Individualized Scaffolding of Scientific Reasoning Development – Complementing Teachers with an Auto-agent

Toi Sin Arvidsson

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

ABSTRACT

Individualized Scaffolding of Scientific Reasoning Development – Complementing Teachers with an Auto-agent

Toi Sin Arvidsson

Building on the success in a previous study in engaging the underserved middle-school population in the practice of science through individualized scaffolding, the current study sought to examine an automated-agent, Astro-world, developed to provide real-time guidance to students in order to increase the scalability of the intervention while maintaining the benefits of the individualized format. Through practices of argument and counterargument in advancing and challenging claims, the agent focused on coordination of multiple variables affecting an outcome, rather than only the foundational and more extensively studied strategy of controlled experimentation, in the context of a scenario in which students had to investigate multiple factors affecting the performance of potential astronauts in a space simulator. The intervention sought to help students see the purpose and value of scientific practices using social science content rather than traditional science topics. In addition to adapting the technology into a regular classroom setting in which the teacher is actively engaged (teacher-involved condition), the study included a second condition to determine if the technology could be used effectively without active teacher involvement (tech-only condition). Delayed far-transfer assessments showed that only students in the teacher-involved condition (but not the tech-only condition) outperformed those in a non-participating control group in recognizing the need for evidence and considering all contributing factors in making predictions. Furthermore, post-hoc analysis showed that these significant differences occurred predominantly among those who mastered the foundational control of variable skills. Possibilities are considered as to why teacher involvement was critical to effectiveness, and implications for classroom practice are addressed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

Besides all the privileges I was born with, which made this milestone attainable, I thank my family who shaped me to become the person I am to pursue my dream: Louisa, who let me tag along her courageous and brilliant self and taught me so much since I was ten; My father, who instills in me that no dream is too far to reach; My mother, who cares so much about my well being that nothing seems to matter more; My little brother, who showed me love when he was nine and I was a heartless teenager.

Furthermore, this journey has been made so precious and enjoyable because of so many inspiring individuals along the way: Stephanie, who is always there for me and who teaches me how to be a scientific mother. Bonnie, who makes me laugh. Nip, who tolerates my much needed check-ins. Valerie, Laurie, Julia, Sweet, Shimin, Mia, Jacklyn, and Luc, who are the perfect companions for entertainment and hard work. Christina and Norm, who are wonderful collaborators. Professor Baker, Miller, Cookston, Lynch, Louie, and Paik who are so stimulating in their teaching and so generous with their advice and time. My committee members, Professor Peverly, Chi, Baker, Keller, & Kuhn who challenged me and whose constructive feedbacks made this thesis better. And all my other friends and families who are too many to name, who I am touched by each of their individual support and love to sustain my life journey.

I also found myself to be one of the luckiest doctoral students to have Professor Kuhn as my advisor, who is not only inspirational in her work and writing but also cares so much both about our overall well-being and our long term success. She also seems to know the exact timing for lighting a fire under us. Thank you, Deanna.

And to my partner in crime—my dear husband, whose ability to grow challenges me to become a better person, I cannot imagine how one can endure me day and night and putting up

with every laughter, scream and tear drop through these years. Together with our beloved daughter, whose curiosity fascinates me and gives me life examples of my own learning, you two keep me going.

I thank you all and am so grateful that you all are in my life.

Chapter I: Introduction

The goals of science education have evolved over time. Besides learning content knowledge, it is widely agreed now that an important goal is for students to understand the value of science as collaborative practices that hypothesize, make inquiries, and debate as a way of constructing knowledge (Osborne, 2014; Sandoval, 2014; Lehrer & Schauble, 2015). Therefore, such education should provide students the experience of science as a practice and one that goes beyond procedural knowledge of conducting experiments.

Such an experience involves epistemological growth in differentiating beliefs from evidence-supported claims – therefore growth in metacognition, including knowing what is not known. Such metacognition lays the way for developing the disposition to engage in investigation and experiment. As students develop strategies to experiment and interpret data, argumentation skills help them see the need to examine one's statements and address opposing claims, which is crucial for advancing dialogues leading to knowledge construction. In sum, metacognition, strategies for experiments and interpreting data, as well as argumentation skills, are all essential parts of science and hence of learning to do science.

In two recent studies (Arvidsson & Kuhn, 2016; Kuhn, Arvidsson, Lesperance, & Corprew, 2017), the author and team implemented an intervention intended to provide such a learning environment, with a particular goal of developing students to be multivariable thinkers. The intervention was problem-based investigating factors that predicted astronaut performance. Besides engaging students in scientific practice, a scaffolding protocol was designed to guide students by questioning without ever giving them explicit instruction regarding strategies. These scaffolds were intended to challenge students cognitively, while providing metacognitive supports. The study showed that, in a delayed-far-transfer posttest, participating students outperformed non-participants in both experimentation and multivariable thinking skills for middle school and high school students.

In addition, for middle-school students, comparing two instructional conditions – classroom versus individualized – we found students in the individualized setting performed better in the posttest, even though they spent less time overall in the intervention (an average of six 24-min sessions) than those in the classroom setting (10 45-min sessions). In the individualized condition, a human facilitator guided one pair of students at a time through all the activities; in the classroom condition, a teacher led the whole class through the activities, with students working in pairs. In both conditions, the same scaffolding protocol and activities were used, except that in the classroom setting, discussions were held as whole-class discussions and students engaged in memo-writing activities rather than the individual discussion between student pair and adult facilitator that took place in the individualized condition.

The better results in the individualized condition, despite lesser time investment, were not entirely surprising. Students benefit from personalized learning in many content areas and contexts, but personalization may be even more crucial for a science learning experience, given the range of skills needing to develop in concert. The inability to scaffold each student within their zone of proximal development constitutes missed opportunities. Yet, personalized learning is difficult at best in classrooms. Besides personalization, the limited success in direct instruction of experimental skills among lower-performing students suggests that the type of scaffolding may also matter (Lorch, Lorch, Calderhead et al., 2010).

The results of the Arvidsson and Kuhn (2016) study suggest that further examination is warranted regarding how best to integrate teacher-led activities with technology to provide personalized scaffolding that is both successful and scalable. Here, we address the question of what combination of instructional activities is optimal if science education is to achieve its goals of developing deep understandings of science process as well as content. We thus devote the present study to further examination of the elements in the recent intervention that seems to be essential or most important. In particular we investigate how the desired learning objectives can be achieved while being scalable through use of an automated-agent.

Chapter II: Literature Review

What Needs to Develop?

Epistemological growth

Epistemological beliefs are mainly defined as ones' beliefs about the nature of knowledge and the nature of knowing (Hofer & Pintrich, 1997). Those who hold a more mature view that knowledge is not merely a collection of facts but rather is complex and evolving are more likely to value knowledge as a result of evaluation of evidence rather than authority or unexamined beliefs.

The developmental trajectory of epistemological growth begins with young children holding a naïve understanding of knowledge, believing that what they know are truths (Hofer, 2008). Kuhn and Dean (2004) called them *realists*, who do not question whether there are differences between reality and what is asserted by someone to be reality. As children grow older, they become *dualists* or *absolutists*, who are aware of possible conflicting ideas and they develop the idea of universal objective truths. Yet they regard knowing as a collection of facts, most often provided by direct observation or by authority figures (Hofer, 2008). Gradually, most come to recognize subjectivities but while they acknowledge the existence of multiple views, there is a lack of differentiation among the quality of these views. They are *multiplists*. Some but not all, however, eventually develop into *evaluativists* who appreciate that certain views are more valid than others and that some information presented as fact is more reliable than other. Such an appreciation of the need for evaluation of ideas is fundamental to the practice of science and therefore an important goal of science education. However, unlike some of the developmental progressions in which children eventually reach the most advanced stage, not all adults are evaluativists; therefore, education is needed to bring about such development (Hofer, 2001; Kuhn & Dean, 2004).

Researchers have observed how differences in epistemological beliefs predict differences in how students learn. Stathopoulou and Vosniadou (2007) argued that when faced with information that does not align with existing knowledge, one's epistemological beliefs may influence how the person learn at two levels: Those who view knowledge as a collection of unchanged facts may be more likely to dismiss the new information as incorrect (Mason, 2003). In addition, because of such epistemological beliefs, their goal for knowledge acquisition of collecting unchanged facts provides little motivation in seeking resolution for conflicts between new information and existing knowledge (Sinatra, 2005). Similarly, Fazey (2010) also suggested that students will be more motivated to seek out deeper understanding of knowledge and to better understand and resolve conflicting information if they believe that knowledge is not just facts that are certain and unchanged.

Previous research found that students' epistemological beliefs about knowledge influence their learning approach as well as predicting academic performance and that those with more advanced beliefs tend to have more reflective thinking (Phan, 2008). Students who have the more mature belief that knowledge is complex and not absolute approach learning more deeply, presumably due to their awareness that knowledge does not constitute discrete facts that can be simply memorized (Cano, 2005).

While more advanced epistemological beliefs seem to predict better learning strategies, not all people develop them (Hofer & Pintrich, 1997). Furthermore, amount of education one has does not necessarily predict more sophisticated beliefs (Tsui, 1999). At the same time, mature understanding that knowledge evolves and depends on evidence was found less likely to change over a short period of time even after multi-session science instruction that included hands-on activities of collecting data and drawing conclusions based on observations (Conley et al., 2004). Science major students presumably have more content knowledge about science but do not necessarily have higher-order critical thinking skills. Students who recognize science knowledge not as fixed facts were more likely to recognize complexities and evaluate contradictory claims (Liu, Lin, & Tsai, 2010). However, more topic knowledge is not necessarily an indication of more advanced epistemological beliefs; the relationship is complex and relies on the context of the judgment about the knowledge (Bromme, Kienhues, & Stahl, 2008). Moreover, among students within the same setting, SES and achievement level did not predict changes in students' epistemological thinking, suggesting that students of all levels may face similar challenges (Conley et al., 2004).

On the other hand, studies have shown that over a period of regular science instruction with hands-on experience, students relied less on authority and more on observations and expressed greater uncertainty about there existing one right answer in science (Conley, Pintrich, Vekiri et al, 2004). Focus on observation and exploration may have aided students in moving away from the reliance on authority; moreover, more focus on argument and reflection may also help students see that knowledge evolves and that evidence is needed to justify conclusions (Conley et al., 2004).

Argumentation skills thus may also be closely associated with one's epistemological thinking. Mason and Scirica (2006) found epistemological thinking to predict better argumentation skills in middle-school students. Felton and Kuhn (2007) believe that argumentation supports the development of more sophisticated thinking because it highlights disagreements in interpretation between well-respected figures, which indirectly show that knowledge cannot be treated as simplistic list of facts. Argumentation also helps make ones' thinking explicit—in defending one's points when being challenged; those who take part in the debate experience the process of evaluating evidence, examining assumption and drawing conclusions, just as someone with sophisticated epistemological thinking would do on their own.

There is evidence that presenting refutational text influences progression to more mature epistemological understanding. However, the opposite also has been reported of a decline (Kienhues, Bromme, & Stahl, 2008), suggesting such belief to be more fluid and less stable. In reality, one rarely relies on first-hand evaluation of evidence in judgments of the reliability of knowledge but rather has some ways of evaluating the trustworthiness of the source of information (Bromme, Kienhues, & Stahl, 2008). Nevertheless, in the context of teaching science as a process, it is an important goal to instill a more sophisticated sense of epistemological belief in realizing that construction of knowledge ultimately relies on evidence; while knowledge is not absolute, some interpretations may be more valid than others.

Metacognitive development

Epistemological beliefs and metacognition are closely related constructs. Some consider sophisticated epistemological beliefs as a motivation for a process of metacognitive monitoring to reach a goal (Bromme, Pieschl, & Stahl, 2010)—if one believes that justification of how evidence is evaluated to construct knowledge is needed, they might be more likely to set the goals for seeking out the evidence and then metacognitively monitor the execution of the plan and the strategies for evaluating the evidence to achieve such a goal.

In the past, metacognition, self-regulation and self-regulated learning have sometimes been used to refer to similar processes, while at other times they are seen as a larger construct that includes the others (Garrisona & Akyolb, 2013; Dinsmore, Alexander, & Loughlin, 2008; Lai, 2011). For example, some researchers consider self-regulated learning to include different processes involving learning and consider metacognition to be one of them (Loyens, Magda, & Rikers, 2008; Schraw, Crippen, & Hartley, 2006) and that metacognition is specifically the process of monitoring during learning (Winters, Greene, & Costich, 2008). Others consider metacognition and self-regulation to be unique constructs that are associated with each other (Fox & Roconscente, 2008). Pintrich, Wolters, and Baxter (2012), on the other hand, define metacognition to include a more static component of metacognitive knowledge, such as what one knows about certain declarative knowledge, along with the more dynamic processes of metacognitive control and regulation, which involve active judgment and control. Here, we consider metacognition to concern one's awareness of one's cognitive processes, allowing one to seek to control these processes (Livingston, 2003). Components of metacognition include judging, regulating and monitoring, which actively evaluate how well one is aware of one's knowledge (Pintrich, Wolters, & Baxter, 2012).

Most researchers agree that metacognition develops naturally as children mature, at least to some extent (Kuhn & Dean, 2004), without training (see Lai, 2011 for review). At the same time, research shows that, education can be beneficial to advance such development (Lai, 2011). For example, Sungur and Tekkaya (2006) found that Problem-Based-Learning, a learner-centered teaching method that empowers students to learn through developing successful solutions for a well-defined problem (Savery, 2015), appears to help students achieve higher levels of metacognition.

In the realm of problem solving and constructing strategies to achieve a goal, just as metacognition is at work when one constructs declarative knowledge, a similar type of judging and monitoring process may help students monitor and evaluate their ability to effectively design and implement their strategies. Students having better metacognition will successfully keep track of their goals while they form strategies to achieve those goals. In addition, once they have devised a strategy, they will evaluate whether the strategy is likely to achieve those goals. Less successful students may lose track of what they intended to achieve; or once they have arrived at a strategy, they may not realize the need to evaluate whether the strategy will actually work.

In the previous problem-based intervention (Arvidsson & Kuhn, 2016) mentioned above, a student's ability to self-monitor was a bigger challenge in the classroom setting compared to the individualized setting. The intervention relies on student pairs to be self-directed learners while the teacher manages the entire classroom. A literature on self-regulated learning notes how the success of such learning goes hand in hand with facilitating students' metacognition (Zimmerman, 2002).

While metacognition is important, it does not work in a vacuum. As students attempt to solve problems, their knowledge and cognitive abilities in the given domain also limit how well they are able to address the problem. In a domain new to students, they may not have sufficient knowledge to fully inspect all aspects of the problems they are trying to solve when they evaluate their strategies against their goals. A strategy might appear to be successful because students fail to see weaknesses of the strategy. In other words, students may have the metacognitive resources to monitor their goals and execution of the strategy they constructed but lack the cognitive skills with regard to the problem they are trying to solve to allow them to come to an accurate conclusion regarding how successful their strategies are in this case.

Consider an imaginary student who was given a problem of determining predictive relationships between multiple factors (say hours of television watching, family income, and age) and an outcome variable (say GPA). Let's say the student chose a flawed strategy of evaluating only a single data point of one case and that the record showed a low level of television-watching and a high GPA. When asked what she found out, a student who has weak metacognition might have lost track of the original goal of figuring out the causal relationships and instead become distracted and answer, “The student has a high GPA.” On the other hand, if she metacognitively keeps track of her plan, she might follow through with her strategy using the single data point and draw the conclusion that hours of watching televisions made a difference to the GPA. When asked how the conclusion was reached, she might say it was because when television-watching was low, the GPA was high. At this point, the student metacognitively kept track of her goal and planned to figure out the causal relationship between one factor and the outcome variable and drew conclusions related to that goal. However, she lacked the cognitive ability to recognize that her inference strategy and conclusion were both flawed. Scaffolding in this scenario may help the student see the aspects of the goal that she had not achieved (due to lack of comparison and therefore insufficient data), leading her to realize that a modification of her strategy was needed. In this way, developments in metacognitive and cognitive skills go hand in hand.

Collaborative learning

Collaborative learning is associated with social benefits such as receiving social support and enhanced motivation (Wentzel & Watkins, 2002). Some studies even found students to feel less anxiety when working with other students (Laal & Ghodsi, 2012). Furthermore, engaging in argumentation and communicating ideas are core elements of science practice (Osborne, 2014; Lehrer & Schauble, 2015). Scientific achievements are rarely products of a sole scientist working alone; working collaboratively with another student allows students to experience discussing their thoughts and hearing about and considering others’ thinking before coming to conclusions.

Besides creating a more realistic scientific practice experience, collaborative learning may also allow students to support each other metacognitively because metacognition as a cognitive process is both individual and social (Tarricone, 2011). For example, in collaborative projects, Winters, Greene, and Costich (2008) found students to help regulate each other while Hennessey (1999) found students to develop metacognitively. As a result of interacting with each other, one might become more aware about their cognitive process than without such interaction (Volet, Vauras, & Salonen, 2009) possibly because of the opportunity of encountering other perspectives in reflecting on the problem at hand (Lajoie & Lu, 2012). Researchers argue that when working in a group, besides invoking metacognition individually within each member, metacognition is "shared" socially, such as by monitoring each other as they work on a problem together with a common goal (Iiskala, Vauras, Lehtinen, & Salonen, 2011).

While collaborative learning has many potential benefits, learning together with another student is most effective if students are genuinely interacting with each other to co-construct knowledge (Chi & Wylie, 2014). Meaningful interactions involve asking good questions and providing feedback. However, that is often challenging for students especially if they are new to the domain. Even among older students, implementing and improving collaborative skills is difficult. Passively providing generic guidelines for how students should interact with each other might not be sufficient in helping students ask good questions, even if it helps them ask more questions (Choi, Land, & Turgeon, 2005).

In recent years, technology assisted learning research has made use of artificial intelligence to provide supports that encourage quality peer interactions between students interacting online (Magnisalis, Demetriadis, & Karakostas, 2011). Short of elaborate integration of artificial intelligence, other work shows success when students are first trained and then practice rules of collaboration – for example, Respect, Intelligent collaboration, Deciding together, and Encouragement (RIDE; Saab, van Joolingen, & van Hout-Wolters, 2011). Saab et al. found students were more likely to plan and strategize as a team if they were first trained on RIDE and then participated in activities online that prompted students to attend to those rules, especially if their inquiry activities began with asking students to hypothesize together. The key to these prior successes might be in prompting students to collaborate in real time and reminding them how to effectively collaborate.

Developing Multivariable Thinkers as a Key Objective

The intervention reported here sought to support students in becoming multivariable thinkers. Science education and cognitive development research involving inquiry learning until recently have focused largely on students' ability to accurately design informative experiments that allow them to identify relations between two variables (Lehrer & Schauble, 2006, 2015; Zimmerman, 2007). To be successful, students must understand the need to keep all other variables controlled while manipulating only the investigated variable and observing changes in the outcome variable—the control of variables skill (COV). COV is one of the key fundamental skills in experimental design emphasized in the Next Generation Science Standards (2013, Appendix F p.7 & 8): "Consider possible confounding variables or effects and evaluate the investigation's design to ensure variables are controlled… Make directional hypotheses that specify what happens to a dependent variable when an independent variable is manipulated." In recent years, various technology-based interventions have had varying degrees of success in helping students develop this skill (Dieterle & Clarke, 2007; Gobert, Sao Pedro, & Baker et al., 2012; Kuhn & Pease, 2008).

Eliminating confounding variables is an initial and essential step in determining whether a relation between two variables exists. However, practicing scientists commonly move beyond a univariable framework of one independent and one dependent variable to examine multivariable relations. Whether in physical science or social science, those who seek to explain phenomena in the real world must take into consideration multiple factors working together in contributing to an outcome (Brusatte, Butler, & Barrett et al., 2015; Zimring, 2006), especially if the goal is to make accurate predictions of future events that can lead to effective policies and practices. For example, having a firm grasp on multiple contributing factors may impact how practitioners treat patients (Phelana, Burgess, Burke et al., 2015). While not yet achieving widespread recognition among science educators, coordinating multiple variables is arguably a higher-order cognitive skill critical to the development of scientific thinking (Kuhn, 2016; Grotzer et al.).

Multivariable thinking, i.e., recognizing that multiple causes contribute to an outcome, both extends beyond the domain of science to everyday thinking and is more challenging than one might expect. Among the general adult population, only a small percentage are able to consistently take into account the multiple known contributors in making predictions (Kuhn, Ramsey, & Arvidsson, 2015). Many appeared to believe that a single cause is sufficient in explaining any outcome. Another source of difficulty is the belief bias that is implicit in most people's thinking. Given the prevalence of belief bias in considering social problems, such as poverty (Cozzarelli, Wilkinson, & Tagler, 2011), beyond understanding science practice, in this age of internet access to complex information and arguments, being able to set aside one's belief bias and consider the complexity of multiple variables having additive or interactive effects on an outcome is an essential 21st century skill, one having the potential to enable the next generation to better fulfill their civic duties in influencing political policy decisions.

Personalized Scaffolding Environment

While students in the classroom setting performed better than a business-as-usual control group in our previous study (Arvidsson & Kuhn, 2016), the students in the individualized setting performed even better, despite spending only about one third of the time on average on the intervention. Essentially, students who received the scaffolding individually throughout the intervention performed better than those who spent more time on the activities but received the scaffolding only through memo-writing and whole-class discussions. All of this suggests that the type, the timing and the amount of scaffolding may all affect students' performance.

Besides easier on-task management, the individualized condition used in the 2016 has the advantage of scaffolding students at the precise, most "teachable" moments, when they constructed strategies, executed them, and drew conclusions. While students in the traditional classroom setting experienced similar challenges through memo-writing and whole-class discussion, the metacognitive scaffolding and feedback they received may not have been of sufficient frequency and/or timed optimally, as students were developing their skills.

Scaffolding students at the moment when metacognitive judging and monitoring process are needed to direct students towards their goals may be effective in helping those who have yet developed stable metacognitive monitoring and control. Scaffolding tied to students' behavior has long been found effective in acquiring declarative knowledge, possibly because it helps facilitate students to self-regulate (Azevedo & Hadwin, 2005). It appears that scaffolding that helps students monitor how well they are learning and evaluate their understanding could be one of the key elements in the success of scaffolding. Moreover, providing scaffolding at just the right learning moment may be equally as important (Hadwin, Wozney, & Pontin, 2005).

Studies on acquiring declarative knowledge show that scaffolding serves to introduce metacognitive tools to help students strengthen the phases of learning—setting goals, executing learning plan, and evaluating their levels of learning (Hadwin, Wozney, & Pontin, 2005; Puntambekar & Stylianou, 2005). Similar mechanisms may be relevant in the case of students developing cognitive skills, rather than only declarative knowledge. When students are constructing strategies to solve problems, scaffolding based on their inputs may help them evaluate how well the strategies will work. However, providing direct metacognitive supports in learning software has been limited and when it happens, the supports may not be specific enough for students to act on (Winne, Hadwin, & Perry, 2013). To address this issue, the protocol in the intervention studied here was designed to scaffold students both cognitively and metacognitively, primarily through questioning.

Blank (2000) found that when metacognitive thinking was emphasized by asking students to reflect on their progress, students showed better concept learning. A similar mechanism may help students construct better strategies. As students' claims are challenged, they are implicitly asked to reflect on their thinking, including the strategy used and the validity of their conclusions from the data selected based on the strategy. The scaffolding in the present intervention is designed to encourage reflections through questioning.

For example, when students falsely draw causal conclusion based on one case of data, the protocol in our intervention asks, "What would happen to the outcome if the factor changes from high to low (or low to high depending on the case selected)?" (Arvidsson & Kuhn, 2016). Here, the scaffolding does not directly point out the error in the student's conclusion; instead, it draws attention to the implication of the student's conclusion and gives the student a chance to re-examine their claim. In order to answer this question, students with better developed metacognition may re-evaluate their claim and realize that they lack sufficient data to answer the question, whereas others may simply respond based on their claim and say that the outcome would also have a lower (or higher, depending on their claims) level. Following such a response, the protocol further scaffolds the students by asking, "Are you sure? Why don't you find out?"

Here, similar to the metacognitive judgment of confidence about one's declarative knowledge, such metacognition helps students evaluate the level of confidence that they have about their strategy. By suggesting students to find out, the scaffold explicitly helps students regulate by suggesting that they confirm whether their claim was correct. Following such a suggestion, students look for a new data point to determine if they were correct. As they do that, students who metacognitively kept track of the original goal of identifying a relationship between the investigated factor and the outcome variable may recognize the usefulness of finding a second data point and re-evaluate their conclusions based on the two data points.

Once students achieve the more advanced strategy of making comparisons, they may lack the awareness or cognitive skill to recognize the need for controlling other factors. More scaffolding will accordingly be needed. Once students construct and apply the optimal strategy and come to correct conclusions, the protocol further scaffolds by challenging students with contradicting claims. These questions challenge students to re-evaluate their solutions from another point of view. Those who have achieved sufficient mastery should be able to point to the weakness of the contradictory claim and explain why the claim is incorrect. Those who have not would require more scaffolding.

As students learn to think in terms of a multivariable model, in which multiple factors working in conjunction with each other to influence an outcome, they must relinquish incorrect mental models, most importantly one that assumes a casual relationship based on the mere co-existence of certain levels of two variables. Constructing mental models is critical for students to be able to effectively monitor their cognition (Hogan & Thomas, 2001). Cognitive disequilibrium can be helpful for triggering and solidifying conceptual changes as students give up incorrect mental models and adopt new models (Chinn & Brewer, 1993; Gunstone &

Mitchell, 1998; Pintrich, Marx, & Boyle, 1993). Similarly, by allowing students opportunities to construct strategies without direct instruction, they are afforded the opportunity of experiencing cognitive disequilibrium, as they are then challenged to realize the insufficiencies of their strategies. We expect this process to encourage deeper growth. Just as the development of an autonomous learner encompasses a range of necessary skills working together (Schraw, Crippen & Hartley, 2006), the development of a scientific thinker, who sees the value of science practices as ways of constructing knowledge and has the skills to do so, requires a range of skills to be developed in concert. By asking questions in real time that draw students' attention to meaningful evaluation towards their goals, challenging students to see various implications of their own claims, and providing counterarguments that students must think hard to refute, the scaffold invokes students' metacognition and challenges students cognitively within their zones of proximal development, to allow them to develop new skills and understanding.

Astro-world – a Technology Integrated Solution

To provide the benefit of personalized scaffolding described in the previous section while helping students develop the range of skills necessary as discussed above, an online program named Astro-world was developed by the author for this study to achieve the scalability of the whole-class format while maintaining the benefits of the individualized format in the previous study (Arvidsson & Kuhn, 2016). Advancement in technology opens the potentials for personalized learning that were not possible in the past. Teachable agents that challenged students were found to promote reflective thinking and deeper learning (Pareto, 2014; Schwartz et al., 2007).

While technology could be helpful, technology that provides students with a lot of control of their learning process may not always work well for students low in self-regulation (Winters, Greene, & Costich, 2008). Therefore, being able to gauge the level of support needed by a student may be crucial to the success of the technology. Technology has potentials for creating meaningful, personalized, conversational-based learning environments; yet, processing natural language can be so resource intensive that its implementation is prohibitive. Instead, programs may be designed to simulate conversations by matching students' responses to a limited set of expected content, an approach that requires familiarity and expertise in the subject domain and in linguistic analysis (Graesser, 2015).

To provide personalized scaffolding in the problem-based intervention, Astro-world is an automated-agent developed to take the place of a human facilitator in guiding students to problem solve and achieve the goal of the activity by anticipating students' responses. It adapted the scaffolding protocol in the previous Arvidsson and Kuhn (2016) study in facilitating students through all the activities. Given the challenges of natural-language processing, it is therefore important to minimize the complexity of natural-language processing and limit the scope of students' responses that the system needs to process open-endedly for our technology. By leveraging existing understandings and observations of student behaviors from previous studies as they engage in solving the problems in the individualized in-person condition in our earlier work, it has proven possible to focus the technology development on common paths seen among students, in order to develop an effective automated-agent.

Summary

Further examination is warranted of the previously developed intervention (Arvidsson & Kuhn, 2016; Kuhn, Arvidsson, Lesperance, & Corprew, 2017) that was found successful with low performing students in a school serving an academically underprivileged population. The previous study found the individualized condition to be more effective than the classroom condition of the intervention that was designed to develop middle-school students' multivariable thinking through a scientific activity that students undertook collaboratively in pairs (Arvidsson & Kuhn, 2016). Given the results, the research reported here further investigates how to integrate technology that provides individualized supports in a classroom setting to make the intervention more scalable. As the first step of achieving scalability, this study investigated whether the intervention remained effective when the paper and pencils activities were replaced with this technology, Astro-world, in a similar classroom setting as in the previous study when the teacher actively led class discussions and engaged students (Arvidsson & Kuhn, 2016).

Technology potentially frees teachers' time, allowing them to tend to students' individual needs. At the same time, how important is teacher involvement? Would the technology be able to help students develop the important skills of multivariable thinking without teachers' active engagement? Therefore, as a second step, the present study also investigated whether the intervention described above as an automated-agent was effective when the program was administered without active teacher involvement.

In the following sections, we reported on the current study to answer the following research questions:

1. Would the intervention be effective in a classroom setting with active teacher engagement when the paper and pencil activities were replaced by a technology of an automated-agent?
2. If the intervention continued to be effective, would students still benefit from the technology if teachers were no longer actively involved?

Chapter III: Method

Participants

Participants were 64 students from two sixth-grade classrooms of a public elementary school in the San Francisco Bay Area. The school population is 45% White, 37% Hispanic, 5% African American, 4% Asian, 10% other including Filipino, American Indian or Alaska Native, Native Hawaiian or Pacific Islander. About 50% of the student population is socioeconomically disadvantaged. About 45% perform at grade level on standardized Mathematics tests and about 51% perform at grade level on standardized English Language Arts/Literacy tests.

Each year, the school has two classrooms in the sixth grade and its policy is to intentionally assign students so that both classrooms share similar student demographics and performance levels. Each class has its own homeroom teacher but the two teachers each teach their subjects to both classes.

One of these two classes participated in each of the two experimental conditions of the intervention (32 students each, 14 females in one and 13 in the other). An additional group of 56 students (47% female) was drawn from two sixth grade classes the following year. They served as a business-as-usual control group. Students of all four classes showed similar demographics. They all shared the same mathematics and science teacher. All four performed similarly academically. On a 10-item written task of basic COV skill (prior to the intervention for experimental groups), there was no significant difference between the two experimental classrooms.

Design and Procedure

The design included two experimental conditions: a technology-with-teacher-involvement (teacher-involved) condition and technology-only (tech-only) condition. In both conditions, the automated-agent guided students through the activities; new problems only appeared when a problem was solved. Students worked in pairs throughout, with pairs chosen by matching gender when possible and matching performance in a 10-item basic COV skill written pretest. Whenever there was absence, students were paired with another student or student pairs so that they did not work alone. Each student pair shared one user account for the Astro-world program. Pairs were instructed to share their ideas and only respond when they had reached agreement.

The problems, goals and activities that constituted the intervention were the same as those in the previously described studies (Arvidsson & Kuhn, 2016; Kuhn, Arvidsson, Lesperance, & Corprew, 2017). The only difference is that all activities were engaged electronically with an automated-agent, using the Astro-world program. In the intervention, additive-correlational relations are investigated between the factors and the outcome variable. Because the intervention is intended to introduce students to multivariable relations and it is a challenge for middle school students to understand additive effects, interaction effect, while an important concept, is beyond the scope of the current intervention.

The intervention consisted of three phases of activities, taking place during between three and nine class sessions (varying by the rate of a pair's progress through the phase) of 45 minutes each. The entire class began each phase at the same time, but within phases students progressed at an individual pace. Pairs who completed the first or second phase ahead of the rest of the class did unrelated work until all other pairs were ready to continue to the next phase.

Throughout all three phases, in the teacher-involved condition, as student pairs interacted with the automated-agent in the program, the classroom teacher and the researcher circled the room actively to look for student pairs who appear to need help and engaged students as needed.

They never, however, suggested strategies for solving the problem. As needed, they further scaffolded students by encouraging them to read the prompts from the automated-agent out loud if students appeared to not be reading the prompt carefully, making sure that students were comprehending the prompts correctly; they may also observe conversations between the student pairs and helped them collaborate more productively, for example by listening to each other or by making sure that they discuss their ideas with each other and come to consensus before responding to the automated-agent. In addition, whole-class discussions were held at the end of each phase to allow students to share reports of how they solved the problems. During the class discussions, the researcher asked similar questions as the automated-agent did but, again, never provided hints or instructions for how the problems were to be solved.

In the tech-only condition, the teacher and the researcher facilitated the activities logistically but did not engage students actively throughout each session except for ordinary classroom management to ensure that students were on task. Throughout all three phases, the teacher and researcher were present in the room but passively waited for students to ask questions, which were mostly related to classroom management issues, such as whether they could go to the bathroom. In a few cases when the questions were about the activities, the researcher redirected students to read the prompt carefully to make sure that they understood what they were asked to do. The researcher also observed the classroom to make sure that students were on task and redirected them back to the activities if they were off task. No whole-class discussions were held for students to share ideas with a larger group.

Intervention

In the first session, the researcher introduced Astro-world to the entire class using a powerpoint presentation She explained that the objective of Astro-world was to figure out what matters to astronauts' performance by reviewing records about astronaut candidates and their performances, the purpose being to predict the performance of new applicants and therefore select the best team. Once the researcher confirmed that students understood the objectives, they were asked to log on to the program online to begin the activities.

At the left side of the online program is a chat window where an automated-agent "chats" with the student pairs. The right side of the program presents the student with the appropriate screen for the activity. Besides prompting students based on their responses for the specific activities, the automated-agent also periodically reminds students to decide together and remind them to talk to each other. Figure 1 shows the introduction to the program as it appeared on the screen. Figure 2 shows the initial screen in which participants began to interact with the agent.

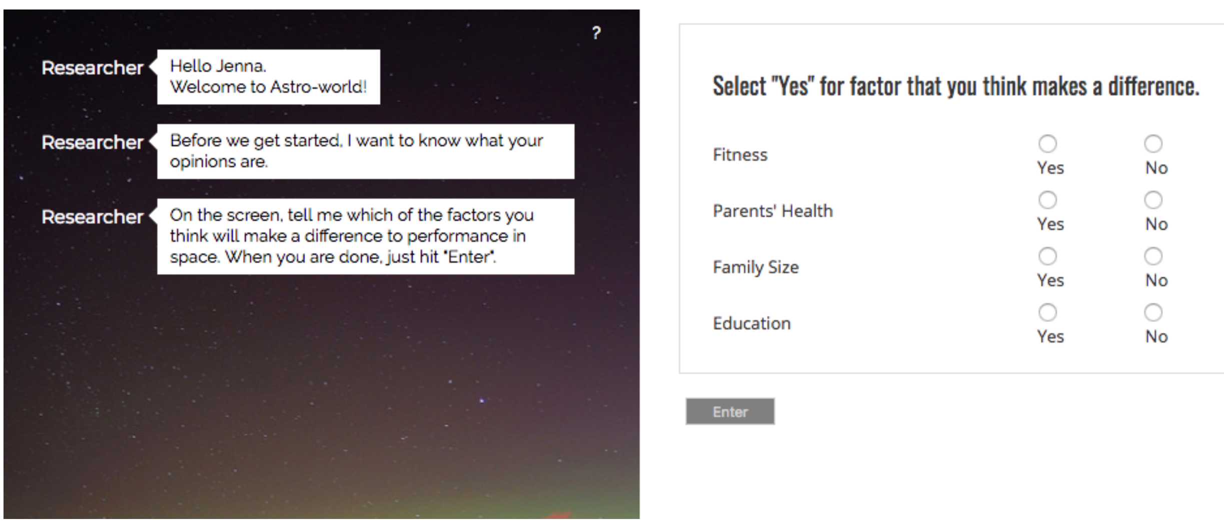


Figure 1. First screen that users see after logging on to the Astro-world. It asks the users about their beliefs

Control of variables phase

In this phase, students are to figure out whether fitness, parents' health, family size and education matter by comparing individual records of astronauts. The program begins by asking students about their beliefs. It then lets students try selecting a record by selecting the levels of each factor. The main activity then allows students to figure out whether a factor matters, one factor at a time, beginning with fitness and followed by parents' health, family size, and education. For each target factor, the automated-agent asks students to choose whether they want to see one or two records and which record or records they want to see (see Figure 2 for an example).

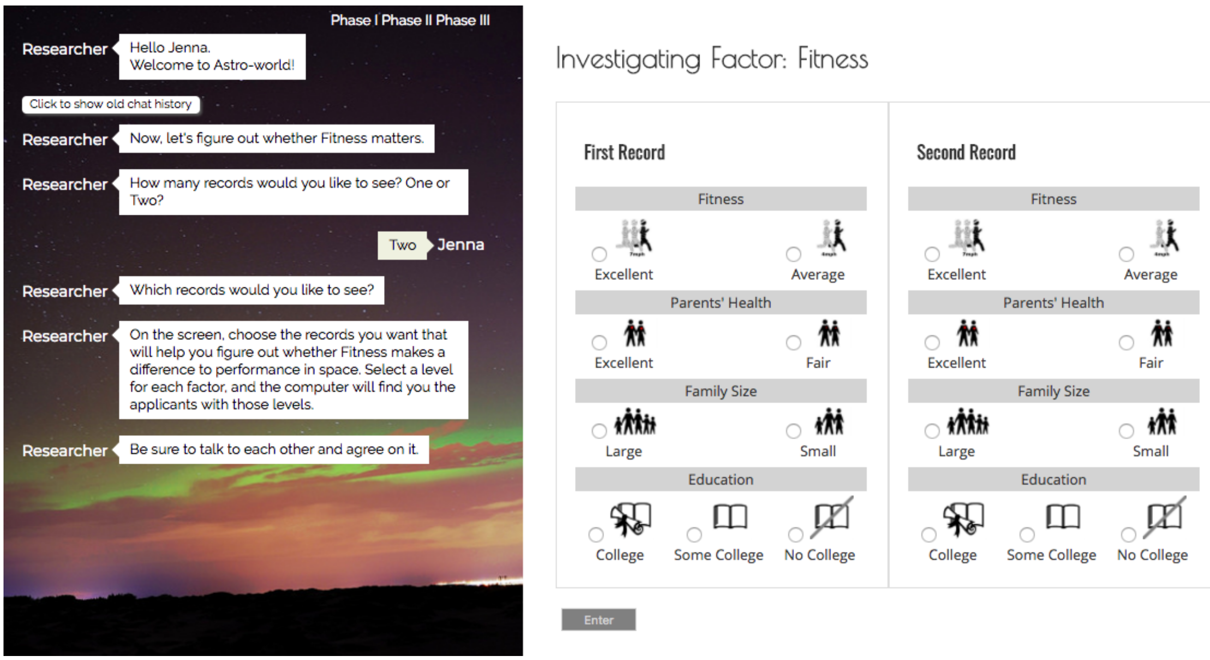


Figure 2. Automated-agent asking students which records they want to see

As students make their selections, records are presented and students are asked to draw conclusions on whether the target factor matters. To successfully complete the activities in this phase, students must be able to select records that constitute a controlled comparison for a particular variable, in order to reach a valid conclusion regarding whether the target factor matters to astronaut performance. Once pairs do that successfully, they are asked to write a memo about what they found out. Only then do they proceed to investigate the next factor.

At each step, the automated-agent asks relevant questions. In the case of an invalid conclusion, the agent challenges students' conclusion by pointing out its potential flaws, i.e., that an alternative (uncontrolled) factor could be responsible for a variation in outcome. Participants are then invited to conduct further investigations. Holding variables constant, however, is never explicitly suggested as a strategy. See Figures 3 to 5 for examples.

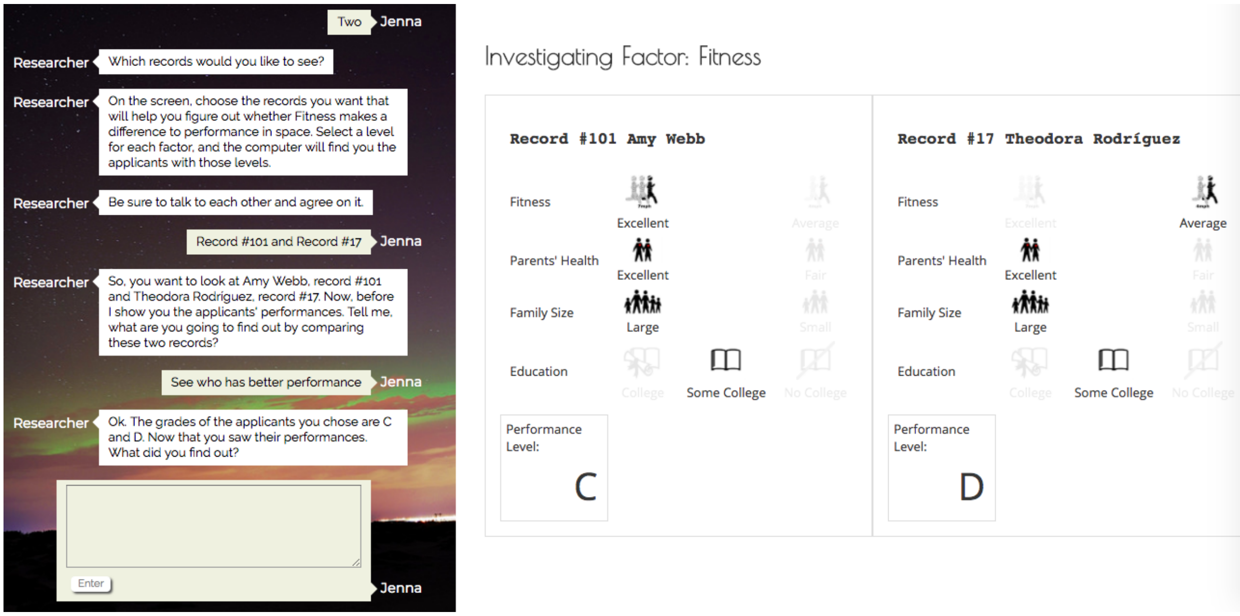


Figure 3. Automated-agent asking students about what they found out

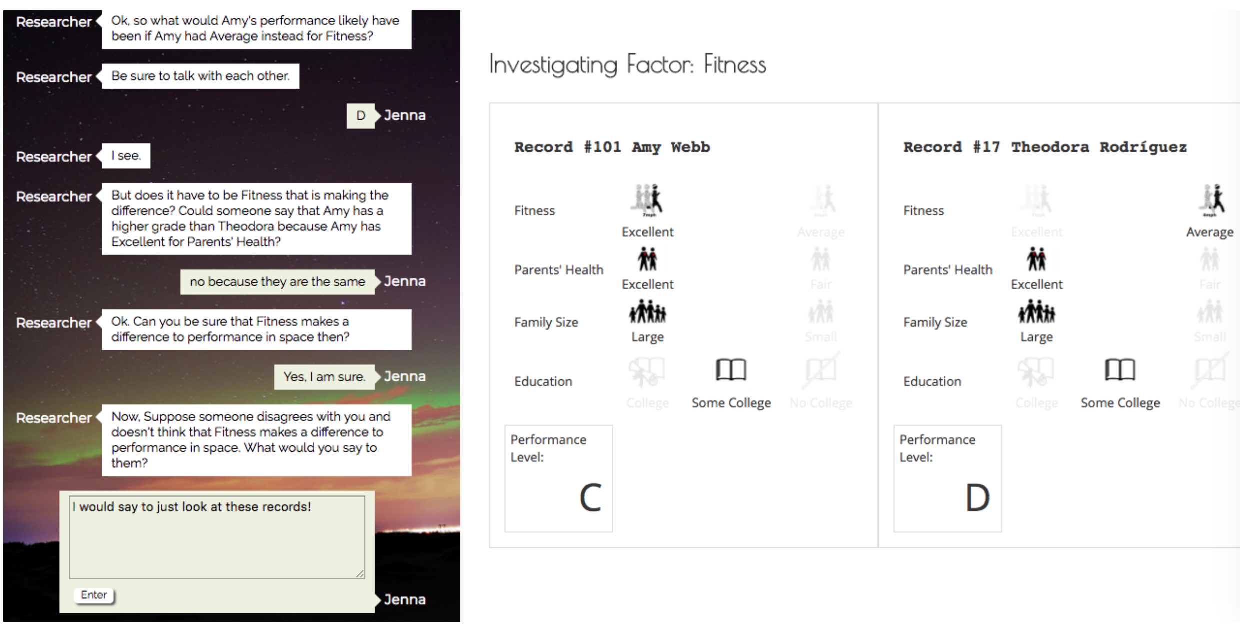


Figure 4. Automated-agent challenging students

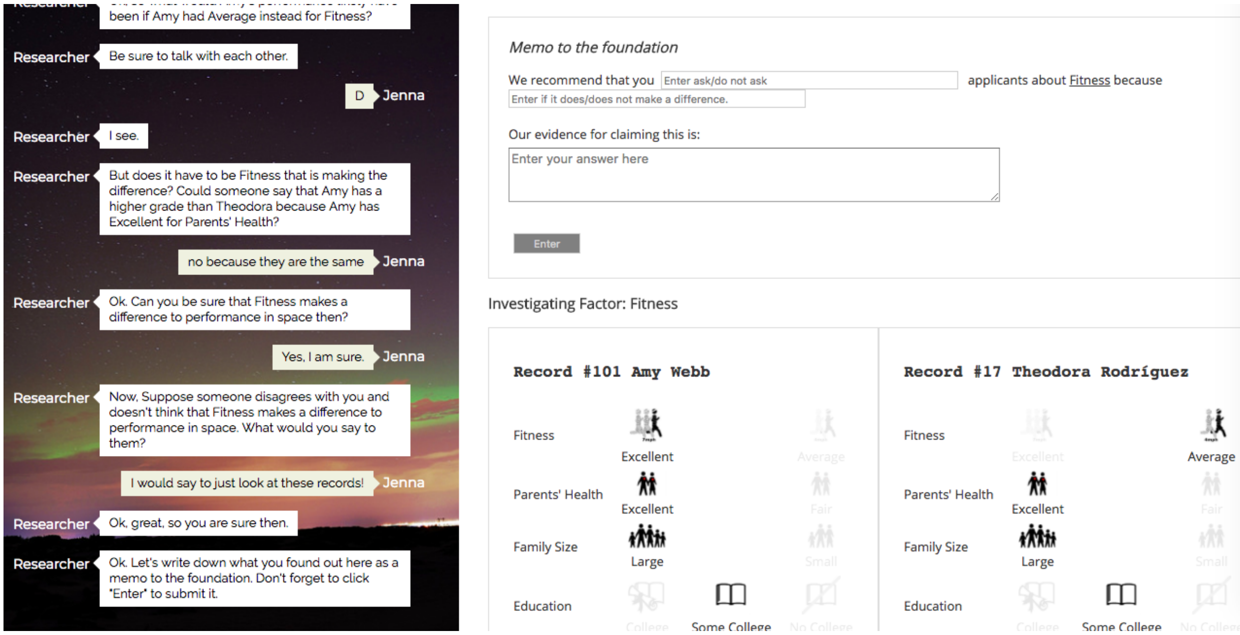


Figure 5. Automated-agent asking students to write a memo about what they found out

This phase ends when all four factors have been investigated and correct conclusions drawn (specifically that fitness, parent's health and education are related to astronauts' performance and family size is not). In the tech-only condition, students were then introduced to the next phase at the next class session. In the teacher-involved condition, at the end of this phase a class discussion was held and students were asked to demonstrate how they chose the records on a whiteboard in order to figure out whether a factor matters. The researcher then held discussions about the conclusions that follow based on students' chosen records, in a manner paralleling the automated-agent in the program. This concluded the end of the first phase for the teacher-involved condition and students were then introduced to the next phase at the next class session.

Chart reading phase

In this phase, in addition to the four factors that students looked at in the previous phase, one more factor is added to be investigated. In both conditions, the researcher introduced students to this phase by discussing with students whether they think firm conclusions can be made based on just the few records that they have looked at. They reach agreement that it would be better to examine a larger number of cases, and the researcher introduces them to a method of examining aggregate data. She told the students that by studying charts of data, it would be possible to see many more data than what they looked at in the previous phase. The automated-agent then introduces and guides the students through interpreting such charts, following which pairs investigate the status of a factor as displayed on the data charts. See Figures 6 to 8 for illustrations of the automated-agent introducing this activity to students.

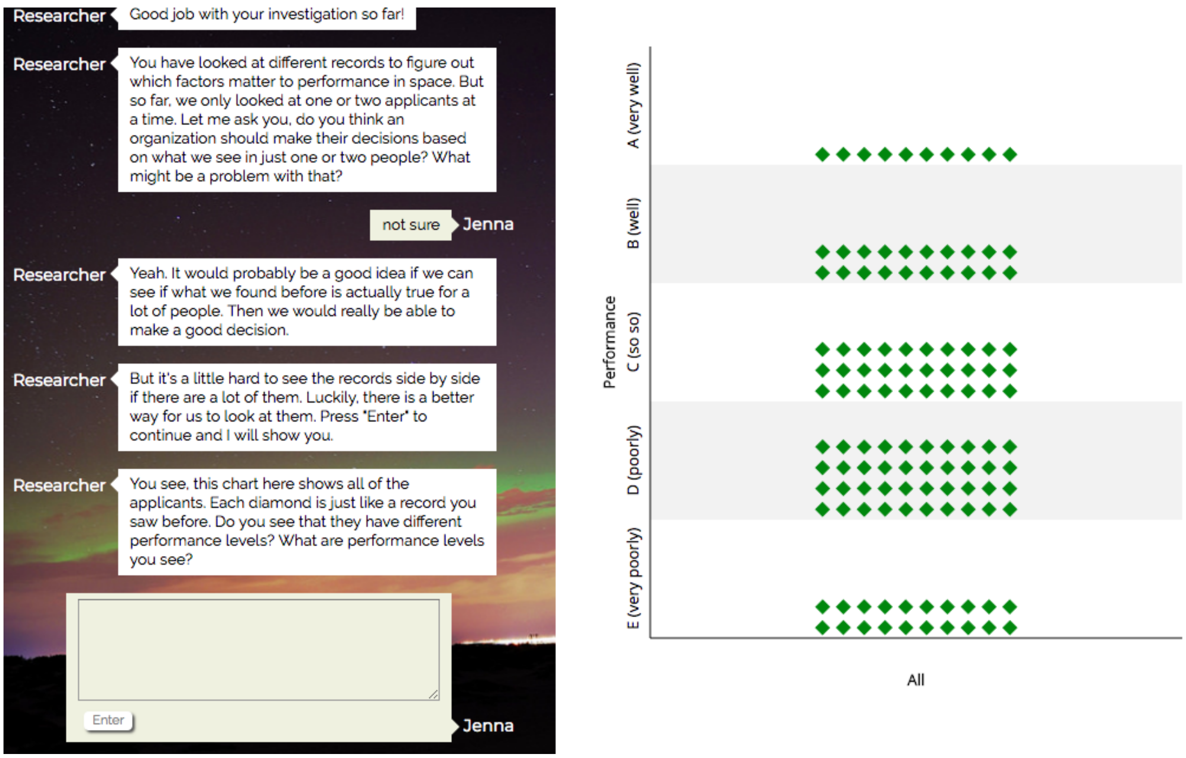


Figure 6. Automated-agent introducing charts to students

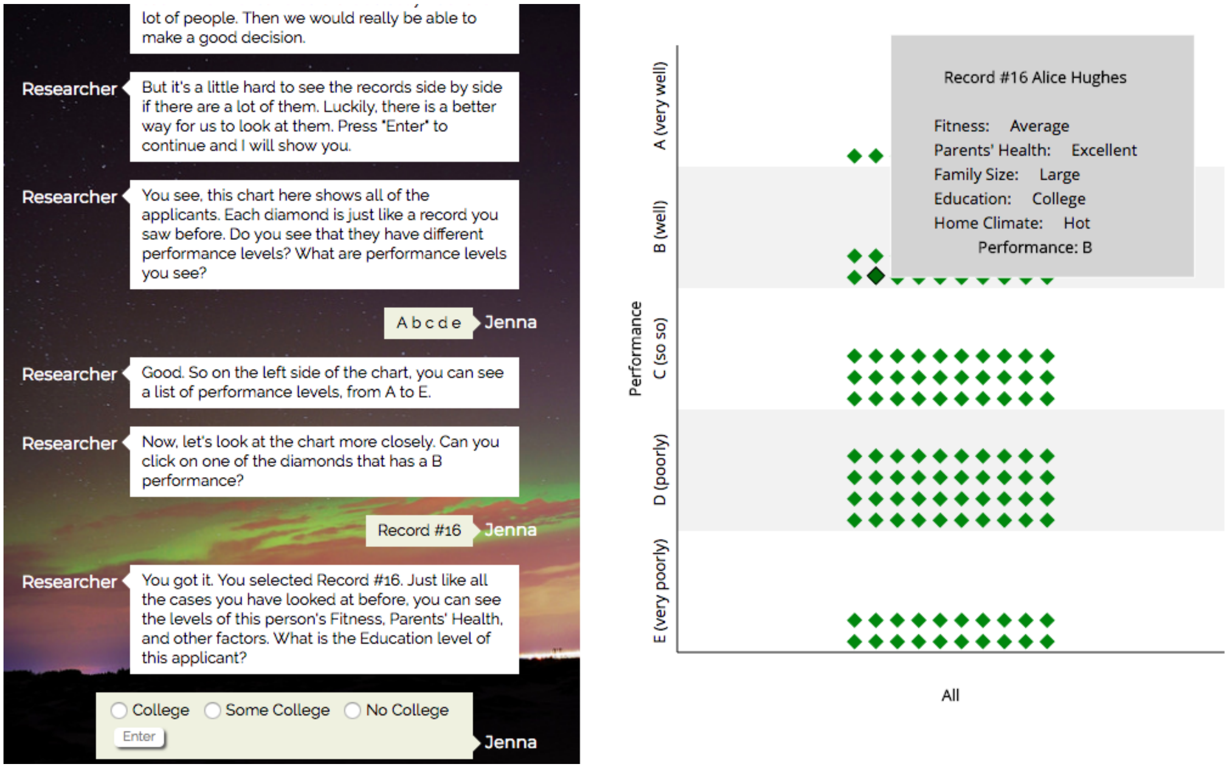


Figure 7. Automated-agent ensuring that students understand how to read the charts

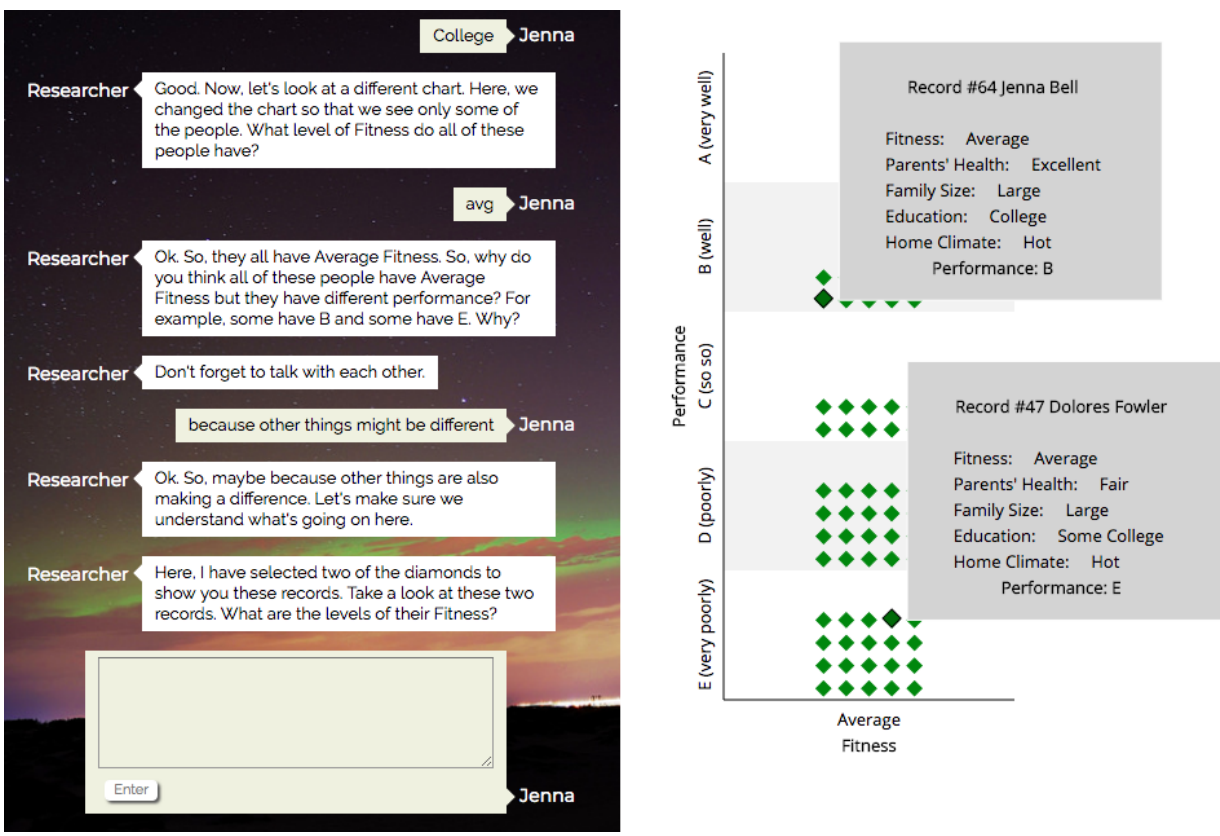


Figure 8. Automated-agent showing students that multiple factors matter to performance

The automated-agent then begins the activities for this phase by showing a chart of the Fitness factor to the students. The automated-agent then queries students about what they can conclude, again challenging them about their conclusions if warranted. Once students draw the correct conclusion based on the data presented, they write a memo about what they found out and then proceed to the next factor. See Figure 9 for an illustration of the activity in this phase.

Once all of the factors have been investigated, the automated-agent then asked the pair to provide a summary of what they found. If students offer a wrong conclusion in their summary, the automated-agent will ask them to review the chart again, querying them appropriately. At the end of this phase, and only in the teacher-involved condition, class discussion was held regarding data interpretation for the education factor (causal) and the home climate factor (non-causal).

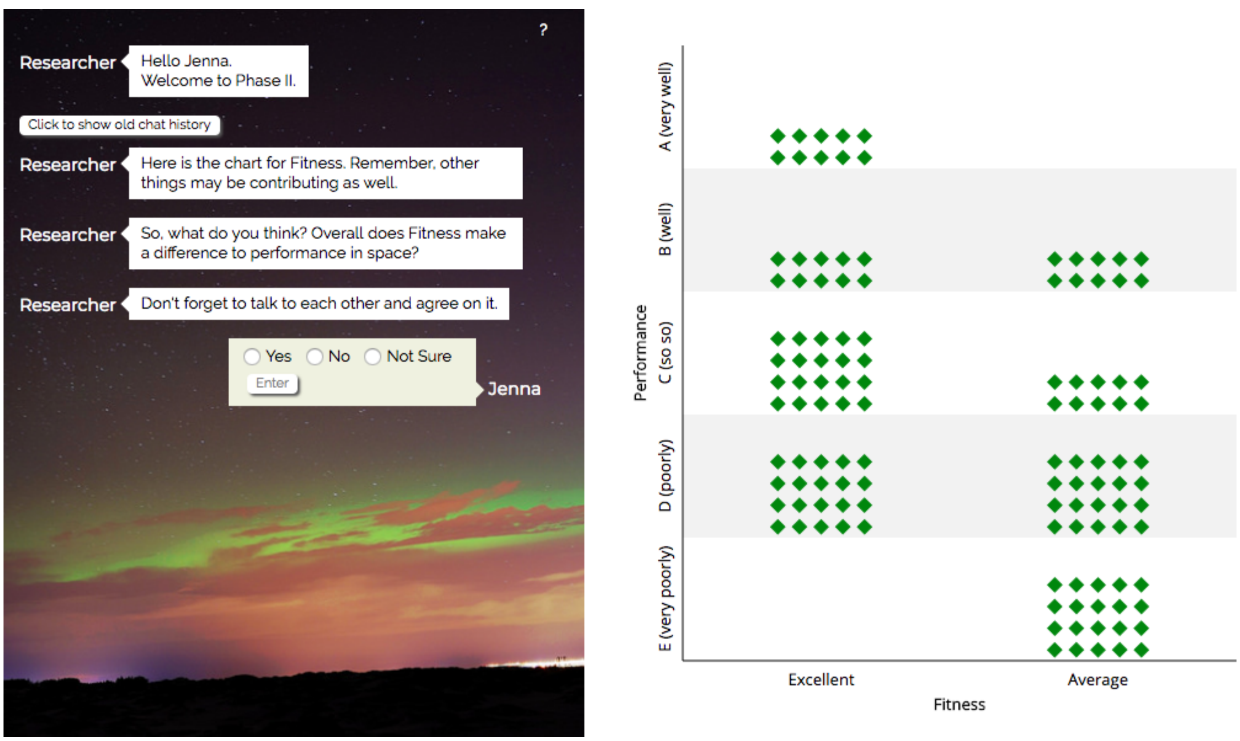


Figure 9. Automated-agent showing students a chart about Fitness

Prediction phase

In the third and final phase, students are asked to make predictions about the performance of 10 new astronaut applicants and, finally, to choose the best team of 5. In both conditions, the researcher introduced students to this phase by telling them that they can now make use of what they have found to make the best predictions about new astronaut applicants . The automated-agent then guided the students through the activities.

Before students begin making predictions, the automated-agent asks students to identify up to four factors that they need information about for an applicant in order to make the best predictions (see Figure 10 for a sample screen). If a pair does not choose all of the causal factors (fitness, parents' health, education), the agent presents the data chart for the missed causal factor and scaffold the students to review the chart again and recognize its relevance.

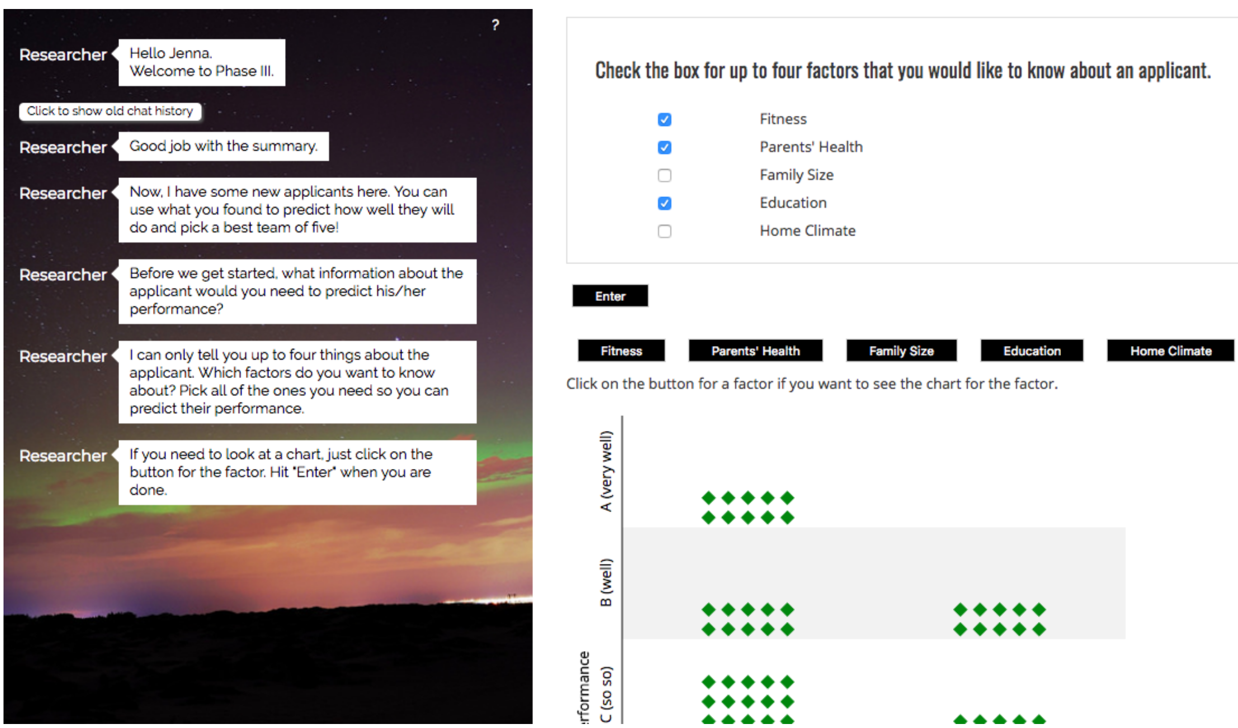


Figure 10. Automated-agent asking students to select factors they would like to see about an applicant in order to predict the performance of the applicant

The automated-agent then presents students with one astronaut applicant at a time, including that applicant's status on all of the relevant factors and one non-relevant (non-predictive) factor. Pairs then predict the performance of the applicant (see Figure 11). Once the prediction is made, pairs are asked about the factors that mattered to their predictions (see Figure 12). If they choose anything other than only the three relevant factors, they are again presented with the corresponding data charts to review. They then have an opportunity to revise their predictions if they wish.

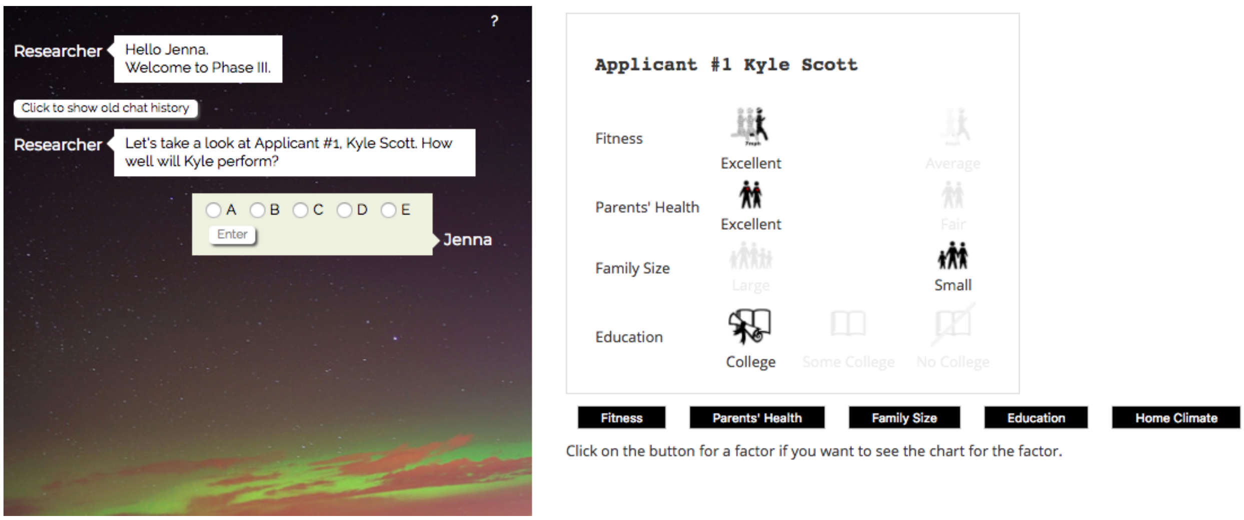


Figure 11. Automated-agent asking students to predict the performance of an applicant

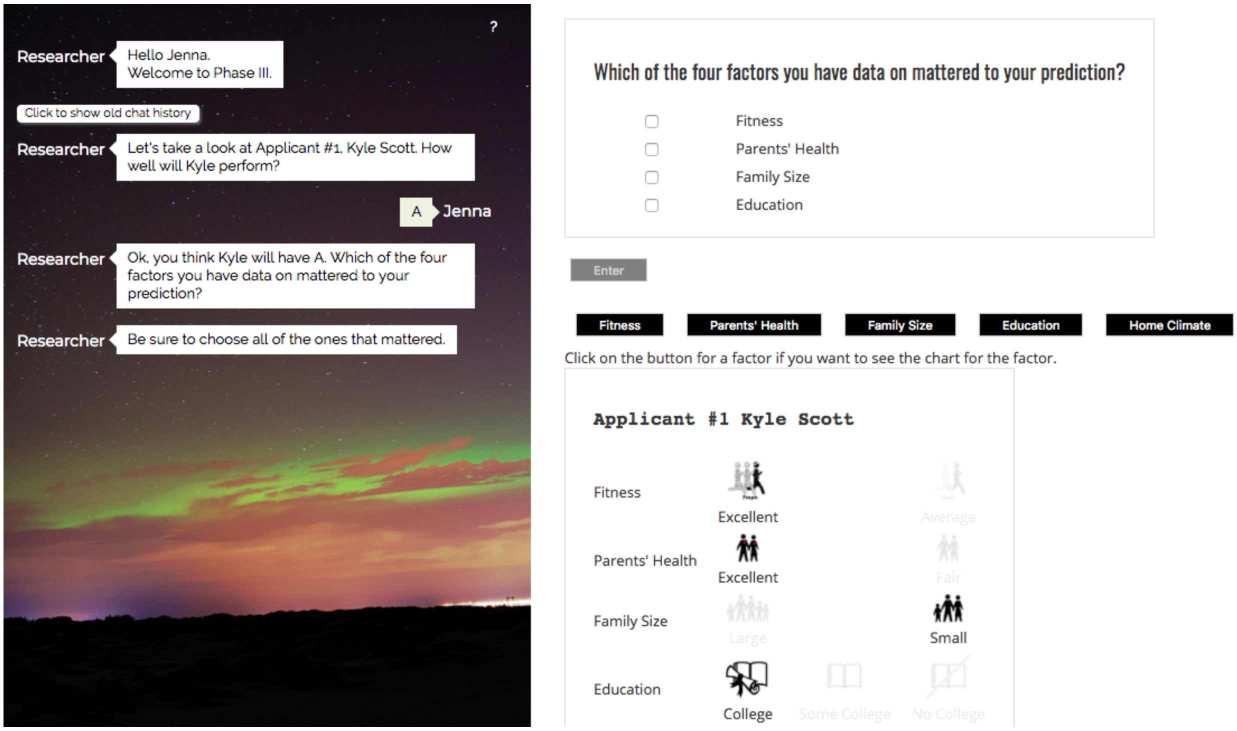


Figure 12. Automated-agent asking students about what mattered to their prediction

Once students completed their predictions for all 10 applicants, they were asked to choose the best team of 5 applicants (see Figure 13). In the teacher-involved condition, a concluding class discussion occurs in which a tally of students' predictions was presented and students were able to see how much agreement existed. Disagreements were discussed, and relevant data charts reviewed together to resolve them. Finally, the researcher reviewed students' initial beliefs about the role of the factors and how many differences in belief there were. Now there were few if any disagreements. In the tech-only condition, a similar tally of students' predictions is presented but no further demonstrations were held on how predictions were made. Finally, in both conditions, the researcher reviewed students' initial beliefs about the role of the factors and how many differences in belief there were. Now there were few if any disagreements. In conclusion, in both conditions, the researcher emphasized the power and importance of investigating and obtaining data, rather than relying only on beliefs.

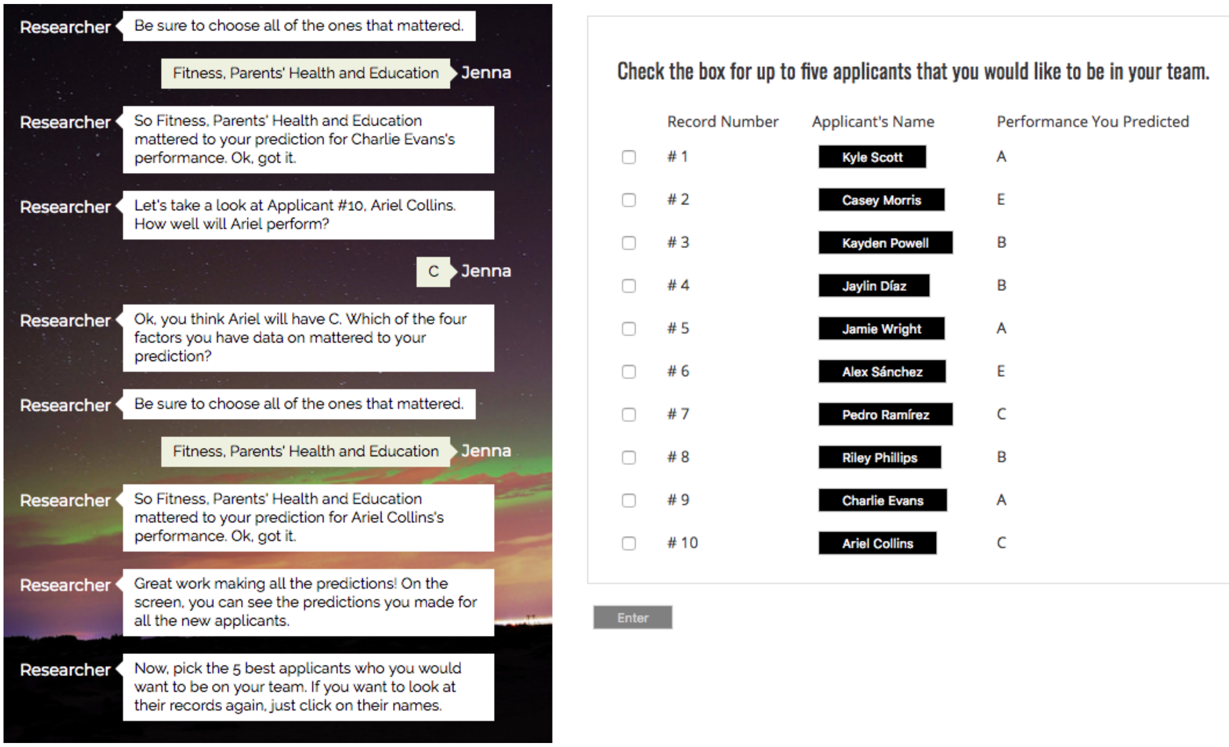


Figure 13. Automated-agent asking students to select the best team of five applicants

Posttests

A far-transfer posttest was administered two weeks after completion of the intervention. The contents of the assessments described below were unrelated to what was in the intervention. Students worked on these paper and pencils tests individually without any scaffolding. The same assessments were administered to the control group during the same period of the academic year.

COV

A three-item task of two levels of complexity was used to assess students' COV skill in a context unrelated to the intervention. For each item, students were asked to select among a multiple choice of two cases in order to investigate the effect of a specified factor on an outcome. There were two variables in the first two items and three variables in the third item. See Figure 14 for an example and Appendix A for the complete task.

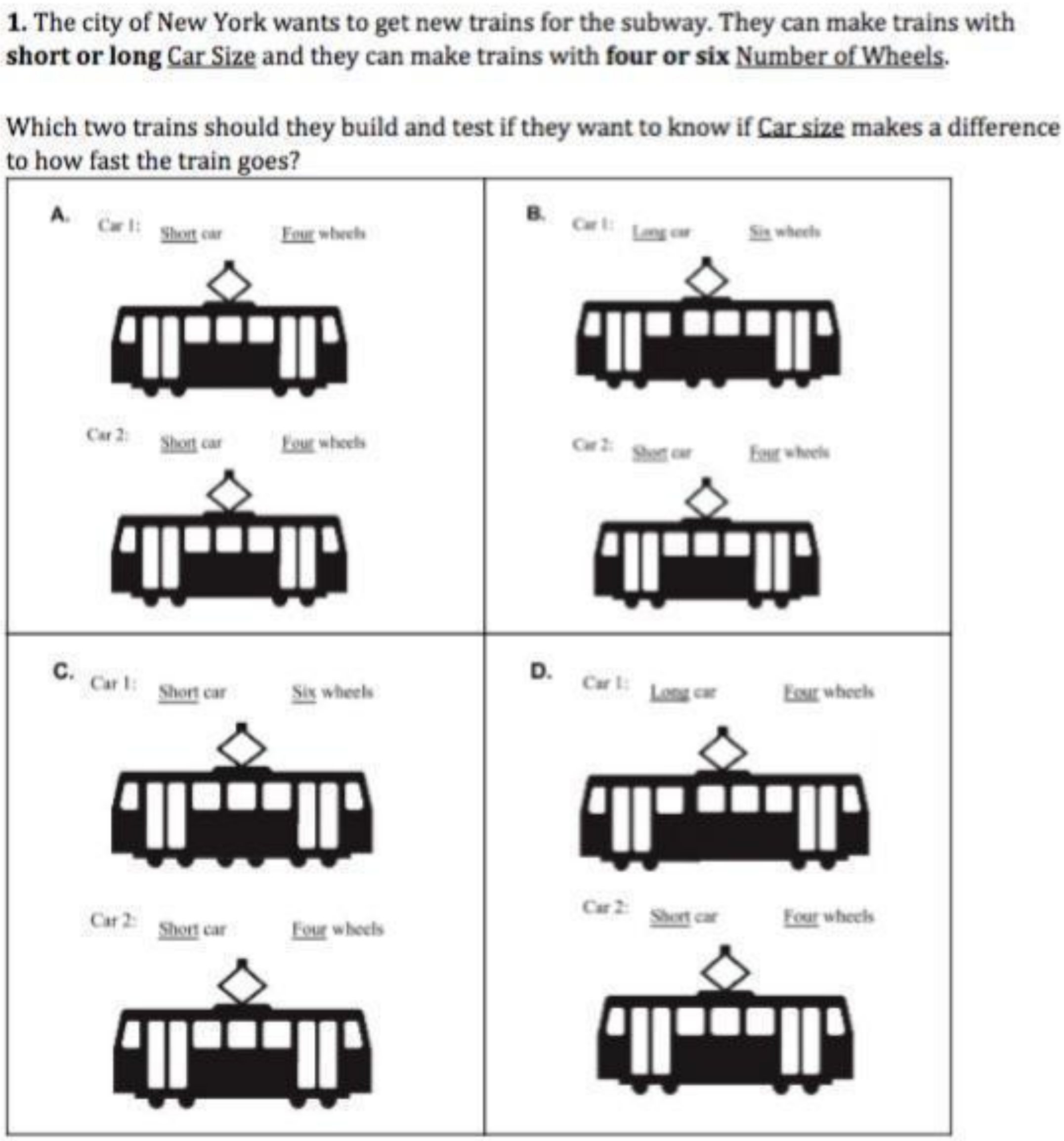


Figure 14. An example of the COV task.

Life Expectancy (LE) posttest

A 6-item task was used to assess multivariable analysis and prediction skills introduced in the intervention with new content. The task was previously used for this purpose among adults (Kuhn, Ramsey, & Arvidsson, 2015), high school (Kuhn, Arvidsson, Lesperance, & Corprew, 2017) and middle school (Arvidsson & Kuhn, 2016) populations. It was previously reported on by Kuhn, Ramsey, and Arvidsson (2015). Using authentic but simplified data from a World Bank 2010 report, the task presented factors (Employment, Family size, Education, Home Climate) found predictive of average life expectancy across countries and one non-contributing factor (Country size). See Figure 15 for the information presented to the students.

Students were asked to make predictions regarding the category of life expectancy (very lo, lo, medium, and hi) of additional countries based on information provided to them about the relations between each factor and average life expectancy of countries. After each prediction was made, they were asked to circle all of the factors that mattered to their prediction. This was to evaluate how well students were able to explicitly attribute all the causal factors to make the best predictions possible. Three practice items were given to students to ensure that they understood the task. Another six items were used to assess their skills. See Appendix B for the complete task.

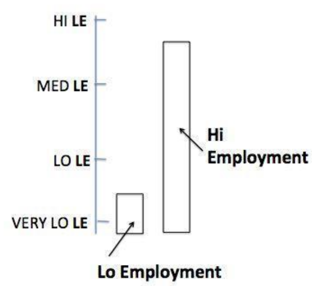
Tear this page

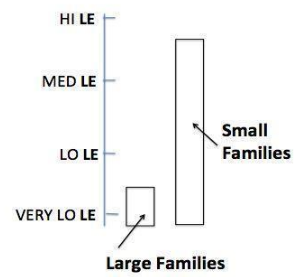
Some people live very long lives. Others die at an early age. What makes the difference? Life Expectancy (**LE**, for short) is the term for how long people on average are expected to live. **LE** differs greatly across different countries. Some countries have a much higher average **LE** than others. What causes the differences? Here are some possibilities that studies suggest.

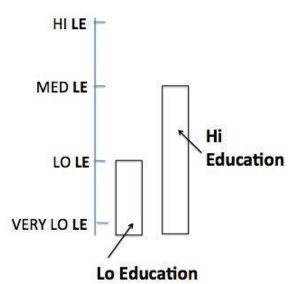
One is employment. As the chart below shows, countries where people have high employment levels have a higher average life expectancy than countries where employment is low (many people are without jobs). As you see, the chart shows that employment makes a very big difference to **LE.**

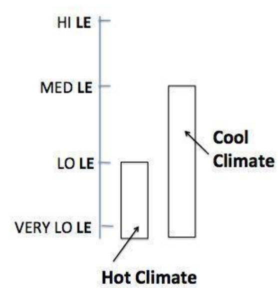
The other charts show other things that can affect **LE.** As you see, family size, like employment, makes a very big difference. Smaller families on average mean higher **LE.**

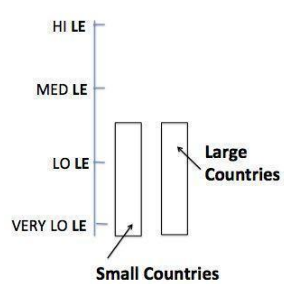
Other things, like education and climate, make a smaller difference. And some, like country size, seem to make no difference at all - **LE** overall is about the same for small and large countries.











Now your task is to make some predictions about **LE** for different countries. You can look back at these charts when you want to. For each country, predict the average **LE** you think that country will have.

5

Figure 15. Introduction to the LE task.

Chapter IV: Results

Progression through the Intervention

Pairs in the teacher-involved condition spent four 45-min class sessions in the first phase, three sessions in the second phase, and two sessions in the final phase. Pairs in the tech-only condition spent three 45 minutes class sessions in the first phase, three sessions in second phase, and one session in the final phase. On average, students in the teacher-involved condition spent a total of 169 minutes engaged with the Astro-world program while those in the tech-only condition spent an average total of 136 minutes.

Posttest Performance

Posttest measures served to assess students' mastery of COV and multivariable reasoning, as well as the control of variables logic that we believe to be a necessary foundation for it.

COV performance

To find out whether the intervention was effective in improving the foundation skills of making controlled comparisons, we compared the percentages of those who successfully made controlled comparisons in all three items of the COV task between the experimental and the control conditions (see Table 1). Three 2x2 Chi-square comparisons with Bonferroni correction to maintain a familywise Type I error rate of .05 were made. Therefore, instead of the standard alpha of 0.05, alpha was set to 0.05 divided by 3, which is 0.017. A higher percentage of students in the teacher-involved condition (60%) consistently selected controlled comparison compared to the percentage of the control group (37.5%), however, the differences did not reach significance, , $X^2(1) = 3.13$, $p = 0.077$. Similarly, the tech-only condition (55.6%) had a higher percentage of students who consistently selected controlled comparison compared to the percentage of the control group (37.5%) but the difference was not significant, $X^2(1) = 1.74$, $p = 0.187$. There was also no significant difference between the two experimental conditions found, $X^2(1) = 0$, $p = 0.944$. Consistently, correlational analyses between students' COV posttest and their COV pretest scores were significant for both conditions, $r(20) = .63$, $p = .002$ for the teacher-involved condition and $r(23) = .44$, $p = .028$ for the tech-only condition.

Table 1. Percentages of Students who Consistently Made Controlled Comparison by Condition

	Teacher-involved condition	Tech-only condition	Control condition
Consistent controlled comparison	60.0% (18)	55.6% (15)	37.5% (21)
Others	40.0% (12)	44.4% (12)	62.5% (35)

Note. N = 30 in the teacher-involved condition, 27 for tech-only condition and 56 for control condition.

In sum, while a higher percentage of students in the teacher-involved and the tech-only group out performed those in the control group in mastering the COV skills, the differences were not statistically significant.

Multivariable thinking and prediction

To assess the accuracy of students' predictions on the multivariable reasoning task, prediction errors for each item were calculated. There are four prediction levels (very lo, lo, medium, hi). The error score indicates the difference between the student's prediction from the correct prediction. There are three possible scores for each item: a score of zero indicates that there is no error in the prediction; an error score of one and an error score of two indicate that the student's prediction is one or two levels, respectively, too low or too high. An error score cannot exceed two. To take into the consideration that students might make careless mistake as they calculate the level of their predictions in some of the items, the modal error score will be used to analyze how well students made predictions. See Table 2 for the percentages of students with modal error score of zero for different conditions.

We compared the percentages of those who had a error score of zero as their modal prediction error score in the LE task between the experimental and the control conditions. Three 2x2 Chi-square comparisons were again analyzed. Similarly, Bonferroni correction was made to maintain a familywise Type I error rate of .05 for the contingency table, therefore alpha was again set to 0.017. Higher percentages of students in the teacher-involved condition (87.1%) were significantly less likely to make errors in their predictions than those in the control group (58.9%), $X^2(1) = 6.14$, $p = 0.013$. There was nearly no difference between the percentage of students in the tech-only condition (59.3%) and the control condition (58.9%) in having zero modal error score, $X^2(1) = 0$, $p = 1$. The teacher-involved condition (87.1%) also made less error than the tech-only condition (59.3%), but the difference was not significant, $X^2(1) = 4.47$, $p = 0.034$.

Table 2. Percentages of Students who Most Frequently Made Correct Predictions by Condition

Modal prediction error score out of 6 items	Teacher-involved condition	Tech-only condition	Control condition
zero error	87.1% (27)	59.3% (16)	58.9% (33)
error score of 1 or more	12.9% (4)	40.7% (11)	41.1% (23)

Note. $N = 31$ in the teacher-involved condition; $n = 27$ in the tech-only condition and $n = 56$ in the control condition.

Also examined was degree of consistency in attributing predictive power to a factor. A score from 1 to 5 was assigned based on pattern of answers across the six items. These were defined as follows:

5 - Chose four contributing factors completely consistently across 6 countries

4 - Chose multiple consistent (but not all four causal) factors across 6 countries

3 - Chose multiple but inconsistent factors across 6 countries

2 - Chose only one consistent factor across 6 countries

1 - Chose only one but inconsistent factor across 6 countries

Table 3 displays the distribution of the attribution scores by condition. Majority of students in the teacher-involved (65.4%) and a portion in the tech-only (36%) and control (28.3%) group , mastered the multivariable thinking skills and scored the perfect score of five. They were able to consistently consider all contributing factors (and only the contributing factors) when they made their predictions. A small portion of students scored 4 (7.7% of teacher-involved, 12% of tech-only, and 13.2% of control group), they were able to consistently refer to the same factors for all their predictions, however, they either omitted some of the contributing factors or included the non-contributing factor, potentially demonstrating successful understanding of consistent relationships between variables and that multiple factors contributing to an outcome and yet, their beliefs of what factors mattered potentially hindered them choosing all of the contributing factors. Majority of those in the tech-only (44%) and the control (54.7%) group scored 3 who inconsistently attributing more than one factor to their predictions. They appeared to have some understanding of multiple factors possibly contributing to an outcome, however, such an understanding was unstable and was potentially influenced by their beliefs. A few students in the tech-only (8%) and control groups (3.8%) only considered one inconsistent factor for all of their predictions demonstrating a lack of understanding for consistent relationships between variables and that multiple factors contribute to an outcome.

Table 3. Percentages of Students in Different Attribution Pattern by Condition

Attribution pattern	Teacher-involved condition	Tech-only condition	Control condition
5 - Chose four contributing factors completely consistently across 6 countries	65.4% (17)	36% (9)	28.3% (15)
4 - Chose multiple consistent (but not all four causal) factors across 6 countries	7.7% (2)	12.0% (3)	13.2% (7)
3 - Chose multiple but inconsistent factors across 6 countries	26.9% (7)	44.0% (11)	54.7% (29)
1 - Chose a single varying factor as predictor across 6 countries	0.0% ()	8.0% (2)	3.8% (2)

Note. $N = 26$ in the teacher-involved condition; $n = 25$ in the tech-only condition and $n = 53$ in the control condition.

To answer the research questions, we compared the percentages of only those who demonstrated mastery in multivariable thinking skills, scoring the perfect score of five, by consistently attributing all (and only) four contributing factors when making their predictions between different conditions. See Table 4 for the percentages of those who acquired score of five in different conditions. Similar to previous analyses, three 2x2 Chi-square comparisons were analyzed. Bonferroni correction was made to maintain a familywise Type I error rate of .05 for the contingency table, therefore alpha was again set to 0.017.

For the first research question, Chi-square comparison showed that the teacher-involved group (65.4%) had a significantly higher percentage of students who mastered the LE task in attributing all (and only) four contributing factors consistently across all six countries than the control group (28.3%), $X^2(1) = 8.47$, $p = 0.004$. This showed that the intervention continued to be effective in developing multivariable thinking when combining an automated-agent with active teacher engagement.

However, for the second research question, no significant difference was found between the tech-only group (36%) and the control group (28.3%), $X^2(1) = 0.18$, $p = 0.671$. In addition, no significant difference was found between the teacher-involved group (65.4%) and the tech-only group (36%), $X^2(1) = 3.31$, $p = 0.069$. This showed that the effectiveness of the intervention was no longer seen when the automated-agent was used without active teacher engagement.

Table 4. Percentages of Students who Attributed Consistently to All Contributing Factors by Condition

Attribution pattern	Teacher-involved condition	Tech-only condition	Control condition
5 - Chose four contributing factors completely consistently across 6 countries	65.4% (17)	36% (9)	28.3% (15)
Others	34.6% (9)	64.0% (16)	71.7% (38)

Note. N = 26 in the teacher-involved condition; n = 25 in the tech-only condition and n = 53 in the control condition.

COV as a foundation for multivariable reasoning

A further post-hoc analysis attempted to understand how students performed in the LE task between the teacher-involved and the control group for those who mastered the COV skills and for those who did not. The analyses were made by making the comparisons between the two conditions separately for those who mastered the COV skills (scoring the maximum of 4) and those who did not. Table 5 displays the percentages of students who mastered the LE task by consistently attributing all contributing factors when making predictions among those who scored 4 for the COV task between different conditions. Table 6 displays the same thing but for those who scored 3 or below for the COV task.

Z tests were used to compare the observed cell frequencies of teacher-involved versus the control group for those who mastered the LE task. To distribute the risk of Type I error, the chi-square critical value from the entire 2x3 contingency table will be used to test for significance

(Sharpe, 2015). The corresponding critical value for degree of freedom of two is z = +/-2.45. Among those who mastered COV (see Table 5), a significantly higher percentage of students in the teacher-involved group (81.3%) achieved mastery of multivariable reasoning (by consistently choosing only and all of the causal factors in their attribution judgments) than those in the control group (42.9%), $z = 2.648 > |+/-2.45|$. On the other hand, among the subgroup who did not master COV (see Table 6), there was no significant difference between the teacher-involved (33.3%) and control conditions (18.8%), $z = 0.859 < |+/-2.45|$ for those who achieved mastery of multivariable reasoning.

Table 5. Percentages of Students who Attributed Consistently to All Contributing Factors by Condition for Students with High COV scores (scores of 4)

Attribution pattern	Teacher-involved condition	Tech-only condition	Control condition
5 - Chose four contributing factors completely consistently across 6 countries	81.3% (13)	53.3% (8)	42.9% (9)
Others	18.7% (3)	46.7% (7)	57.1% (12)

Note. N = 16 in the teacher-involved condition; n = 15 in the tech-only condition and n = 21 in the control condition.

Table 6. Percentages of Students who Attributed Consistently to All Contributing Factors by Condition for Students with Low COV scores (less than 4)

Attribution pattern	Teacher-involved condition	Tech-only condition	Control condition
5 - Choose four contributing factors completely consistently across 6 countries	33.3% (3)	10% (1)	18.8% (6)
Others	66.7% (6)	90.0% (9)	81.3% (26)

Note. *N* = 9 in the teacher-involved condition; *n* = 10 in the tech-only condition and *n* = 32 in the control condition.

In sum, students performed best on the multivariable task when they had both achieved the conceptual foundation of COV and received the benefit of teacher-involved interactions. Students were unlikely to perform well when they had neither.

Chapter V: Discussion

It is no easy task to effectively engage middle-school students in science practice. Besides designing materials and choosing contents that develop the appropriate skills, to establish a curriculum that will work well in a classroom, there are many challenges, such as creating activities that students find engaging and finding ways to scaffold students as they need help. The goal of this study was to address some of these challenges by adapting a curriculum that was successful in engaging middle school students into a technology that makes individualized scaffolding possible even in a classroom setting. In addition, the study examined whether students would benefit from the technology even without teacher's active engagement.

We saw here that the Astro-world intervention continued to be effective when the paper and pencil activities were replaced with an online technology in a regular classroom setting. The three phases of activities allowed students to experience science as a practice in a collaborative effort with their partners: The COV phase helped students strengthen their foundation of experimental design while the chart reading phase helped them see the need for more data. Finally, the prediction phase solidified the need to refer to empirical data rather than relying only on beliefs in order to make accurate predictions. In addition, the social science context of choosing astronauts to go to space seemed to have kept the students interested and engaged.

According to the teachers, students enjoyed working on the activities:

> *The students enjoyed the activity as it was a change from their usual routine. The process was informative.*
>
> *The program was easy to use and implement with our 1:1 computer to student ratio. It could easily be used during our computer lab time. In fact, if possible, I would like to use it in the future.*

> *The overall reception was positive… Sustained concentration like that is not easy for some. However, students had positive things to say about the program... and it was a new way for them to engage with material.*

At the same time, we also saw the effectiveness disappear when teachers were no longer actively engaged. In addition, students who benefited most were predominantly among those who mastered the foundational control of variable skills.

Advancing Metacognition and Epistemological Thinking

The intervention was effective in helping students develop multivariable thinking. By mastering the LE task, a delayed far-transfer task new to the students, in consistently attributing only and all contributing factors when they made their predictions, the most successful students potentially demonstrated more advanced level of epistemological thinking and strong metacognitive skills in solving these tasks. These students were able to persistently refer to the data provided as they performed the task at hand.

Having the epistemological understanding that knowledge is constructed by evaluating evidence and the recognition that the relationships between contributing factors hold across scenarios and must always be taken into consideration; and the metacognitive skills to regulate and repeatedly put aside their own preconceived ideas of what factors mattered to life expectancy might have helped these students perform consistently well in the LE task. These skills are not readily displayed even in the untrained adult population (Kuhn, Ramsey, & Arvidsson, 2015). Belief bias has long been known to influence one's ability to think critically (Evans, 2003); in the face of evidence, when it contradicts one's belief, evidence tends to be rejected or ignored (Kuhn, Garcia-Mila, & Zohar, 1995). The difficulty of making judgment based on evidence especially when it contradicts one's belief was found to possibly stem from the heuristic processes to respond based on intuition while logical reasoning requires higher effort and cognitive load (Evans, 2003). The intervention might have helped students retain the higher effort of making logical reasoning by inhibiting their beliefs even weeks later when performing new tasks in a different context.

COV as Foundation

Although the intervention helped students develop multivariable thinking, our results showed that mastering COV skills was crucial to the success. This provided further support to the need to developing scientific thinking not as isolated procedural knowledge but rather as a practice, by providing a context for investigation (such as to solve a social problem) and an incentive for doing science. While the result did not reach significance, possibly due to the small sample size and a lack of power, about 20% more students in the experimental groups performed better than the control group in the delayed COV far-transfer posttest task. The multi-phase activities provided practices and strengthened the COV skills that are fundamental to the understanding of designing experiments and testing hypotheses. All of these might have contributed to enhancement of epistemological thinking by showing students what they believed may not align with what the data showed. Mastering the COV skills may imply that students have also better grasp in the crucial role of empirical data.

Argumentative Scaffolding at the Right Time

The success of the intervention involving the automated-agent among 6th graders was an encouraging first step in our attempt to scalably provide individualized scaffolding to students in developing the critical scientific thinking skill of multivariable thinking. While auto-tutors have been successful in improving students' domain specific knowledge acquisition, less investigated is how similar technology may be effective in developing students' scientific thinking in general.

Success was found in technology that provided customized just-in-time hints/instructions in helping students develop COV skills (Sao Pedro, Gobert, & Baker, 2014). However, given the limited success of direct instruction in developing deep understanding of COV skills among lower-performing students (Lorch, Lorch, Calderhead et al., 2010) and that feedbacks and explanation from tutors tend to predict shallow learning (Chi, Siler, & Jeong et. al., 2001), our results add insight into how an automated-agent can provide scaffolding that challenges students to reflect on their own thinking. Such scaffolding may be crucial to success in a school with only 45% students performing at grade level on standardized tests. Most importantly, the current results showed that technology can support students' developing more than simply procedural knowledge, instead, help advance students' epistemological and metacognitive thinking as discussed earlier.

The teachers found the technology to have served well in engaging in dialogues with the students. For example, here was some of the feedback from the teachers:

> *Overall, I was impressed with the program's ability to anticipate student answers and generally guide them to understand the concept of how to control variables.*
>
> *The program was good in that it was simple for the students to understand and allowed them to interact with their partners to create responses.*

Based on what they observed, the teachers also felt that scientific thinking skills improved:

> *I'd expect them to do a little better at not taking their opinions at face value, but realize that they have to test opinions as hypotheses. I would hope that they would be able to transfer the argument work into their class writing, and vice versa.*

I would expect the students to generalize the concept of how to test for single variables within the context of science.

Much of the success may be due to the technology's ability to scaffold students at the time of the activities to help them reflect on their thinking. An example is student pair A's early dialog below:

Auto-agent: Ok. So you think that Fitness makes a difference to performance in space. What makes you say that?

Student pair A: You have to be strong to be in space

Auto-agent: That may be true. But can we really tell for sure that Fitness makes a difference to performance in space? Couldn't someone say that it is because John has Excellent for Parents' Health and that is why Jan has a lower grade than John? Can you really know for sure that Fitness gave Jan a lower grade?

Student pair A: No, I am not sure.

In the above dialog, these students drew their conclusions based on their beliefs. But then later, after a few tries, it can be seen that the answers of this same pair to a similar question changed; instead, they pointed to the data as justification for their conclusion. It appeared that their epistemological thinking advanced as the automated-agent pointed to the flaws in their conclusions.

Auto-agent: What makes you say that?

Student pair A: the records told us Stewart is better at fitness and so had a higher grade

In addition, students also had the chance to practice making counterarguments as you can see in the dialog below from the same students above:

Auto-agent: Ok. Can you be sure that Fitness makes a difference to performance in space then?

Student pair A: Yes, I am sure.

Auto-agent: Now, Suppose someone disagrees with you and doesn't think that Fitness makes a difference to performance in space. What would you say to them?

Student pair A: Fitness matters and we know because the records say that it does even if both people have the same parents' health, family size, and education.

In this above response, the same students who advanced their thinking to draw their conclusions based on what they learned from the data were also able to articulate their strategy and point to the data as evidence that supported their argument to justify their claim and counter their opponents.

These dialogues demonstrate the way the automated-agent helped students advance their own thinking, at their own pace. Rather than providing hints, by questioning and challenging students' conclusions, the agent enabled students to reflect on the flaws in their own thinking and possibly revise it.

Let Teachers Teach – Automated-agent Complementing Teachers

While the intervention was found successful, the effectiveness disappeared when active teacher involvement was removed. Even though students in the tech-only condition completed all the activities, they did not perform better than the control group in the delayed far-transfer tasks. In other words, because the activities were structured so that the automated-agent only advanced students to the next task once they exhibited the skills needed to complete the task at hand, students in both experimental groups achieved all of the skills during the intervention.

However those in the tech-only group were less able to retain them and transfer them to the new tasks at the delayed posttest.

There are many possible reasons why the intervention lost effectiveness without active teacher involvement. First of all, when the teacher was actively participating in the intervention by leading class discussions and circulating the classroom observing students and asking questions, the amount of scaffolding experienced by students would likely be more than those experienced in the tech-only classroom. The increased frequency of scaffolding, especially the kind that prompted reflections rather than the ones provided feedbacks, might have led to more constructive responses from students, which may in turn lead to more effective learning (Chi, Siler, & Jeong et. al., 2001). In this environment, students may have learned more directly from interacting with the teacher and indirectly from the perceived active presence of the teacher. The following factors might be at work: Social context may have increased students' motivation for deeper engagement and metacognition, which may have led to deeper learning. Modeling may have strengthened what students had learned in the activities while supporting better collaborations. Direct interaction with the teacher may have provided regulation to students that struggled to self-regulate.

Social context – problem solving together

In the teacher-involved condition, teacher led discussions about how to solve the problems at the end of each phase. Teachers' active presence may have conveyed a positive attitude towards to the problems the class worked together to solve and projected onto the students the importance of learning. It creates a shared sense of why these activities were worth participating in.

During these discussions, teacher scaffolded students in front of the class and called on different students to help solve the problem. In this way, the whole class worked together in finding the best strategy and providing better supported arguments to the teacher's challenges. Students observed the mistakes that other students made when discussing with the teacher and they sometimes participated in discovering the flaws of a proposed solution or coming up with better solutions. This social environment of class discussions afforded students the chance to solidify their learning by reflecting back on their own activities with the automated-agent and then practiced the best strategies by interacting with the rest of the class. Excerpt from a class discussion in a previous study demonstrated how students might benefit from a class discussion.

Student H.: I think that home climate does matter… Because, what if their home climate… When they are doing the test, what if it's like hot and then, say, they are not so used to being cold, maybe that can also like give them a bad performance.

Teacher: Remember how we talked about opinion versus data? What does the data show in terms of home climate. *(Showed the relevant chart on whiteboard)*. So, what does the data tell us about home climate.

Student H.: The people… oh…

(Teacher described what were on the axes of the chart)

Student H.: This is the same amount. Everybody who scored like different levels… it's the same amount of people in cold climate.

Teacher: So according to this data, I hear what you are saying, you made a good point. But according to this data, does it matter?

Student H.: No.

Other students calling out unsolicited: no. no.

(Teacher described the chart again and repeating the conclusion made by the students).

Teacher: And it seems like, as you just said, that it doesn't make a difference.

Student H.: We… me and (partner) decided that home climate does matter.

Teacher: Ok, what evidence did you use to support that? How did you decide… How did you conclude that home climate matters?

Student H.: So, because... I mean... we have no evidence.

Teacher: So you use your opinion to say that… you don't have any evidence to support that, right, actually the evidence contradicts what you are saying, right? And that's ok. We made predictions and then we try to go to the data to see if it's right or wrong….

Teacher: So does anyone want to add anything here.

Student B.: At first I had a B, now, I think he got an A because he had cold home climate. But, home climate doesn't matter. So, I think he got an A.

In the above discussion, Student H lacked the metacognition to put away his beliefs even after the teacher worked with him to review the data and that he acknowledged that data contradicted his belief. He insisted on what he believed was true and only came to the realization of the weakness of his claim when the teacher challenged him, which led him to finally recognize that he had no evidence to support his claim, requiring a more advanced epistemological thinking. At the same time, Student B observed the discussion and learned of her own mistake and corrected her solution. This type of development through observations of others being scaffolded is consistent with previous findings that observations of students being tutored also helped the observers learn effectively especially when the observers were afforded with the opportunity to collaboratively solve similar problems by interacting with a partner (Chi, Roy, &

Hausmann, 2008). The missing class discussions in the tech-only condition may have robbed the students the chance to this kind of overall reflections and practice in a social environment.

Encouraging better self-regulation

As the teacher circulated the room and observed student pairs interacting with the program, if the teacher saw that students were struggling, the teacher might redirect students to re-read the prompts or to read them aloud if the students appeared not to have comprehended them correctly or have repeatedly misread them. There is also the challenge of repetition. Some lower performing students expressed the view that the questions were repetitive. In fact, they hadn't read the prompts carefully. For example, a pair of students showed frustration, saying "we answerd this like 3 times already". They did not realize that the cases they chose to compare were different each time. One time, they chose a pair of cases reflecting an uncontrolled comparison; the second time, they chose a pair of cases with unvarying levels for all factors, and then the third time, they chose a pair of cases with the investigated factor unvarying. Although the automated-agent tried to guide the students differently each time, the students perceived the prompts as repetitions.

If students perceive the teacher to be nearby, concerned with everyone's progress, they may have greater motivation to continue to think deeply. Although the automated-agent was designed to detect situations where students needed extra prompting and provide them accordingly, by verbally rephrasing the questions, a live teacher may have helped alleviate frustration and reassure students that the issue at hand was worth attending to. Some eacher below comments support this interpretation:

My quick learners got a little bored with the repetitive nature of the task - because they had 4 variables, they had to go through the same procedure 4 times. I'm not *sure how one could fix or change this though. My slower workers also go a little frustrated, partly, I think because once they were finished with the first attribute, they didn't realize they would need to go through all 4. Sustained concentration like that is not easy for some...*

Repetition was the biggest one[challenge], though some did not realize that they were having to repeat because they were getting the wrong answer/not understanding the concept. Knowing when you don't know something is hard for them.

Improving collaboration

During the class discussions, while those who answered questions profited from directly interacting with the teacher by explaining their solutions, students who observed these interactions may have learned from modeling, both with respect to how the problem was solved but also in how the teacher asked questions. The same was true when teachers interacted directly with individual pairs. This kind of modeling may have helped students to become better at collaborating to solve the problems at hand. Even if students are talking to each other and are asking each other questions, it does not mean that they are asking good questions (Choi, Land, & Turgeon, 2005). Besides asking good questions, more teacher engagement may also help encourage students to better work together. This was one of the challenges noted by a teacher:

The greatest challenge is ensuring that the partners were indeed working together and one student was not taking over.

As the teacher challenged students through class discussions and while interacting with student pairs, these verbal interactions enable the teacher to promote better argumentation by reinforcing the types of argumentative reasoning exhibited by the automated-agent. For example, one student pair repeatedly provided no counterarguments when the automated-agent challenged them (by asking, "Suppose someone disagrees with you… What would you say to them?"). Their answers were repeatedly, "no" and "they are wrong". While students tended to eventually provide better counterarguments toward the end of the intervention, if a teacher observed these repeated responses early on, the teacher might have been able to intervene sooner and therefore students would have had more effective practice of argumentation skills.

More teaching time

Perhaps one of the most important features of individualized scaffolding through automated-agent is to allow teachers to better utilize their time in the classroom. When the teacher was actively engaged in the intervention, students not only learned from the activities and interacting with the automated-agent, they were also afforded more valuable time with the teacher. Besides the potential benefits of better self-regulation and better collaboration and argumentation, other social benefits of interacting with the teacher may have also indirectly encouraged learning. Below feedback from a teacher showed how the teacher appreciated more time with the students:

Yes. Roaming is always necessary, but this allowed me to spend more time with struggling students and know that those that were able to work independently were engaged.

Limitations and Future Work

While the comparisons between the two experimental conditions provided useful insights in how technology may be employed in a classroom, given the limitation of a quasi-experimental design, it is not possible to draw definitive conclusion that the differences in posttest performances were due to the differences between the experimental conditions. Without a randomized controlled experiment, no causal relations can be inferred. To minimize the impact of this limitation, the school involved in this study has the policy to assign students so that two sixth grade classrooms have similar demographics and performance levels; and that they share the same teacher for each subject. In addition, COV pretest comparison was made for the two classrooms, which showed that there was no significant difference between the average COV pretest scores between the two classrooms.

As described in the feedbacks from the teachers, perceived repetition is one of the biggest weakness in the technology that potentially results in disengagement from the students. Improving the scaffolding to detect boredom and that students may be perceiving repetition in order to more explicitly direct students to the flaws of their solutions may help reduce confusions that students experienced by feeling that they were asked to do the same thing over and over again.

In addition, improving the technology by adding feature to help teachers easily review students' progress or to highlight students who may be struggling could potentially help direct teachers' attention to students in need. This type of features also may help teacher hold better class discussions to address struggles that are common among students.

As pointed out the feedback from the teacher, sometimes one student in the student pair may take over the task or that one student may be disengaged and therefore resulting in a lack of collaboration. Finding ways to improve the technology so that it better reinforces more balanced participations between the pairs and helping them learn how to ask better questions may help students achieve better collaborations.

Because the technology captures all of the input from the students, there are potentials in further analyzing students' responses that can help better understand how students progress through the technology. Such analyses can help discover ways that can improve the intervention.

In addition, the results showed encouragement in how intervention of this kind potentially helps students regulate against their beliefs. Additional studies with assessment tasks that present students with information in which students have higher stakes in their beliefs may help better gauge the level of effectiveness in helping students regulate against biases that they hold more dearly than neutral topics such as life expectancy.

Conclusion

The encouraging results from this study showed that technology can help make personalized scaffolding more scalable. Argumentative scaffolding such as the type used in this intervention, when delivered just at the right time, can help students develop beyond narrow procedural knowledge and instead acquire more advanced epistemological thinking and metacognitive skills that are needed to solve more complex problems. In an era when information of all levels of qualities are over abundantly available, the ability to put aside belief bias to evaluate evidence is an important skill to becoming a productive citizen.

As technologies become more powerful, our results provide insight into how technology can be effectively integrated into classrooms. Made clear here is that while technology can be helpful, it merely complements teachers' teaching. Social context afforded by whole-class discussions is an important factor that cannot be easily omitted. Active engagement from a teacher was not only more effective but was essential to the success of learning. Technology is a tool that allows teachers to spend more valuable time with students. Technology affords teachers with more chances to have meaningful interactions with the whole class and individually. Technology may also help capture data that can help us better understand how teachers can improve their teaching.

References

Adamson, D., & Rosé, C. (2012). Coordinating multi-dimensional support in collaborative conversational agents. In *Proceedings of intelligent tutoring systems* (pp. 346–351). Berlin: Springer.

Arvidsson, T. S., & Kuhn, D. (2016). Introducing Academically Low-Performing Young Science Students to Practices of Science. In *Proceedings of International Conference of the Learning Sciences*, Singapore (pp. 82-89).

Azevedo, R., & Hadwin, A. F. (2005). Scaffolding self-regulated learning and metacognition–Implications for the design of computer-based scaffolds. *Instructional Science*, *33*(5), 367-379.

Blank, L. M. (2000). A metacognitive learning cycle: A better warranty for student understanding? *Science Education*, *84*(4), 486-506.

Bromme, R., Kienhues, D., & Stahl, E. (2008). Knowledge and epistemological beliefs: An intimate but complicate relationship. In *Knowing, knowledge and beliefs* (pp. 423-441). Springer Netherlands.

Bromme, R., Pieschl, S., & Stahl, E. (2010). Epistemological beliefs are standards for adaptive learning: a functional theory about epistemological beliefs and metacognition. *Metacognition and Learning*, *5*(1), 7-26.

Brusatte, S. L., Butler, R. J., Barrett, P. M., Carrano, M. T., Evans, D. C., Lloyd, G. T., Mannion, P. D., Norell, M. A., Peppe, D. J., Upchurch, P., & Williamson, T. E. (2015). The extinction of the dinosaurs. *Biological Reviews*, *90*(2), 628-642.

Cano, F. (2005). Epistemological beliefs and approaches to learning: Their change through secondary school and their influence on academic performance. *British Journal of Educational Psychology*, *75*(2), 203-221.

Chi, M. T., Roy, M., & Hausmann, R. G. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive science, 32*(2), 301-341.

Chi, M. T., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*(4), 471-533.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219-243.

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

Choi, I., Land, S. M., & Turgeon, A. J. (2005). Scaffolding peer-questioning strategies to facilitate metacognition during online small group discussion. *Instructional science*, *33*(5-6), 483-511.

Conley, A. M., Pintrich, P. R., Vekiri, I., & Harrison, D. (2004). Changes in epistemological beliefs in elementary science students. *Contemporary educational psychology*, *29*(2), 186-204.

Cozzarelli, C., Wilkinson, A. V., & Tagler, M. J. (2001). Attitudes toward the poor and attributions for poverty. *Journal of Social Issues, 57*(2), 207-227.

Dieterle, E., & Clarke, J. (2007). Multi-user virtual environments for teaching and learning. *Encyclopedia of multimedia technology and networking*, *2*, 1033-44.

Dinsmore, D. L., Alexander, P. A., & Loughlin, S. M. (2008) Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. *Educational Psychology Review, 20*(4), 391-409.

Evans, J. S. B. (2003). In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences, 7*(10), 454-459.

Fazey, I. R. A. (2010). Resilience and higher order thinking. *Ecology and Society*.

Felton, M. K., & Kuhn, D. (2007). "How Do I Know?" The Epistemological Roots of Critical Thinking. *Journal of Museum Education, 32*(2), 101-110.

Fox, E., & Roconscente, M. (2008). Metacognition and self-regulation in James, Piaget, and Vygotsky. *Educational Psychology Review, 20*(4), 373–389.

Garrison, D. R., & Akyol, Z. (2013). Toward the development of a metacognition construct for communities of inquiry. *The Internet and Higher Education, 17*, 84-89.

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

Graesser, A. (2015). Authoring conversation-based tutors. In R. Sottilare, A. Graesser, X. Hu, & K. Brawner (Eds.), *Design Recommendations for Intelligent Tutoring Systems: Authoring Tools and Expert Modeling Techniques.* Vol. 3. Orlando, FL: U.S. Army Research Laboratory. p. 195-198.

Grotzer, T. A., Derbiszewska, K., & Solis, S. L. (2017). Leveraging fourth and sixth graders' experiences to reveal understanding of the forms and features of distributed causality. *Cognition and Instruction, 35*(1), 55-87.

Gunstone, R., & Mitchell, I. J. (1998). Metacognition and conceptual change. In J. L. Mintzes, J. H. Wandersee, & J. D. Noval (Eds.), *Teaching for science education: A human constructivist view* (pp. 133–163). San Diego, CA: Academic Press.

Hadwin, A. F., Wozney, L., & Pontin, O. (2005). Scaffolding the appropriation of self-regulatory activity: A socio-cultural analysis of changes in teacher–student discourse about a graduate research portfolio. *Instructional Science*, *33*(5-6), 413-450.

Hennessey, M. G. (1999). *Probing the dimensions of metacognition: Implications for conceptual change teaching-learning.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, Boston, MA.

Hofer, B. K. (2001). Personal epistemology research: Implications for learning and teaching. *Educational Psychology Review, 13*(4), 353-383.

Hofer, B. K. (2008). Personal epistemology and culture. In *Knowing, Knowledge and Beliefs* (pp. 3-22). Springer Netherlands.

Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Review of educational research*, *67*(1), 88-140.

Hogan, K., & Thomas, D. (2001). Cognitive comparisons of students' systems modeling in ecology. *Journal of Science Education and Technology*, *10*(4), 319-345.

Iiskala, T., Vauras, M., Lehtinen, E., & Salonen, P. (2011). Socially shared metacognition of dyads of pupils in collaborative mathematical problem-solving processes. *Learning and Instruction, 21*(3), 379-393.

Jewett, E., & Kuhn, D. (2016). Social science as a tool in developing scientific thinking skills in underserved, low-achieving urban students. *Journal of experimental child psychology*, *143*, 154-161.

Kienhues, D., Bromme, R., & Stahl, E. (2008). Changing epistemological beliefs: The unexpected impact of a short‐term intervention. *British Journal of Educational Psychology*, *78*(4), 545-565.

King, P. M., & Kitchener, K. S. (2004). Reflective judgment: Theory and research on the development of epistemic assumptions through adulthood. *Educational psychologist, 39*(1), 5-18.

Kuhn, D. (2016). What Do Young Science Students Need to Learn About Variables?. *Science Education*, *100*(2), 392-403.

Kuhn, D., Arvidsson, T. S., Lesperance, R., & Corprew R. (2017). Can engaging in science practices promote deep understanding of them? Science Education, 101(2), 232-250.

Kuhn, D., & Dean, Jr, D. (2004). Metacognition: A bridge between cognitive psychology and educational practice. *Theory into practice, 43*(4), 268-273.

Kuhn, D., Garcia-Mila, M, & Zohar, A. (1995) Strategies Of Knowledge Acquisition. Monographs of the Society for Research in Child Development, 60(4).

Kuhn, D., & Pease, M. (2008). What needs to develop in the development of inquiry skills? *Cognition and Instruction, 26*(4), 512-559.

Kuhn, D., Ramsey, S., & Arvidsson, T. S. (2015). Developing multivariable thinkers. *Cognitive Development*, *35*, 92-110.

Laal, M., & Ghodsi, S. M. (2012). Benefits of collaborative learning. *Procedia-Social and Behavioral Sciences, 31*, 486-490.

Lai, E. R. (2011). Metacognition: A literature review. *Always Learning: Pearson Research Report.*

Lajoie, S. P., & Lu, J. (2012). Supporting collaboration with technology: Does shared cognition lead to co-regulation in medicine. *Metacognition and Learning, 7*(1), 45–62.

Lehrer, R., & Schauble, L. (2006). Scientific thinking and scientific literacy: Supporting development in learning contexts. In K. A. Renninger & I. Sigel (Vol. Eds.) & W. Damon (Series Ed.), *Handbook of Child Psychology*. Vol. 4. Hoboken, NJ: Wiley. (6th ed.)

Lehrer, R., & Schauble, L. (2015). The development of scientific thinking. In L. Liben (Vol. Ed.) & R. Lerner (Series Ed.), *Handbook of Child Psychology and Developmental Science*. Vol. 2. Hoboken, NJ: Wiley. (7TH ed.)

Liu, S. Y., Lin, C. S., & Tsai, C. C. (2010). College students' scientific epistemological views and thinking patterns in socioscientific decision making. *Science Education*, *95*(3), 497-517.

Livingston, J. A. (2003). *Metacognition: An Overview.*

Lorch Jr, R. F., Lorch, E. P., Calderhead, W. J., Dunlap, E. E., Hodell, E. C., & Freer, B. D. (2010). Learning the control of variables strategy in higher and lower achieving classrooms: Contributions of explicit instruction and experimentation. *Journal of Educational Psychology*, *102(1),* 90-101.

Loyens, S. M., Magda, J., & Rikers, R. M. (2008). Self-directed learning in problem-based learning and its relationships with self-regulated learning. *Educational Psychology Review, 20*(4), 411-427.

Magnisalis, I., Demetriadis, S., & Karakostas, A. (2011). Adaptive and intelligent systems for collaborative learning support: a review of the field. *IEEE Transactions on Learning Technologies*, *4*(1), 5-20.

Mason, L. (2003). Personal epistemologies and intentional conceptual change. In G. M. Sinatra & P. R. Pintrich (Eds.), *Intentional Conceptual Change*. Mahwah, NJ: Lawrence Erlbaum Associates. p. 199-236.

Mason, L., & Scirica, F. (2006). Prediction of students' argumentation skills about controversial topics by epistemological understanding. *Learning and Instruction, 16*(5), 492-509.

Osborne, J. (2014). Teaching scientific practices: Meeting the challenge of change. *Journal of Science Teacher Education*, *25*(2), 177-196.

Pareto, L. (2014). A teachable agent game engaging primary school children to learn arithmetic concepts and reasoning. *International Journal of Artificial Intelligence in Education*, *24*(3), 251-283.

Phan, H. P. (2008). Predicting change in epistemological beliefs, reflective thinking and learning styles: A longitudinal study. *British journal of educational psychology*, *78*(1), 75-93.

Phelan, S. M., Burgess, D. J., Burke, S. E., Przedworski, J. M., Dovidio, J. F., Hardeman, R., Morris, M. & van Ryn, M. (2015). Beliefs about the causes of obesity in a national sample of 4th year medical students. *Patient education and counseling*, *98*(11), 1446-1449.

Pintrich, P. R., Marx, R. W., & Boyle, R. A. (1993). Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Review of Educational research*, *63*(2), 167-199.

Pintrich, P. R., Wolters, C. A., & Baxter, G. P. (2000). Assessing Metacognition and Self-Regulated Learning. In S. Gregory & J. C. Impara (Eds.), *Issues in the Measurement of Metacognition* (pp. 43-97). Lincoln, NE: Buros Institute of Mental Measurements.

Puntambekar, S., & Stylianou, A. (2005). Designing navigation support in hypertext systems based on navigation patterns. *Instructional Science*, *33*(5-6), 451-481.

Saab, N., van Joolingen, W., & van Hout-Wolters, B. (2012). Support of the collaborative inquiry learning process: influence of support on task and team regulation. *Metacognition and learning*, *7*(1), 7-23.

Sandoval, W. (2014). Science education's need for a theory of epistemological development. *Science Education, 98,* 383-387.

Sao Pedro, M. A., Gobert, J. D., & Baker, R. S. (2014). The impacts of automatic scaffolding on students' acquisition of data collection inquiry skills. Roundtable presentation at *American Educational Research Association*.

Savery, J. R. (2015). Overview of problem-based learning: Definitions and distinctions. In A. Walker, H. Leary, C. E. Hmelo-Silver, & P. A. Ertmer (Eds.), *Essential readings in problem-based learning: Exploring and extending the legacy of Howard S. Barrows*. West Lafayette, Indiana: Purdue University Press. p. 5-15.

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education*, *36*(1-2), 111-139.

Schwartz, D. L., Blair, K. P., Biswas, G., Leelawong, K., & Davis, J. (2007). Animations of thought: Interactivity in the teachable agent paradigm. In R. Lowe &W. Schnotz (Eds.), Learning with animation: *Research and implications for design* (Vol. 18, pp. 334–352). UK: Cambridge University Press.

Sharpe, D. (2015). Your chi-square test is statistically significant: Now what?. *Practical Assessment, Research & Evaluation*, 20.

Sinatra, G. M. (2005). The “warming trend” in conceptual change research: The legacy of Paul R. Pintrich. *Educational Psychologist, 40*(2), 107–115.

Stathopoulou, C., & Vosniadou, S. (2007). Exploring the relationship between physics-related epistemological beliefs and physics understanding. *Contemporary Educational Psychology, 32*(3), 255-281.

Sungur, S., & Tekkaya, C. (2006). Effects of problem-based learning and traditional instruction on self-regulated learning. *The Journal of Educational Research, 99*, 307–317.

Tarricone, P. (2011). *The Taxonomy of Metacognition*. NY: Psychology Press.

Tsui, L. 1999. Courses and instruction affecting critical thinking. *Research in Higher Education, 40*(2),185-200.

Volet, S., Vauras, M., & Salonen, P. (2009). Psychological and social nature of self- and co-regulation in learning contexts: an integrative perspective. *Educational Psychologist, 44*(4), 215-226.

Wentzel, K. R., & Watkins, D. E. (2002). Peer relationsips and collaborative learning as contexts for academic enablers. *School Psychology Review, 31*(3), 366.

Winne, P. H., Hadwin, A. F., & Perry, N. E. (2013). Metacognition and computer-supported collaborative learning. *International handbook of collaborative learning*, 462-479.

Winters, F. I., Greene, J. A., & Costich, C. M. (2008). Self-regulation of learning within computer-based learning environments: A critical analysis. *Educational Psychology Review, 20*(4), 429-444.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into practice*, *41*(2), 64-70.

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

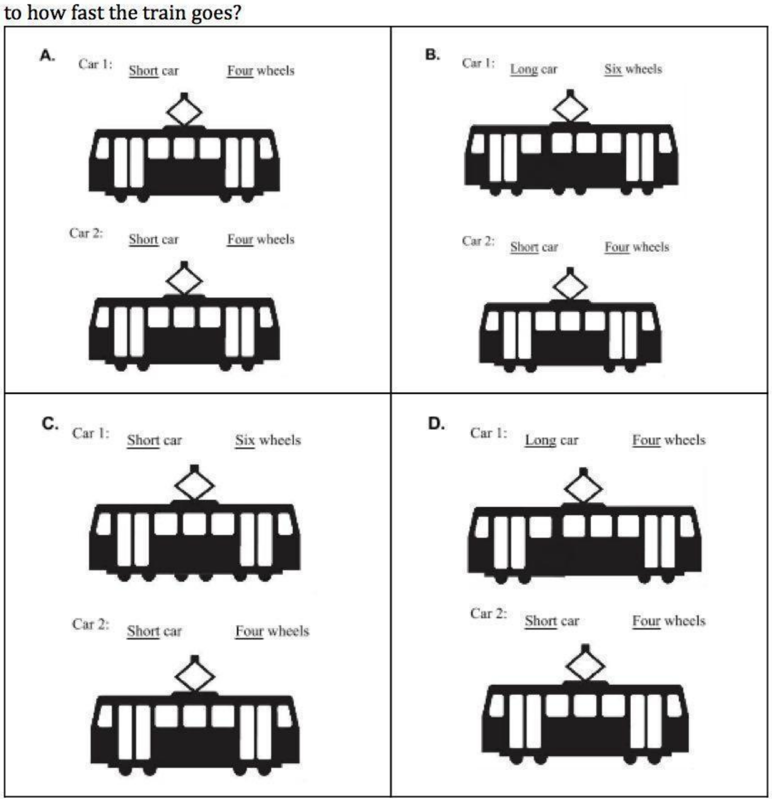
Zimring, F. E. (2006). *The great American crime decline*. Oxford University Press, USA.

Appendix A - COV Posttest

Name: ____________________________ **Class:** ______________ **Date:** _____________________

1. The city of New York wants to get new trains for the subway. They can make trains with **short or long** Car Size and they can make trains with **four or six** Number of Wheels.

Which two trains should they build and test if they want to know if Car size makes a difference to how fast the train goes?



2. Why did you choose that set of trains?

A. The two trains are completely different

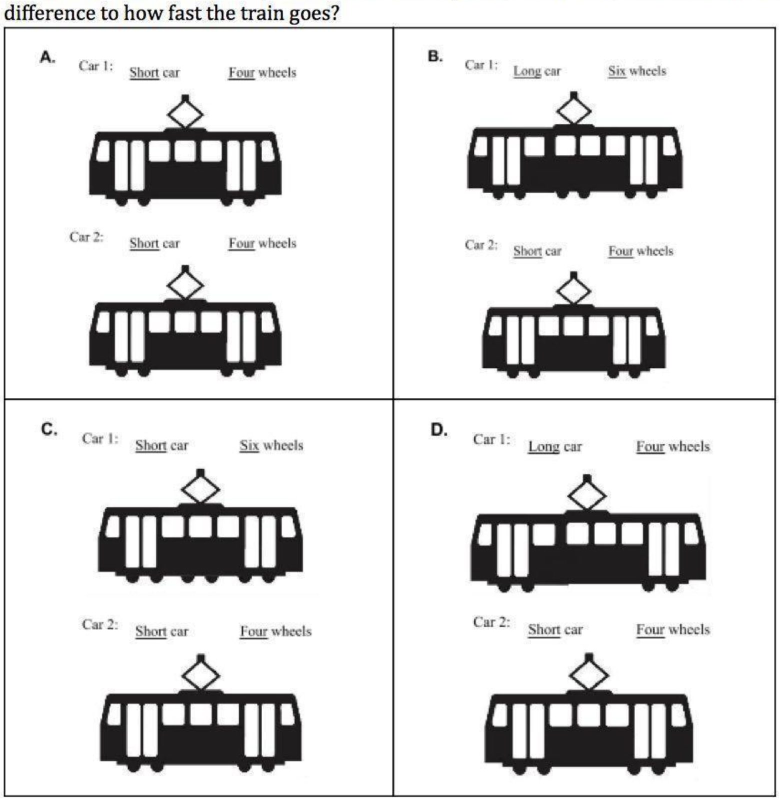
B. The two trains are completely different except for car size

C. The two trains are completely the same

D. The two trains are completely the same except for car size

3. The city of New York wants to get new trains for the subway. They can make trains with **short or long** Car Size and they can make trains with **four or six** Number of Wheels.

Which two trains should they build and test if they want to know if Number of Wheels makes a difference to how fast the train goes?



4. Why did you choose that set of trains?

A. The two trains are completely different

B. The two trains are completely different except for number of wheels

C. The two trains are completely the same

D. The two trains are completely the same except for number of wheels

5. A sailor wants to make a new boat for a race. The sailor can make boats with **short or long** Size of Body, with **one or two** Number of Sails and make them with **big or small** Size of Sail.

Which two boats should the sailor build and test if the sailor wants to know if Size of Body makes a difference to how fast the boat goes?

A. Boat 1: Short body Two sails Small sail

Boat 2: Short body Two sails Small sail

B. Boat 1: Short body Two sails Small sail

Boat 2: Short body Two sails Big sail

C. Boat 1: Short body Two sails Big sail

Boat 2: Long body One sail Small sail

D. Boat 1: Short body Two sails Big sail

Boat 2: Long body Two sails Small sail

E. Boat 1: Short body Two sails Small sail

Boat 2: Long body Two sails Small sail

3

Appendix B - LE Posttest

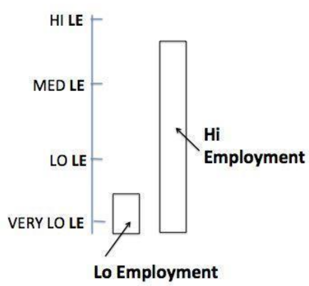
Tear this page

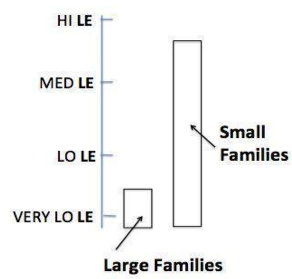
Some people live very long lives. Others die at an early age. What makes the difference? Life Expectancy (**LE**, for short) is the term for how long people on average are expected to live. **LE** differs greatly across different countries. Some countries have a much higher average **LE** than others. What causes the differences? Here are some possibilities that studies suggest.

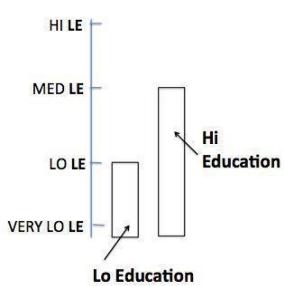
One is employment. As the chart below shows, countries where people have high employment levels have a higher average life expectancy than countries where employment is low (many people are without jobs). As you see, the chart shows that employment makes a very big difference to **LE.**

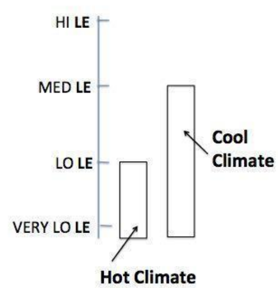
The other charts show other things that can affect **LE.** As you see, family size, like employment, makes a very big difference. Smaller families on average mean higher **LE.**

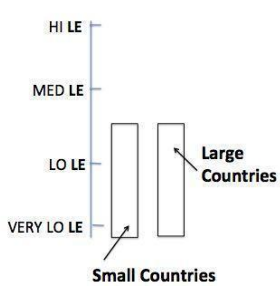
Other things, like education and climate, make a smaller difference. And some, like country size, seem to make no difference at all - **LE** overall is about the same for small and large countries.











Now your task is to make some predictions about **LE** for different countries. You can look back at these charts when you want to. For each country, predict the average **LE** you think that country will have.

5

Which of these do you need to know to predict the average **LE** for **Country 1**?
(circle one or more)
Employment Family size Education
Climate Country size

Country 1:
(Cross out what you do not need to make your prediction)
Employment: HI
Family size: SMALL
Education: HI
Climate: HOT
Country size: SMALL

Circle the average **LE** you expect for this country:
VERY LO LO MEDIUM HI

Which of these do you need to know to predict the average **LE** for **Country 2**?
(circle one or more)
Employment Family size Education
Climate Country size

Country 2:
(Cross out what you do not need to make your prediction)
Employment: LO
Family size: LARGE
Education: LO
Climate: COOL
Country size: LARGE

Circle the average **LE** you expect for this country:
VERY LO LO MEDIUM HI

Which of these do you need to know to predict the average **LE** for **Country 3**?
(circle one or more)
Employment Family size Education
Climate Country size

Country 3:
(Cross out what you do not need to make your prediction)
Employment: LO
Family size: SMALL
Education: LO
Climate HOT
Country size SMALL

Circle the average **LE** you expect for this country:
VERY LO LO MEDIUM HI

Country 4: Employment LO, Family size SMALL, Education HI, Climate COOL, Country size SMALL

Circle the average **LE** you expect for this country:
VERY LO LO MEDIUM HI

Which of the five things mattered to your prediction?
(circle one or more)
Employment Family size Education
Climate Country size

Country 5: Employment HI, Family size LARGE, Education LO, Climate HOT, Country size LARGE

Circle the average **LE** you expect for this country:
VERY LO LO MEDIUM HI

Which of the five things mattered to your prediction?
(circle one or more)
Employment Family size Education
Climate Country size

Country 6: Employment HI, Family size SMALL, Education LO, Climate COOL, Country size SMALL

Circle the average **LE** you expect for this country:
VERY LO LO MEDIUM HI

Which of the five things mattered to your prediction?
(circle one or more)
Employment Family size Education
Climate Country size

Country 7: Employment LO, Family size LARGE, Education HI, Climate HOT, Country size LARGE

Circle the average **LE** you expect for this country:
VERY LO LO MEDIUM HI

Which of the five things mattered to your prediction?
(circle one or more)
Employment Family size Education
Climate Country size

Country 8: Employment HI, Family size LARGE, Education HI, Climate HOT, Country size LARGE

Circle the average **LE** you expect for this country:
VERY LO LO MEDIUM HI

Which of the five things mattered to your prediction?
(circle one or more)
Employment Family size Education
Climate Country size

Country 9: Employment HI, Family size LARGE, Education HI, Climate HOT, Country size SMALL

Circle the average **LE** you expect for this country:
VERY LO LO MEDIUM HI

Which of the five things mattered to your prediction?
(circle one or more)
Employment Family size Education
Climate Country size