This teacher was our “content expert.” A second teacher had 10 years science teaching experience, and one was an early career teacher. Neither of these teachers had prior expertise with the complex systems concepts that the content expert had. All three teachers received professional development on the PF instructional approach that also covered the weather and climate and the complex systems content and using the NetLogo computer models for the two teachers without that prior knowledge. The content expert taught two classes, a One Model class and a Two Models class. The early career science teacher taught a Two Models class, and the more experienced teacher taught a One Model class. We assigned the teachers to these conditions to ensure that any benefit of having Two Models did not depend on either having content expertise nor at least 10 years of teaching experience. Qualitative classroom observations, which were videotaped, indicated that the teachers implemented the two different treatment conditions as requested by the research team.

---

2 We note in this study that there are no control conditions that did not employ PF or did not use computer models. In our view, there have already been several studies that have demonstrated significantly higher learning and transfer outcomes of PF compared to non-PF treatments in noncomputer-based interventions (e.g., Kapur, 2008, 2011; Kapur, 2012, 2014) and in one computer model-based interventions (Jacobson et al., 2015). Given the results of these prior studies, we did not include a non-PF control condition in this study, so we could better focus on the research questions with a simpler two treatment group experimental design. Also, there has been research into students learning climate change ideas using more conventional instructional approaches (i.e., no use of computer models), and the findings have been consistently disappointing (for a review of this literature, see Shepardson, Choi, Niyogi, and Charusombat (2011)).
Table 3
Sources, types of information, and sample questions.

<table>
<thead>
<tr>
<th>Source</th>
<th>Type</th>
<th>Sample Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-test survey on background information</td>
<td>Personal information (7 items)</td>
<td>What is your date of birth? Is English the primary language spoken at home?</td>
</tr>
<tr>
<td>Question 1 to 5 (Pre-test and Post-test)</td>
<td>Climate systems ideas (5 questions): All 5 were coded for both declarative knowledge explanatory knowledge</td>
<td>Please explain what the carbon cycle is. Why is it important?</td>
</tr>
<tr>
<td>Question 6 to 9 (Pre-test and Post-test)</td>
<td>Complex systems ideas (4 items): Explanatory knowledge</td>
<td>What are examples of emergent properties in climate systems? Please explain.</td>
</tr>
<tr>
<td>Problem Solving 1 (Pre-test and Post-test)</td>
<td>Solving a climate and complex systems problem (one question for near within domain transfer)</td>
<td>“Butterfly Effect” Problem: It has been said that a butterfly flapping its wings in Brazil can jiggle the air and thus can help cause a snowstorm in Alaska. Is this possible? If so, how? If not, why not?</td>
</tr>
<tr>
<td>Problem Solving 2 (Post-test)</td>
<td>Complex Systems transfer questions (one question for far across domain transfer)</td>
<td>“Robot Mining” Problem: How can autonomous robots on a remote planet effectively and efficiently mine gold?</td>
</tr>
</tbody>
</table>

3.4. Data sources

Assessments of declarative and explanatory knowledge and of near within domain and far across domain transfer were the main sources of data for this study. See Table 3 for the sources, types of information, and sample questions associated with these assessments. Five items were used on the pretest and posttest to assess declarative and explanatory knowledge about weather and climate ideas (each of the five questions had two parts, and were coded for both kinds of knowledge), and four items for explanatory knowledge about complexity ideas were used (declarative knowledge about complexity was not assessed as these topics are not in the current Australian science curriculum). These nine questions each assessed a different concept (e.g., the carbon cycle, positive feedback, respectively). A question about the “butterfly effect” was used as a near within domain transfer assessment since an appropriate solution would require relevant concepts about both weather and complex systems, which were ideas focused on in the instructional intervention. Far across transfer domain transfer was assessed with the robot gold mining on another planet problem (Resnick, 1994), which requires the use of complexity ideas such as positive feedback and self-organization but in a domain that has surface features completely different than weather or climate phenomena and different than the complexity models the Two Models Group students worked with. This item was only included on the posttest because of research that suggests pretest items can influence learning outcomes by providing cues to the students about what to focus on during a learning intervention (Murry, 1990).

3.5. Data analysis and item reliability

All items were open-ended problems requiring written explanatory answers, which were scored using a rubric on a scale of 0–3 scale (see Table 4) that had been validated by a Ph.D. physics expert. Examples of the participant posttest transfer problem solutions and rubric scores are shown in Table 5. Twenty-five percent of the responses were independently coded by two trained raters, with an inter-rater reliability of 97% (see Appendix A for the link to the online supplemental materials for inter-rater reliabilities for each assessment). After discussions, the two raters reached 100% agreement. The remaining responses were then split between the raters and independently coded.

To assess the reliability for each composite score used for research questions 1 to 3, we computed McDonald’s ω and the Greatest Lower Bound (GLB) for the five items assessing declarative climate knowledge, for the five items assessing explanatory climate knowledge, and for the four items assessing explanatory complexity knowledge. Both reliability measures are argued to be better alternatives to Cronbach’s alpha in assessing test reliability under naturalistic conditions, particularly when the tests have few items, contain items that are skewed, and items that differ in quality due to qualitative responding (Trizano-Hermosilla & Alvarado, 2016). Both measures are on a scale between 0 and 1. We also performed principal components analysis on each test, extracting the first principal component, and then examining how much variance the first component predicted in the total test. Finally, we calculated test-retest reliability by examining the bivariate correlations between pretest and posttest performance for each test. All reliability metrics are shown in Table 6 (see Appendix A for the link to the online supplemental materials for further breakdowns of the correlations between pretest and posttest by condition).

We argue that all three composite scores show acceptable reliability for an educational assessment that exclusively uses open-ended questions where the goal is to analyze the effects of instruction, not to create a psychometric instrument to discriminate between student skill levels (i.e., wherein reliability standards must be higher). First, the majority of the tests show ω and GLB values over 0.4, indicating that internal

<table>
<thead>
<tr>
<th>Rubric Score</th>
<th>Answer type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Incorrect response</td>
<td>Scientifically irrelevant</td>
</tr>
<tr>
<td></td>
<td>Frivolous answer</td>
<td>Incorrect idea</td>
</tr>
<tr>
<td></td>
<td>Irrelevant ideas</td>
<td>Shows lack of seriousness in ideas presented</td>
</tr>
<tr>
<td>1</td>
<td>Partly correct</td>
<td>Saying same thing as in questions</td>
</tr>
<tr>
<td></td>
<td>No scientifically correct terminology used</td>
<td>Use of relevant terms without explanation or further elaboration or relating one ideas</td>
</tr>
<tr>
<td></td>
<td>Incomplete explanations</td>
<td>Using terms correctly but having these terms explained incorrectly</td>
</tr>
<tr>
<td>2</td>
<td>Correct answer with some scientifically correct terminology used</td>
<td>Use of two or more scientific terms that are explained correctly</td>
</tr>
<tr>
<td></td>
<td>Correct answer with complete explanations</td>
<td>Scientifically correct explanation on how those terms are related</td>
</tr>
<tr>
<td>3</td>
<td>Correct answer with scientifically correct terminology used</td>
<td>Reasonably full and complete expert explanation, which shows understanding and further relationships to relevant ideas</td>
</tr>
<tr>
<td></td>
<td>Demonstrate full understanding of the concepts with complete explanations and examples</td>
<td>Use of scientifically correct terminology</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scoring</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Scientifically irrelevant</td>
</tr>
<tr>
<td>1</td>
<td>Incorrect idea</td>
</tr>
<tr>
<td>2</td>
<td>Shows lack of seriousness in ideas presented</td>
</tr>
<tr>
<td>3</td>
<td>Saying same thing as in questions</td>
</tr>
<tr>
<td>4</td>
<td>Use of relevant terms without explanation or further elaboration or relating one ideas</td>
</tr>
<tr>
<td>5</td>
<td>Using terms correctly but having these terms explained incorrectly</td>
</tr>
<tr>
<td>6</td>
<td>Use of two or more scientific terms that are explained correctly</td>
</tr>
<tr>
<td>7</td>
<td>Scientifically correct explanation on how those terms are related</td>
</tr>
<tr>
<td>8</td>
<td>Reasonably full and complete expert explanation, which shows understanding and further relationships to relevant ideas</td>
</tr>
<tr>
<td>9</td>
<td>Use of scientifically correct terminology</td>
</tr>
<tr>
<td>10</td>
<td>Presence of more terms which explains further the concept</td>
</tr>
</tbody>
</table>
Table 5
Examples of participant posttest transfer problem solutions and rubric scores.

<table>
<thead>
<tr>
<th>Rubric Score</th>
<th>Problem Solutions</th>
<th>Robot Mining (Far transfer)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>I do not believe that a butterfly flapping its wings in Brazil will help cause a snowstorm. Firstly, the butterfly is too far away from both Alaska and any form of snow to have an effect on both snow, or Alaska. Secondly, the snow would probably not give any feedback in response to the little amount of ‘jiggles air’ caused by the butterfly. Also, the butterfly would be too small to be able to cause much of an impact upon the large volume.</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>A butterfly flapping its wings in Brazil can affect the air and thus contribute to the cause of a snowstorm in Alaska. This probably could be possible, because even the slightest movement could affect wind movements on a small scale. In this way, it is possible that the butterfly disrupted a small amount of air, which would disrupt a larger amount, and so on so forth until the winds move in a way so that a snowstorm could be caused in Alaska.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Although butterflies are small creatures, they can cause big things to happen in the world by triggering other forces of nature. Thus, it is possible that they may cause a snowstorm. The butterfly can cause a snowstorm because on a micro level, this positive feedback can disrupt a dynamic equilibrium. If the current motion of its wings disturbs something else, this can cause the snow to move because snow is loosely packed, meaning that it can easily be disturbed. This dynamic equilibrium that has been disturbed will cause other forces of nature to react, thus causing more influential objects to have a bigger impact on the snow, which might cause a snowstorm in Alaska. It is clear that a butterfly may have small power but its influential skills affect many creatures and forces around it. This goes to prove that it can cause a snowstorm.</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>This depends whether you’re taking the statements literally or figuratively. It literally, it’s impossible, if figuratively, it’s possible to an extent. The actual flapping of a butterfly’s wings creates so little movement that it can’t be felt a metre away. This would be immediately absorbed/taken over by the wind, which in itself doesn’t help cause an Alaskan snowstorm. An actual butterfly flapping its wings could not possibly make any difference in Alaska. However, if the phrase taken figuratively to mean that ‘small’ changes can have big impacts, then it could be quite possible. If another quarter of the Amazon were to be chopped down or burned, that would release an awful lot of carbon, making the air heat faster and more. This could, in turn, generate much stronger winds that could actually affect Alaska and help create a snowstorm. In conclusion, a literal butterfly could not possibly make a difference in Alaska. However, if the phrase is taken literally, then ‘small’ changes can have large impact.</td>
<td></td>
</tr>
</tbody>
</table>

Table 7 shows the means and standard deviations on the main consistency of each test is acceptable, especially given the low number of test items and the small sample size (when compared to more psychometrically-focused studies). Second, the majority of the first principal components extracted for the composite scores were able to explain more than 30% of the total test variance and had screen plots strongly suggesting a one-factor solution. This suggests that the composite scores are measuring one construct. Finally, all the test-retest correlations were positive and significant, indicating consistent performance over time to the same test. In summary, these reliability metrics suggest that our composite test scores are reliably measuring the knowledge constructs we intended to measure.

4. Results

Table 7 shows the means and standard deviations on the main assessments of conceptual understanding related to weather and climate change and to complex systems on the pretest and posttest, as well as the results of the far across domain transfer posttest item. Fig. 3 shows a plot of the total means and standard deviations for the posttest results. Due to the quasi-experimental design of our experiment, a linear mixed-model effects analysis (implemented using maximum likelihood) was used to assess how our hybrid productive failure and analogical comparison design affected performance on the knowledge assessments. To control for teacher and student variability, we entered Time (Pretest versus Posttest), Condition (One Model Group versus Two Models Group), Teacher Expertise (Content-expert vs. Non-expert A + B) as well as the two- and three-way interactions as fixed effects for each model (see Table 8 for the means and standard deviations for experienced and early career teachers in the one model and two models groups). Subject nested within Classroom was entered as a random
Researchers who might be interested in further analyses, such as comparing subjects and within the different classrooms, and subsuming the variance statistically significant using Wald’s test, \( p < 0.003 \), indicating variability within classroom as a random effect (AICc = 142.734, BIC = 174.094) than for the model when it was not (AICc = 142.734, BIC = 174.094). The random effects covariance parameter estimates of subject nested within classroom was 0.056 (SE = 0.015, [95% CIs = 0.027–0.084]), and statistically significant using Wald’s test, \( p < 0.001 \), indicating variability within subjects and within the different classrooms, and subsuming the variance from our fixed effects analysis. For the fixed effects factors, there was an effect of Time, \( F(1,92) = 10.873, p = 0.001, \eta^2 = 0.12 \), indicating that both groups demonstrated learning of weather and climate knowledge, as evidenced by higher performance on the posttest (\( M = 0.935, SD = 0.26 \)) compared to the pretest (\( M = 0.835, SD = 0.24 \)). There was no effect of Condition, \( F(1,92) = 1.001, p = 0.319, \eta^2 = 0.01 \), nor was there an interaction between Condition and Time, \( F(1,92) = 0.059, p = 0.809, \eta^2 > 0.01 \). There was also no effect of Teacher Content Expertise on climate declarative knowledge, \( F(1,92) = 0.319, p = 0.573, \eta^2 = 0.003 \), nor was there an interaction between Teacher Content Expertise and Time \( F(1,92) = 0.019, p = 0.889, \eta^2 > 0.01 \), or Teacher Content Expertise, Time, and Condition, \( F(1,92) = 0.538, p = 0.465, \eta^2 = 0.006 \).

A separate linear mixed model was used to assess explanatory knowledge. Again, there was a better fit for the model containing subject nested within classroom as a random effect (AICc = 142.734, BIC = 174.094) than for the model when it was not (AICc = 142.734, BIC = 174.094). The random effects covariance parameter estimates of subject nested within classroom was 0.056 (SE = 0.015, [95% CIs = 0.027–0.084]), and statistically significant using Wald’s test, \( p < 0.001 \), indicating variability within subjects and within the different classrooms, and subsuming the variance from our fixed effects analysis. For the fixed effects factors, there was an effect for Time, \( F(1,92) = 11.363, p = 0.001, \eta^2 = 0.11 \), with both conditions improving in their ability to explain climate knowledge, as evidenced by higher performance on the posttest (\( M = 1.056, SD = 0.3371 \)) compared to the pretest (\( M = 0.896, SD = 0.3795 \)). There was no effect of Group, \( F(1,92) = 0.085, p = 0.735, \eta^2 = 0.001 \), or Teacher, however, the two-way interaction between Group and Teacher Content Expertise was significant \( F(1,92) = 4.242, p = 0.042, \eta^2 = 0.04 \).

To examine this interaction in more detail, we performed all possible pairwise comparisons, using Student’s t tests Bonferroni-corrected. Despite a trend for participants exposed to the content-expert in the Two Models Group to perform higher, and a separate trend for participants exposed to the early career non-expert teacher to perform higher in the One Model Group, neither of the pairwise comparisons were significant. To further confirm that this interaction effect was spurious, we compared each factor level in the linear model to that of the overall average of the model. That is, we compared the One Model Group with content-expert teacher to the overall average of all the comparisons, and so on, for each comparison. Similarly, to the pairwise comparisons, there was no difference from the overall average for any of the factor levels, indicating that the interaction effect is likely to be spurious because the only pairwise comparisons that were significant were contradictory to each other (e.g., content-expert teacher and one model performs highest, followed by inexperienced non-expert with two models). Furthermore, when all the pairwise comparisons were individually compared to the mean of all the pairwise comparisons, there was no varying effect of teacher expertise or experience between the two groups (i.e., explanatory knowledge did not differ between the two groups as a function of teacher expertise). In summary we concluded that the “significant” interaction was spurious and entirely driven by the pairwise results that were both in the opposite direction to each other. These results show that both treatment Groups significantly improved in their

---

**Table 7**

<table>
<thead>
<tr>
<th>Assessment Type</th>
<th>Test</th>
<th>One Model Group</th>
<th>Two Model Group</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Climate Declarative Knowledge</td>
<td>Pretest</td>
<td>0.821 (0.217)</td>
<td>0.856 (0.261)</td>
<td>0.839 (0.239)</td>
</tr>
<tr>
<td></td>
<td>Posttest</td>
<td>0.914 (0.275)</td>
<td>0.964 (0.240)</td>
<td>0.940 (0.258)</td>
</tr>
<tr>
<td>Climate Explanatory Knowledge</td>
<td>Pretest</td>
<td>0.907 (0.373)</td>
<td>0.880 (0.377)</td>
<td>0.894 (0.423)</td>
</tr>
<tr>
<td></td>
<td>Posttest</td>
<td>1.031 (0.379)</td>
<td>1.022 (0.253)</td>
<td>1.027 (0.316)</td>
</tr>
<tr>
<td>Complex Systems Explanatory Knowledge</td>
<td>Pretest</td>
<td>0.240 (0.161)</td>
<td>0.338 (0.285)</td>
<td>0.291 (0.223)</td>
</tr>
<tr>
<td></td>
<td>Posttest</td>
<td>0.597 (0.417)</td>
<td>0.650 (0.324)</td>
<td>0.624 (0.371)</td>
</tr>
<tr>
<td>“Butterfly Effect” Problem (Near Transfer)</td>
<td>Pretest</td>
<td>0.250 (0.456)</td>
<td>0.205 (0.473)</td>
<td>0.228 (0.464)</td>
</tr>
<tr>
<td></td>
<td>Posttest</td>
<td>0.465 (0.731)</td>
<td>0.380 (0.585)</td>
<td>0.423 (0.658)</td>
</tr>
<tr>
<td>“Robot Mining” Problem (Far Transfer)</td>
<td>Posttest</td>
<td>0.725 (0.747)</td>
<td>1.235 (0.775)</td>
<td>0.980 (0.761)</td>
</tr>
</tbody>
</table>

Notes: Complex Systems Explanatory Knowledge was coded on scale of 0–2. All remaining items were coded: 0–3. Standard deviations are presented in parentheses.

---

Researchers who might be interested in further analyses, such as comparing models in a more stepwise manner, may access our raw data from the link provided in Appendix A. Supplementary data.
weather and climate declarative and explanatory knowledge from pretest to posttest, regardless of which classroom they were in, which teacher they had, or whether they learned with one or two models.

4.2. Research questions 2: learning complex systems knowledge

For RQ2, questions 6, 7, 8, 9, and 10 involved explanatory knowledge about complex systems principles. Using an average of these items, the same linear mixed-model design was run as for RQ1. The model that incorporated student nested within classroom had a better fit to the data (AICc = 107.338, BIC = 138.697) than the model without (AICc = 109.777, BIC = 138.105), however, the difference was small, and non-significant using $\chi^2$, $p > 0.200$. We decided to use the model containing the random effects analysis, as the parameter estimate of subject nested within classroom was 0.021 ($SE = 0.014$, [95% CIs = 0.0008 -0.0418]), and statistically significant using Wald's test, $p = 0.041$. This result indicates that there was variability within subjects and within the different classrooms, and subsampled the variance from our fixed effects analysis, however, compared to RQ1 the variability between the students within the different classrooms was small, highlighting the lower level of complex systems knowledge students possessed before the educational intervention.

For the fixed effects analysis, there was a large effect for Time, $F(1,92) = 66.886, p < 0.0001, \eta^2 = 0.421$, with participants of both treatment groups showing improved performance on the posttest ($M = 0.614, SD = 0.391$) compared to the pretest ($M = 0.280, SD = 0.224$). There was no main effect of Condition, $F(1,92) = 2.126, p = 0.148, \eta^2 = 0.421$, nor was there an interaction between Condition and Time, $F(1,92) = 0.273, p = 0.603, \eta^2 = 0.003$. There was, however, an effect of Teacher Content Expertise, $F(1,92) = 6.652, p = 0.011, \eta^2 = 0.07$, with students exposed to the content-expert showing higher performance ($M = 0.523, SD = 0.450$) compared to students exposed to a teacher with less expertise ($M = 0.392, SD = 0.322$). The interaction between Teacher Content Expertise and Group however, was not significant, $F(1,92) < 0.001, p = 0.985, \eta^2 < 0.001$.

These results are similar to those of RQ1 in that a significant improvement from pretest to posttest in complex systems explanatory knowledge was found for both Treatment Groups, regardless of which classroom they were in, or whether they learned with one or two models. There was, however, now an effect of Teacher, indicating that the students exposed to the teacher with expert knowledge on complex systems (Ph.D. in Physics) showed significantly higher learning of complex systems knowledge.

4.3. Research question 3: near within domain transfer problem solving

RQ3 focuses on the problem-solving performance by the two treatment groups on the butterfly effect question that was used to assess near within domain transfer. Again, the same linear mixed-model was run as for RQ1 and RQ2 for this dependent variable. The model including the random effect of subject nested within classroom had a better model fit (AICc = 342.475, BIC = 373.835) than the model without the random effect (AICc = 351.295, BIC = 379.623). The random effects covariance parameter estimates of subject nested within classroom was 0.113 ($SE = 0.037$, [95% CIs = 0.039-0.187]), and statistically significant using Wald's test, $p = 0.003$, indicating variability within subjects and within the different classrooms, and subsampling the variance from our fixed effects analysis.

For the fixed effects analysis, there was a main effect for Time, $F(1,92) = 7.585, p = 0.007, \eta^2 = 0.076$, but no main effect for Condition, $F(1,92) = 0.425, p = 0.516, \eta^2 = 0.005$, nor an interaction between Condition and Time, $F(1,92) = 0.067, p = 0.796, \eta^2 = 0.001$. There was no effect of Teacher Content Expertise, $F(1,92) = 0.730, p = 0.395, \eta^2 = 0.008$, however, the Condition by Teacher interaction was significant, $F(1,92) = 3.967, p = 0.049, \eta^2 = 0.041$. To further examine this interaction effect, we performed all possible pairwise comparisons, using Student's $t$ tests Bonferroni-corrected. There was a pairwise difference between the content-expert and the non-expert for participants in the One Model Group ($M = 0.283, SE = 0.127, t = 2.23, p = 0.028$, with students exposed to the content-expert teacher showing higher performance for the near within domain transfer. However, there was a different pattern for the Two Model group, which showed no effect of teacher expertise. Comparing all levels of the Two Models Group by Teacher Content Expertise interaction to the overall average of the model with Nelson's adjustment, revealed none of the four levels (One Model Group, Teacher A; One Model Group, Teacher B, etc.) were significantly different to the overall average across all the conditions. This indicates that the Group by Teacher Content Expertise interaction was solely due to the difference between new and experienced teachers for the group exposed to one model. Similarly, to the previous research questions, these results indicate that the students in the two treatment groups demonstrated significantly higher performance in their problem solutions on this item by the posttest regardless of whether they learned with one or two models, but interestingly, the One Model groups benefit was more dependent on a teacher with physics expertise.

4.4. Research question 4: far across domain transfer problem solving

RQ4 focuses on the problem-solving performance by the two treatment groups on the robot space mining problem that was used to assess far across domain transfer. This question only appeared on the posttest. A mixed-linear model without the repeated-measures component of Time was run. The best fitting model did not include subject nested within classroom (AICc = 228.125, BIC = 242.567), and was the best performing model out of a series of models that included and omitted Class, Subject, and Teacher Content Expertise. There was a main effect of Condition, with students in the Two Models Group who were exposed to two models significantly
outperforming students in the One Model Group, \( F(1,92) = 10.403, p < 0.001, \eta^2 = 0.11 \). However, there was no effect of Teacher, \( F(1,92) = 2.724, p = 0.102, \eta^2 = 0.029 \), nor was there an interaction between Group and Teacher, \( F(1,92) = 0.003, p = 0.958, \eta^2 < 0.001 \).

Given that the mixed-model analysis revealed that Condition was the only significant effect in the model, we sought to further examine this effect in more detail, as it implies that participants exposed to two models showed a very specific and large beneficial effect on far transfer performance. Comparing the One Model Group (\( M = 0.719, SD = 0.750 \)) to the Two Models Group (\( M = 1.231, SD = 0.777 \)) using a Bayesian independent samples \( t \)-test revealed that the data were 19.110 times (BF\(_{10} = 0.052 \)) more likely to fit the model that postulated there was a difference between both treatment groups than the model that postulated there was no difference between the groups. In Bayesian terms, this is considered strong evidence for the alternative hypothesis (Jeffreys, 1961). Examining the robustness of the Bayes factor revealed that the evidence for the difference between the treatment groups remained constant across many different Cauchy prior widths (0.25–1.5), indicating that the result was not dependent on any of the user-specified priors inputted by the researchers. In addition, sequential analyses revealed that 70 participants are required to obtain a Bayes factor indicating strong evidence for the alternative hypothesis. This is consistent with the results of a conventional frequentist independent samples \( t \)-test that revealed a significant difference between the groups, \( t_{94} = 3.234, p = 0.002 \), with a medium-to-large effect size, \( d = 0.672 \) (95% CI = 0.252–1.089). The Two Model group showed better far transfer regardless of teacher expertise. The content expert teacher’s Two Model class outperformed the One Model class, \( t_{44} = 2.17, p = 0.035 \), with a medium-to-large effect size, \( d = 0.654 \). The early career Non-Expert’s Two Model class outperformed the experienced Non-Expert’s One Model class, \( t_{44} = 2.4, p = 0.02 \), with a medium-to-large effect size, \( d = 0.693 \). It is important to note that even though the content expert built the climate models, the Two Model condition provided learning advantages in classrooms where the science teachers had no involvement with the design of the content. In summary, despite no main effects of group in RQs 1 to 3, for our final far across domain transfer question there was a large and significantly higher score for the Two Models Group compared to the One Model Group, which is also shown in Fig. 3.

5. Discussion

In this section of the paper, we first discuss the research questions to consider how and why the findings were similar to or different than our expectations. Second, we discuss theoretical and practical implications.

5.1. Answers to the research questions

The results related to RQ1—learning weather and climate knowledge—were not consistent with our expectations. There was no advantage nor disadvantage to using one or two computer models to learn the targeted weather and climate declarative knowledge, perhaps because the students were expected to have similar relevant prior knowledge in these areas to then build upon with the information provided in either treatment group. That there was no disadvantage to students in the Two Models Group is important as it suggests that these students were not confused or did not experience increased cognitive load from using two models versus one model.

For RQ2 and RQ3, the results showed no significant differences between the two treatment groups on explanatory knowledge for complex systems concepts and the near within domain transfer problem, which were not in the expected direction of higher performances by the Two Models Group students on these two types of assessment. The use of a PF design with either two models or a single model in the Generation and Exploration phase followed by the teacher led Consolidation providing instruction about the targeted complex systems concepts were equally effective for learning complex systems explanatory knowledge and preparation for solving a near within domain transfer problem. However, we note that there was not a lower performance on these two types of assessments by the Two Models students, which suggests (similar to our RQ1 analysis) that confusion or high cognitive load was not experienced by students in the Two Models Group and thus did not contribute to poorer performance in these areas.

For the last research question, RQ4, the results found the Two Models Group students performed at a significantly higher level on the far across domain transfer problem, which was in the expected direction. This result is also not consistent with a view that using two computer models would be confusing or increase cognitive load, and thus would lower the performance of the two models students compared to the single model students. In the next section, we discuss reasons as to why we had these patterns of results and what implications there might be.

5.2. Theoretical and practical implications

There are theoretical and practical implications for these findings. The main theoretical question is: Why did the Two Models Group outperform the other group on the far across domain transfer assessment? One possible explanation is that participants in the Two Models Group interacted with four additional computer models than did the participants in the One Model Group, and that the use of these additional models provided more information to the Two Models Group and conveyed a benefit for the far transfer assessment. However, we note that both groups worked on the same climate and complexity problems during the PF Exploration and Generation phases. For both treatment groups during the PF Consolidation and Knowledge Assembly phase, all teachers used the same lesson scripts to explain solutions to each of the climate and complexity problems and to discuss the key climate and complexity topics related to the problem solutions. This experimental design was intended to have the students in both treatment groups access the same amount of information related to the relevant climate and complexity topics and the solutions to the climate and complexity problems in the PF Exploration and Generation phase. If there was more information provided to the two models group (e.g., more salient representations of the targeted complexity concepts by the second model), then one would expect this group to score more highly on the complexity explanatory knowledge assessments. However, there was no difference on the posttest between the two groups on these assessment items, which suggests the “more information” explanation does not account for the far across domain transfer finding.

Another possible explanation is the NetLogo Ants Foraging model used on Day 3 by the Two Models Group had greater surface feature similarities to the Robots Mining problem than did the Climate with Feedback model used by the One Model Group. This possible explanation is based on the research of Holyoak and Koh (1987) who found that similar surface features in analogs can enhance the transfer of common structural features. However, we believe this alternative explanation is not correct given research by Resnick (1994) (who originally designed the Robots Mining transfer problem used in this study), where high school students used an earlier version of the Ants Foraging model and then were asked to solve the Robots Mining problem. He found none of the students provided solutions using the relevant complex systems concepts (e.g., self-organization, positive feedback, random behaviors) that were demonstrated in the Ants Foraging model, and instead, “the students’ strategies were almost always centralized, relying on a leader to make decisions” (Resnick, 1994, p. 128). We propose in Resnick’s study that the students did not spontaneously find surface and structural similarities between the Ants Foraging model and the Robots Mining problem as in fact the surface features are quite different, such as ontological differences (i.e., biological versus mechanical), different communication methods (i.e., chemical pheromones versus radio), size differences (i.e., small ants versus large robots), and so on. Because the Ants Foraging model and the Robots Mining problem have virtually no surface feature similarities (even though there are shared structural features based on complexity concepts), we believe the alternative explanation of shared surface features promoting far across domain transfer does not account for our findings.
Instead, we argue that the critical factor for the far across domain transfer findings in this study was the way the students learned with the two models. Kapur (2014) proposed that “prior knowledge activation and differentiation [during the Generation and Exploration phase] may afford greater opportunities for comparisons between student-generated solutions and correct solutions, thereby helping students’ attend to and better encode critical features of the new concept” (p. 1009).

In line with Kapur’s assertion, we propose two explanations for the far transfer efficacy of the PF design used by the Two Models Group. The first explanation is by using two highly contrasting computer models in different domains in the Generation and Exploration phase that the students were able to activate a greater quantity of their prior knowledge. Consequently, by activating more prior knowledge than their One Model peers, the Two Models students were better able to align their prior knowledge to—and benefit from—the ideas presented by the teacher in the Consolidation and Knowledge Assembly stage.

The second explanation, which we prefer, is about both the type and the quantity of activated prior knowledge. A key difference in the Generation and Exploration phase between the One Model and the Two Models Groups is that students in the latter had the opportunity to make model comparisons between climate and four distinct non-climate domains when they wrote out the similarities and differences between the climate and complexity models. This comparison learning activity would help the Two Models students activate more prior knowledge about structural level concepts shared by both models than would the One Model students, as well as surface level ideas that differentiate the models (Gentner et al., 2003). By activating relevant structural level prior knowledge (i.e., complex systems ideas), the Two Models students would be better “primed” than the One Model students to benefit from comparing their generated solutions to the correct solutions and structural level ideas presented by the teacher in the Consolidation and Knowledge Assembly phase.

We further illustrate our cognitive theorizing with a table that links

| Hypothesized Cognitive Processes for the One Model and Two Models Groups for Day 3 Mapped to Transfer Problems. |
|---|---|---|
| Note: In columns three, four, and six, an “*” represents a concept an idealized student has that is relevant to the problem and the available model. An “x” represents a concept that is relevant to the problem and the available model, but that the student does not know. In column five, the “+” indicates the concepts the teacher covered in the PF Consolidation phase. In columns seven and eight, “NA” indicates that a complexity of climate concept is not applicable for the correct transfer problem solution. The grey shading of the “*” is provided to highlight the different pattern of correct concepts the students in the two groups the students demonstrated in the daily activities and on the posttest transfer problems. |

### Table 9

#### One Model Group

<table>
<thead>
<tr>
<th>Cognitive Process</th>
<th>Activate Prior Knowledge (Incomplete Schema)</th>
<th>Schema Abstraction</th>
<th>Within Domain Transfer</th>
<th>Across Domain Transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activities</td>
<td>Complexity Problem 1</td>
<td>Climate Problem 2</td>
<td>Teacher Instruction</td>
<td>Climate Problem 3</td>
</tr>
<tr>
<td>Models</td>
<td>Climate Model</td>
<td>Climate Model</td>
<td>Climate Model</td>
<td></td>
</tr>
</tbody>
</table>

#### Complexity Concepts

1. Inputs
2. Outputs
3. Positive Feedback
4. Negative Feedback
5. Organization
6. Tipping Points
7. Non-linear

#### Climate Concepts

1. Temperature
2. Effect
3. Cloud Cover
4. Atmospheric Water
5. Feedback
6. Albedo

(continued on next page)
more directly to the details of the study and a figure that is a more general schematic representation. Table 9 describes the hypothesized relevant cognitive processes, their alignment with instructional features, and the predicted learning outcomes, using Day 3 concepts as examples (see Table 1). The relevant climate and complexity concepts are listed for the complexity and climate problems the students worked on and correct solutions for the near within domain and far across domain transfer items on the posttest. This difference between the One Model and the Two Models Groups is shown in Table 9, where an “*” represents a concept an idealized student has that is relevant to the problem and the available model, an “x” represents a concept that is relevant to the problem and the available model, but that the student does not know, and a “+” indicates the concepts the teacher covered in the PF Consolidation phase. We propose participants in the Two Models Group activated more relevant structural level prior knowledge (i.e., Inputs, Outputs, and Negative Feedback) in the Exploration and Generation phase than the One Model Group because of using the second complexity model and the scaffolding prompt to compare the two models, a technique the previous research demonstrated can help with schema abstraction and transfer (Alfieri et al., 2013). Since both groups used the same climate models, we assume similar climate concepts would be activated from the students’ prior knowledge (Global Temperature and Cloud Cover), but not all relevant concepts. The teacher’s instruction then covered all relevant complexity and climate concepts. PF theorizing has proposed that during Consolidation and Knowledge Assembly phase, the students would compare their concepts to the concepts provided by the teacher, and that this would elicit the cognitive process of schema abstraction. We suggest that both groups benefitted by comparison of their concepts in the Generation and Exploration phase with the teacher’s concepts, as shown in Table 9 where
both groups incorporate three additional concepts about climate to their initial two ideas, to have a more complete understanding of the five targeted climate concepts and to be able to use an appropriate concept (Cloud Cover) on the near within domain transfer problem. However, because the Two Models group had activated three complexity concepts during the Generation and Exploration phase, they would be able to construct a more complete generalized schema about complexity with six of the seven targeted concepts for use on both the near and far transfer problems.

As another way to illustrate our theorizing, we provide a schematic representation of the PF and analogical comparison processes in Fig. 4, where the circles represent surface level knowledge about weather and climate and the stars represent structural complexity knowledge relevant for weather and climate phenomena as well as other phenomena. The correct teacher knowledge about climate has the circles that form a triangle, and the correct complexity ideas are stars that form a rectangle (implied by the positioning of the circles and stars to minimize visual clutter). On the left side of the figure, the One Model Group has prior correct knowledge of one star and two circles generated in the Exploration and Generation phase. On the lower left side of the figure, through schema abstraction activated during the Consolidation and Knowledge Assembly phase, the One Model Group has generated a partial complexity structural features schema with two stars that forms a line (i.e., an incomplete rectangle) and a complete climate surface features schema, with three circles linked as a triangle. The line and triangle are linked by a solid line to show the partial structural knowledge is tightly connected to the surface features, and thus would be available for near within domain transfer (i.e., butterfly effect problem) but not far across domain transfer for problem solving (i.e., Robots Mining problem). In contrast, on the right side of the screen, the comparison activity of Two Models Group has activated more relevant prior knowledge consisting of two stars and two circles in the Exploration and Generation phase. The lower right side of the figure shows through schema abstraction in the Consolidation and Knowledge Assembly phase that the Two Models Group has constructed a complete complexity schema with four stars forming a rectangle and a complete climate surface features schema shown by the triangle. The rectangle and triangle are linked by a dotted line to show the complexity structural knowledge is a generalized schema that is loosely connected to the surface features of the climate knowledge, and thus would be flexibly available for both near within domain and far across domain problem solving.

Also, we propose that the Two Models Group would have undergone two phases of schema abstraction: first, when they compared the two models, and second, when they compared their ideas to those of the teacher. As a result of this first phase (shown in Table 9 and Fig. 4), the Two Models Group would have generated more relevant structural level prior knowledge and thus were able to compare and contrast these additional structural level ideas with those of the teacher for a second, deeper round of schema abstraction and consolidation. We propose that this “double round” of schema abstraction helps explain the differences between the two treatment groups on the far across domain transfer task.

From a practical perspective, we note there is a challenge for instructional approaches that require comparisons as research has found learners often do not make appropriate comparisons between two cases even if they are presented temporally close together (Alfieri et al., 2013). One approach researchers have explored to overcome this challenge is to present students with minimally different cases that have some degree of overlapping surface features in order to increase the probability that learners “spontaneously draw appropriate comparisons.” There has been some success with this approach, such as with progressive alignment where subsequent comparisons could be made between increasingly more dissimilar cases, thus helping with the construction of a progressively more abstract schema (e.g., Braithwaite & Goldstone, 2015; Kots LVsky & Gentner, 1996; Thompson & Opfer, 2010). With the progressive alignment approach, the surface features and deeper structure are correlated in minimally different cases, and then the common surface features help learners recognize that there is similarity, which draws the learners to compare and then to notice the deeper structural commonalities.

In contrast, an important finding in this study is that the hybrid PF learning design insured successful comparisons using maximally different cases in the Two Models Group that did not rely on spontaneous comparison. Our reading of the comparison literature is that this has not been previously demonstrated. First, as a scaffold, students wrote in their Student Guidebooks about similarities and differences as part of the Idea Generation and Exploration phase. Second, in the Consolidation and Knowledge Assembly phase, the teacher explicitly provided instruction about the deep structure for the comparisons, which were the targeted complex systems concepts. The hybrid PF learning design approach is in line with the Schwartz et al. (2011) research, which found when the exposition of an expert-like solution happens after students have explored novel deep structures between two contrasting cases, they are more likely to be able to transfer the underlying structure to the new targeted problem.

We believe that additional research is warranted to explore the integration of analogical comparison with PF learning designs in terms of their effectiveness for helping students learning difficult scientific knowledge such as was the content focus in this study. As a practical advantage, we suspect that the use of the hybrid PF learning design with highly contrasting cases will lead to successful learning outcomes in a shorter period of time than approaches with a greater number of cases as used in progressive alignment learning designs, although research is needed to study this issue. Given limitations of a classroom quasi-experimental design, we also recommend future experimental research could systematically vary treatment conditions in order to better “unpack” the theoretical mechanisms that underlie across domain knowledge transfer and to perhaps refine theories of PF and analogical comparison for near within and far across domain transfer. One such study could do a detailed analysis of student responses to the challenge problems in the PF Generation and Exploration phase and the post-Consolidation and Knowledge Assembly challenge problems to see if the empirical data supports our proposed theoretical model discussed for Table 9 and Fig. 4. Other research could explore instructional issues, such as the use of one or two computer models with a PF Generation and Exploration phase and Consolidation and Knowledge Assembly phase versus the use of the same Generation and Exploration phase but without a Consolidation and Knowledge Assembly phase. Further research also could investigate a possible “preparations for future learning” type of transfer with the PF learning designs.
Another practical implication is that this classroom study shows there is efficacy for using a PF learning design to have students learn advanced scientific ideas about weather, climate, and complex systems using agent-based computer models. This is important given the need for students to understand conceptually challenging scientific ideas such as those in the United States Next Generation Science Standards. The hybrid PF design for learning about complexity and climate change also seems to help solve the practical instructional issue of the sequencing of topics. That is, rather than needing to determine if teaching complexity ideas should happen before or after climate change ideas, our findings suggest that concepts about both areas can essentially be learned simultaneously, and presumably requiring less instructional time as well.

It is also a nice finding that successful learning outcomes were not due to the content-expertise or experience levels of the teachers. Indeed, the Two Models Group class with the early career non-expert teacher outperformed the more experienced non-expert teacher in the One Model Group on the far across domain transfer problem. Also, the effect sizes were virtually identical for the advantage of the Two Model group between the two classes of the content expert and between the two non-experts. This suggests the PF Two Models learning design and the teacher lesson plans we provided are likely to be viable in regular classroom settings where there is often variability in the background knowledge and experience of science teachers. A further practical implication of using a PF design with two contrasting computer models is that teachers could achieve the more typical learning goals of efficiency of learning and retention as well as helping students to learn for transfer and to use their knowledge in new situations and contexts.

We note certain limitations to this study. As a quasi-experimental study, there are limits to internal validity, and participants could not be randomly assigned to groups so that convenience sampling was employed. Also, our conclusions are provisional given they are based only on assessments of learning about complex systems and climate change that had been used in previously published research into learning about complex systems (Jacobson et al., 2017). We acknowledge that the reliability indices in this study and the low and skewed scores for certain posttest measures are perhaps insufficient for the development of a psychometric instrument. Still, based on the results of this study we believe further research is warranted in at least three main areas: (a) the development of a psychometrically validated instrument for assessing declarative and explanatory knowledge about complex systems concepts, and near and far transfer of complex systems knowledge, (b) the need for research in a larger number of schools and with a wider range of students in terms of their academic performance, and (c) using PF designs for learning other science topics with and without computer models.

6. Conclusion

Overall, we propose that there are three main novel contributions of this research. First, this study involved a productive failure learning design that investigated learning for transfer from computer models in a single domain versus from comparisons across two contrasting computer models representing different domains. Second, we articulate a theoretical integration of the cognitive processes that have been previously proposed for productive failure and analogical comparison. Third, this study replicates and extends earlier research involving the use of a productive failure learning design with computational model-based learning in real classrooms (Jacobson et al., 2015).

In concluding this article, we hope the results of this study suggest viable practical approaches to help students better learn challenging scientific knowledge and skills that are now being required, such as in the United States Next Generation Science Standards (National Research, 2012) and the Australian Curriculum – Science (Acara, 2013). Also, we hope teachers will welcome innovative instructional approaches such as we have explored in this study, both because students were found to learn and to be able to transfer challenging scientific knowledge and because (based on our qualitative classroom observations) the students were very engaged and enjoyed using the computer models to learn with. Finally, we hope this research will stimulate further interest in theoretical and research based understandings and approaches for achieving the McKeough et al. (1995) goal of solving “one of teaching’s most formidable problems,” that is the transfer of learning.

Acknowledgement

The research discussed in this article has been funded in part by a grant to the first author from the Australian Research Council (ARC) Discovery program DP150102144, an ARC grant DP150104267 to the first and second authors, and grants to the first and third authors from the ARC Linkage program, LP100100594 and the New South Wales Department of Education and Communities.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.learninstruc.2019.101222.

References


Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. Journal of Educational Psychology, 95(2), 393–408.


