EFFECTS OF INNOVATION VERSUS EFFICIENCY TASKS ON COLLABORATION AND LEARNING

A DISSERTATION

SUBMITTED TO THE SCHOOL OF EDUCATION

AND THE COMMITTEE ON GRADUATE STUDIES

OF STANFORD UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

David Andrew Sears

June 2006

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

_____

Daniel Schwartz (Principal Adviser)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

_____

Brigid Barron

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

_____

Edward Haertel

Approved for the University Committee on Graduate Studies.

ABSTRACT

What makes for a naturally productive collaborative task? Some researchers have suggested that optimal tasks for productive collaboration are ill-structured and allow for exploration and construction of multiple possible solutions (e.g. Cohen, 1994). Others have suggested that tasks should have one solution and be well-defined such that everyone can agree on their answers (e.g. Steiner, 1972). In a search for a way to reconcile this dilemma, two dimensions—innovation and efficiency—were examined for their effects on collaboration and learning in two experiments with university students.

Innovation involves the use of prior knowledge to construct solutions to unfamiliar problems. The goal is to prepare students to perceive and appreciate how an expert solution works when they receive instruction on it. Efficiency involves being given the canonical solution and then having an opportunity to practice it. The goal of efficiency is to promote speed and accuracy in applying the expert solution.

These dimensions were recently found to be informative to the field of transfer. Transfer is the generalization of learning from one situation to another. Schwartz, Bransford, and Sears (2005) suggested that optimal instruction for promoting transfer involves cycles of innovation and efficiency, rather than just one or the other approach. Thus, rather than viewing these dimensions as polar opposites, they described them as complementary components for promoting thorough understanding.

For the two experiments reported in this dissertation, it was hypothesized that tasks with an innovation component would yield more productive interactions and learning than tasks with strictly efficiency components. The first experiment compared dyads working on an Innovation version of a concept-mapping task to dyads working on an Efficiency version of that task. It was an exploratory study designed to be an initial test of the Innovation and Efficiency framework. While it found few significant learning differences between conditions, it revealed that the Innovation task promoted more knowledge-sharing behaviors than the Efficiency task, as expected. Through a novel method of analysis, moment-to-moment interactions were related to learning outcomes.

The second experiment built upon the findings of the first. Individuals and dyads were randomly assigned to the Innovation condition or the Efficiency condition.

Participants learned about the chi-square formula, and their understanding of it was assessed with basic calculation questions, comprehension questions, and difficult transfer problems. As part of the transfer problems, a preparation for future learning (PFL) assessment was used to measure participants' ability to adapt their knowledge of the chi-square formula (Bransford & Schwartz, 1999).

PFL assessments include a resource question and a target question. The resource introduces a new type of problem that is related to the initial instruction. The target builds upon the resource. If instructional conditions vary in their effects on students' abilities to learn from the resource, these difference should appear on the target problem. Participants in the Innovation condition scored significantly higher on the transfer problems, and Innovation dyads showed the greatest performance on the target PFL question. The strongest indicator that tasks with innovation components might naturally support collaborative learning came from the finding that Innovation dyads exceeded nominal dyads (dyads modeled on individuals' scores) on the PFL problem.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dan Schwartz, for his immense contribution to my graduate education. It seems that one defining feature of a mentor is his or her ability to impact multiple facets of one's life. The lessons I learned from Dan have certainly made him a mentor to me. All of his students can joke about his appreciation of the 2 x 2 experimental design—yes, that's lesson number one. More important was his encouragement to pursue my interests, his allowance for failure, and his periodic challenges. All of these different kinds of support helped me learn to handle initial defeat, find the interesting phenomena, and trust my developing instincts. As he says, the results will come. These are lessons that have been an invaluable part of my education and will extend well beyond the classroom and laboratory. Thanks Dan! I hope I will someday be as helpful to my students as you have been to me.

I would also like to thank Brigid Barron for starting me on the path of analyzing collaborative group discourse and interaction. Even in my first quarter of graduate school, she allowed me to participate in a conference and present a poster on collaborative interaction. Thank you for making my introduction to the educational research community so exciting! Your thoughtful feedback on early drafts of my dissertation and your encouragement to speculate on what mechanisms might underlie collaborative benefits were also much appreciated. Thank you for all of your help!

Ed Haertel is another person I will cherish for greatly enhancing my Stanford experience. Always thoughtful and generous, Ed started me working with two medical doctors a few years ago on an educational project of great interest to me. He also demonstrated wonderful ways of teaching that I hope to emulate in the future. His course assessments led to insights that helped solidify a topic, his questions led to productive searches for answers, and his guidance when I was a teaching assistant for his Introduction to Test Theory course helped me really enjoy teaching. Thank you, Ed, for sharing your wisdom and kindness so generously!

To my other committee members, Shelley Goldman and Byron Reeves, thank you for coming through for me on my oral defense when the deadline was so near. Shelley, your energy is contagious, thank you. Your perspective during my defense was uniquely

valuable because of my limited background in ethnographic methods. Byron, your questions during the defense and comments subsequently were very helpful as well. Thank you for the nice stroll to the library subsequently, too!

I would also like to thank one other professor and an elementary school teacher and her third grade class for their help and expertise. Professor Bryk provided valuable insight regarding the possibility of HLM analyses. Roberta very generously invited me to conduct research on Innovation and Efficiency in her third grade class. That experience helped propel my thinking on the topic while also giving me a much better sense of how to work with a classroom full of energized students. Thank you for being so kind!

Many friends and colleagues also volunteered their time during pilot testing for this research and contributed greatly to my graduate school experience. Thank you, students of Education 350, for your time and feedback! All my friends in the AAAlab, thank you for your tremendous support! Janet, your careful inter-rater reliability coding made my day when I was in the thick of data analysis. Kevin and Kristen, your explanation and instruction on HLM calculations was the best two hour "workshop" I could imagine. Sashank, Lee, Sandra, Ugochi, Robb, Dylan, and Girija, thank you for your questions and suggestions in lab meetings, and being such a joy to be around.

I would also like to give special thanks to Jammie. Your insight on how the expected value calculation works was invaluable. Our conversations over lunch and walks back from the lab were always such a delight. Thank you! Emma, your camaraderie this year during the dissertation process and job search was such a blessing. Apparently the benefits of collaboration extend beyond learning to include perseverance even when something seems impossible. Best wishes to you as you continue your dissertation! Pieter, my great friend and roommate, you really came through for me at the end and filled the past five years at Stanford with great memories. Sharing dinners with me during the final writing phase, getting me out to the tennis court and gym to keep healthy, and always providing such great humor—thanks buddy!

Lastly, for my family, you are my foundation. Leala, thank you for periodically reminding me what it means to be a classroom teacher. More importantly, thank you for your words of wisdom, your kindness, and your courage. I love your spirit! Thank you, Mom and Dad, for always being there. This would not have been possible without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Collaborative learning has gained prominence as an instructional technique in recent years, but we still lack understanding of when and how it naturally contributes to learning. Currently, rather than being able to give students tasks to do in groups with the expectation that they will interact well and learn the material, we must use artificial scaffolds. Many teachers report using collaborative activities in their classrooms, yet few implement scaffolds found to promote successful learning in groups (Antil, Jenkins, Wayne, & Vadasy, 1998). Some of the reasons teachers give for not using proven techniques are that they are difficult or cumbersome to implement, or that they seem artificial and might not prepare students for the real world. Common techniques for promoting collaborative learning include roles and scripts, motivational systems, and training for group behaviors (e.g. Coleman, 1998; Gillies & Ashman, 1998; Johnson & Johnson, 1999; King, 1999; Slavin, 1996). While these approaches have shown notable successes, teachers often prefer using unstructured group-work. This can be problematic because classrooms that did not implement a proven system showed no significant learning gains from collaboration (Slavin, 1996). Rather than trying to structure the interactions in collaborative groups through training, rewards and punishments, or formal scripts, an alternative approach is to understand what makes a naturally productive collaborative learning activity. This would allow teachers to select appropriate tasks for groups without needing to implement extensive scaffolds to ensure learning. By searching for a way to characterize more and less productive tasks for groups, we can gain a better sense of what the benefits of collaboration are to learning and when we can expect them. This dissertation describes two experiments that tested a recently developed framework from the transfer literature for its ability to characterize productive collaborative tasks (Schwartz, Bransford, & Sears, 2005).

By examining the effects of tasks on collaborative interactions and learning, we should be able to identify features of tasks that naturally afford knowledge-sharing, joint-attention, or other productive interaction patterns. In the remainder of this dissertation, I will review the literature on collaborative learning to further justify why an investigation of the effects of different types of tasks on group interaction and individual learning is

appropriate at this juncture. This review will first describe features that categorize the literature on collaborative learning and help reveal gaps in our current knowledge. It will then note the progress made in understanding the effects of different scaffolds on collaborative interactions and learning. Next, the relative lack of research on the impact of different tasks on outcomes other than group performance will be highlighted. This section will include a detailed review of a few of the limited number of studies that have found strong effects of different tasks on group performance in an effort to describe common features of such tasks. Finally, I will describe two dimensions, innovation and efficiency, that may provide a useful characterization of tasks that support productive collaboration.

Collaborative Learning Literature Review

Before reviewing the literature, three features that characterize studies of collaboration will be described: outcome measures, social characteristics, and task effects. Common outcome measures of collaborative interaction are tests of problem solving and learning. For my purposes here, problem solving refers to the work a group does together. Learning is measured by what an individual can do alone after working in a group, and it presumably entails transfer of some sort (see Bransford & Schwartz, 1999 for why transfer might be re-conceptualized as learning).

Collaborative learning is often considered successful compared to individual learning if the average of the group performance exceeds the average individual performance on a posttest. This could be considered a lower-bound for considering group interaction successful because if groups do not exceed the average of individuals, then people would gain more by working alone. More stringent methods exist that involve comparing group performance or learning to that of mathematically aggregated individuals, known as "nominal groups." This approach has been used to see if groups perform better than even the best individuals (e.g. Laughlin et al., 2003), or better than individuals aggregated under truth-wins assumptions in which perfect knowledge-sharing is assumed (e.g. Schwartz, 1995). Under truth-wins assumptions, if any individual has the correct answer, it is assumed they would share it with the "group" and everyone would recognize its correctness. Truth-wins models represent an upper-bound for what

groups can do if each individual operated separately and then combined in perfect fashion. The only way real groups can surpass nominal groups aggregated under truth-wins assumptions is if something in their interaction leads to new knowledge construction, such as partners coming to new understanding through discussion that neither could have discovered alone (in the allotted time). Outperforming nominal groups is rare. Tasks that have shown such results may contain key features of productive collaborative tasks, and for this reason, the relevant literature will be reviewed in detail in the section on task-structures and collaborative learning.

Social characteristics are also a critical component of studies of collaboration. If we can identify the key behaviors associated with successful group learning, we can attempt to find more effective ways of supporting those behaviors. Studies of social interaction include those that examine the types of behaviors associated with productive group or individual outcomes (usually studies involving video or discourse analysis). For instance, Webb (1983) found that giving explanations was associated with greater learning. Other studies of social interaction manipulate how individuals interact by using social scaffolds, such as roles or prompts, or motivation systems. One recognized system employs interdependence and individual accountability to promote productive collaboration. Interdependence means that the individual's success depends (in part) on the group's success. Individual accountability means that the group's success depends on each individual. As an example, group recognition can be given based on the average test score of individuals in a group. Reviewing 99 studies that compared collaborative learning treatments to control conditions in schools ranging from the elementary level to the secondary level, Slavin (1996) found that when individual accountability and interdependence were both employed, collaborative learning produced an effect size of +.32. Without both scaffolds, collaboration only produced an average effect size of +.07.

Beyond outcome measures and social characteristics, a third factor receiving attention in the literature involves the task around which people collaborate. Studies of task effects include those that explicitly compare the effects of different tasks on collaborative outcomes (e.g. Laughlin et al., 2003) and those that test the effects of a given task on groups versus individuals (e.g. Laughlin et al., 2003; Schwartz, 1995). For example, Steiner (1972) reviewed many studies of task effects ranging from tug-of-war

and brainstorming to syllogisms and Eureka problems. The former tasks showed considerable losses compared to nominal groups whereas the latter often showed group performance equal to nominal groups (and better than individuals alone).

The research on collaborative learning can be broadly divided into studies that relate social characteristics to outcome measures and those that relate tasks to outcome measures. The former have tended to show the effects of different social scaffolds on learning whereas the latter have tended to show the effects of different types of tasks on group problem-solving. At this point, there is little or no overlap between the two. Figure 1 reflects the characterization of the literature provided above and includes studies to be reviewed in each area. What should be noticed in this figure is that relatively few studies have examined the effects of different tasks on learning, and none that I am aware of have examined the effects that different tasks have on social interaction and thereby individual performance and learning. These factors are not the only way to divide the literature but they highlight an area that has been relatively understudied and should offer insights into the unique contributions collaboration can make to learning.

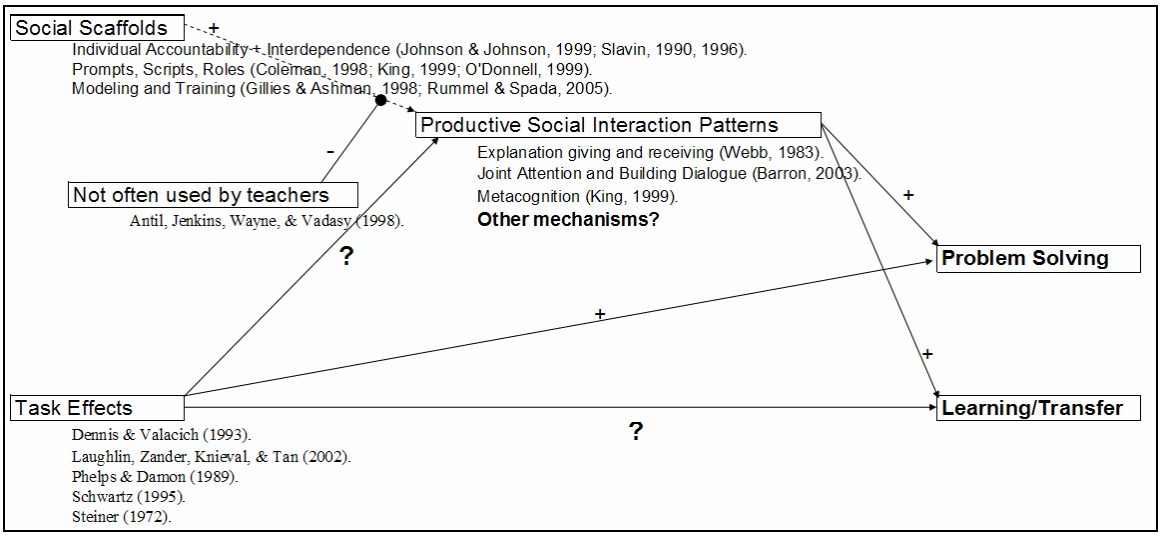

Figure 1. A concept map of the collaborative learning literature review. Social scaffolds have been shown to increase collaborative learning; however, teachers do not often implement these scaffolds. Some tasks have been shown to promote greater problem-solving success for groups compared to individuals, but less is known about their effects on group interaction patterns, and individual learning and transfer.

Studies of Social Interaction and Collaborative Learning

From the many studies of social interaction and collaborative learning, researchers generally agree that under the right scaffolds, such as individual accountability and interdependence (Slavin, 1990; Johnson & Johnson, 1999), individuals can learn more by working in groups than by working alone. Questions that remain for studies of social interaction include what processes are associated with productive learning outcomes (Azmitia, 1996; Slavin, 1996) and what underlies successful group-to-individual transfer (Olivera & Straus, 2004). As will be reviewed below, many researchers have made significant progress in understanding key mechanisms behind productive collaboration.

Webb (1983) described four studies of collaboration on mathematics units conducted with students of various ages from middle-school to upper-high school. She coded student interactions according to whether statements were given or received, requested or not, and whether they consisted of explanations, no responses, or "terminal responses" (i.e. short, yes/no types of statements). Giving explanations was correlated with students' subsequent performance, even after partialing-out a measure of their ability. Receiving explanations after requesting them was also positively correlated with individuals' subsequent performance. "Terminal responses," such as a simple "Yes/No," and non-responses were associated with worse learning.

Further evidence for the detrimental effects of terminal responses comes from Barron (2003). She studied sixth graders solving problem-based math scenarios in three-member groups. She found that groups that did poorly on the task gave more terminal and ignoring responses and did not connect their solution proposals to prior group discussion. Groups that did well on the task showed building dialogue where partners proposed correct solutions to the topics their partners' had just been discussing. Perhaps because of this joint attention and mutuality, these proposals were often accepted or discussed rather than being ignored or rejected by the other group members. Those correct proposals that did not build upon the immediately preceding topic of discussion were never accepted without further review, and they were often rejected or ignored.

Notably, Barron (2003) related these interactional-behaviors to subsequent individual recall and transfer. When individuals solved the same problem alone that they previously solved in groups, participants from high-performing groups did better than those from low-performing groups. On the transfer task the results were similar. Importantly, on average, participants who were assigned to groups outperformed those that were assigned to an individual condition (Barron, 2000). Those in the low-performing groups, however, scored lower on the original problem than those who worked alone despite scoring similarly on the mastery and transfer tasks. In other words, productive learning outcomes were not guaranteed with collaboration; knowledge-sharing and discussion appeared critical to positive outcomes.

Moving from observational studies to experimental tests, Gillies and Ashman (1998) examined the effects of manipulating the amount of certain behaviors they thought were crucial to productive collaborative learning. They trained a random sample of 25 classes of first and third graders to collaborate well with their partners by providing constructive feedback and sharing tasks and resources. Students in the trained group used higher-order cognitive language strategies such as explanations with evidence significantly more often than the untrained group. These findings are noteworthy because the trained participants also performed significantly better on a standardized word-reading test than the untrained participants.

While the training methods of Gillies and Ashman (1998) were fairly broad, others have tried more targeted interventions, such as scripts and roles, to promote greater collaboration and learning. For instance, Coleman (1998) tested the effects of explanatory prompts on fourth and fifth graders' academic achievement. In her experiment, students who were average on an "intentional learner" scale were randomly assigned to a treatment or control group. These two groups were compared to each other and to a high "intentional learning" group on achievement in a science course on photosynthesis. The treatment group received a sheet of prompts to facilitate evaluation of their thinking (a metacognitive skill) and to encourage use of high quality explanations when justifying their answers. For example, one prompt said, "Okay explain why you believe that your answer is correct or wrong?" (Coleman, 1998, p. 406).

All groups performed nearly equally on a pretest measure of photosynthesis knowledge; however, the treatment group performed significantly better than the control group at posttest (though they did not differ significantly from the "high intentional learning" group). Not only did the treatment group perform better than the control group on the posttest of photosynthesis understanding, they also performed significantly better on a conceptual-mapping and a problem-explanation task. Thus, it appears that the prompts facilitated the treatment group's learning and concept development.

Process analyses suggested that the prompts led to further discussion and ultimately to a constructed explanation in 80% of the cases where they were used. In less than 10% of cases, no further discussion occurred after the prompt was stated. Thus, Coleman (1998) found experimental support for the observations of Webb (1983) and Barron (2003) by showing that prompts for explanations and meta-cognitive awareness resulted in elaborated discussion and greater learning.

Following on Coleman's line of investigation, King (1999) found evidence for the importance of metacognition prompts on scripted collaborative learning. Specifically, her model, "ASK to THINK—TEL [sic] WHY," has the tutor ask questions and provide brief feedback and encouragement (with a primary focus on questioning). The tutee, on the other hand, could only provide explanations. Thus, in this program, the questioner starts with a Review question, to which the explainer gives a response. Then the questioner asks a Thinking question that builds on the explainer's response. The explainer responds to it and the questioner can then follow-up with a probe or hint if the explainer's answer was not complete. After the explainer gives a satisfactory answer, the questioner asks a metacognitive question, such as "how did you come up with that answer?" Again, the explainer answers, and then the two students can switch roles.

King noted that this program fosters productive discussions characterized by positive interdependence because the explainer must respond to the questioner while the questioner must tailor his or her questions to the needs of the explainer. In comparison versions of this program with and without the metacognitive question, students who had the metacognitive prompts were better at generating new knowledge and describing their thinking processes both at posttest and at a 12 week follow-up—although the statistical significance of these findings was not clear.

In sum, much of the research on collaborative interaction suggests that productive interaction comes from social scaffolds such as scripts and roles, interdependence and individual accountability, and modeled or trained behaviors. The resulting interaction is characterized by explanations, building dialogue, and active listening. Perhaps equally important, it is also characterized by few ignoring or rejecting responses. Finally, some models have suggested that educational benefits of collaboration may come from greater metacognitive activity, though it is unclear whether that is a behavior facilitated by collaboration or by scripts.

Studies of Task Effects and Collaborative Learning

Social scaffolds provide one way to promote successful collaborative learning. Unfortunately, a recent survey indicated that despite 81% of teachers reporting using collaborative learning every day in their classrooms, few of them used well-tested scaffolds in their group-work activities (Antil, Jenkins, Wayne, & Vadasy, 1998). As one teacher in the study was quoted as saying, "When I was trained [in cooperative learning] it sounded so wonderful but so complex the way they laid it out. Every kid had to have a job and they were so prescriptive. Through my teaching, I have learned that cooperative learning, for me, is just to have the kids discuss things with each other and put together a product" (p. 431). Thus, as the push for collaboration in schools increases, more demands are placed on teachers to enact appropriate scaffolds to ensure students interact productively, but these supports can be difficult to implement successfully. Without these supports, however, collaborative learning has shown little educational benefit (Slavin, 1996). An alternative approach involves finding tasks that naturally pull for productive collaboration. Researchers taking this approach have found some tasks that have produced outstanding group problem solving results, but none that I am aware of have looked at the effects on interaction patterns and subsequent individual learning.

Schwartz (1995) compared dyads to individuals on rule induction or abstract representation tasks. In the first of three experiments, Schwartz had high school students predict the direction of rotation of the final gear in a sequence of gears. A much greater percentage of dyads discovered a parity rule (e.g. if it is an odd number gear it rotates clockwise) for solving the gear problems than individuals (58% vs. 14%). The dyad

performance even exceeded what one would predict using a truth-wins model to predict the performance of nominal groups.

What was it about the task that allowed dyads to perform so much better than individuals? Schwartz's (1995) observational data indicated that students tended to represent the gears with their hands initially to make their predictions. As the number of gears in the problems increased, this strategy would become more time consuming because each gear in sequence would have to be modeled until the last gear was reached. Apparently this did not bother individuals tremendously because very few switched to a parity-rule conception of the problem. Dyads, however, had to come to consensus about their answer, so they had to communicate effectively about their predictions. Again observations indicated that partners discussed which gear in the sequence their partner was modeling with their hands, leading to an enumeration strategy. Schwartz provided evidence that this enumeration of the gears was likely the source of the dyads much greater induction of a parity rule for solving the gear problems.

Only in communicating with a partner were the need and the mechanism (i.e. enumeration) for finding a parity rule realized at a relatively high rate. In other words, a shift of the conceptualization of the problem from a modeling situation to a rule-based situation resulted in group performance that exceeded most-competent-member assumptions (and even truth-wins assumptions). The use of dyads seemed to provide sufficient motivation to promote this shift in problem conceptualization from physical modeling to enumeration.

Steiner (1972) divided tasks along a number of dimensions including whether they were "divisible" (multi-step) tasks versus "unitary" (single-step) tasks (p. 15). Unlike social-interaction researchers who often focus on how participants' knowledge sharing may promote their partner's development as well as their own, Steiner suggested that tasks will best be accomplished by groups when group members recognize who is the best individual to solve the problem and follow that person's lead. If the task involves multiple specializations, then letting the best member in each area solve that sub-problem should be optimal.

According to Steiner, "process loss" occurs when group members interfere with each other's efforts such that the best-member's correct answer is not followed by the

group. In other words, something in the group interaction, whether acquiescence by a correct member, dominance by an incorrect member, social loafing, or some other effect, interferes with the group recognizing and producing the correct answer (Chiang & Guo, 1999). Steiner (1972) placed less emphasis on "process gain," the phenomenon where groups exceed the performance of nominal groups, and this may be because so few examples exist (the study by Schwartz above is a notable exception). Thus, a goal for Steiner was to characterize tasks that showed little process loss, and he suggested that problems for which the correct answer could be clearly demonstrated, such as insight or "Eureka" problems, were less likely to show process loss than tasks where the correct answer was less visible, and thus more susceptible to social rejection.

Some evidence supporting the potential importance of demonstrability comes from contemporaries of Steiner. Lamm and Trommsdorf (1973) reviewed studies of brainstorming, a task that is notable for its lack of demonstrability due to its emphasis on generation without critique, and found that groups typically produced fewer ideas than individuals who brainstormed alone and then combined their answers. These disappointing results are likely due to process loss variables such as production blocking in which one member must hold on to her idea and forestall further production while others report their ideas to the group (Dennis & Valacich, 1993).

In an impressive reversal of such results, Dennis and Valacich (1993) used technology to produce beneficial results for real groups compared to nominal groups on a brainstorming task. For their experiment, half of the participants worked alone on two 15-min. brainstorms using paper and pencil to record their ideas. The other half used a computer program designed specifically to enhance group benefits and decrease potential pitfalls.

According to Dennis and Valacich (1993), the software limited production blocking by making all communication electronic, simultaneous, and visible/review-able, so that each individual could participate as they generated new ideas rather than having to wait for their partners. Evaluation apprehension is a second pitfall that occurs when someone feels too shy to participate because their ideas may be criticized. The software made all communication anonymous to reduce evaluation apprehension. Free riding or social loafing is another common pitfall that can occur when people feel that their

contributions are not needed or wanted. The software did not address this component. To enhance potential benefits of collaboration, the software made partners' contributions visible (though anonymous). This feature meant that participants could more easily avoid redundancy, and they could read what their partners said to get new insights if they started to run out of ideas.

Dennis and Valacich (1993) found that large collaborative groups (12 members as opposed to six members) showed significantly more unique ideas per group and per individual than 12-member nominal groups. Informal observations suggested that those in the software condition took advantage of being able to check their partner's ideas when running out of their own (e.g. some laughed while reading, suggesting an unexpected idea). In addition, the 12-member electronic groups showed significantly less redundancy than the 12-member nominal groups (13 vs. 48 redundancies on average). Thus, by structuring the activity to block process loss and enhance process gain (especially by introducing more potential unique ideas by increasing group size), they found a way to turn brainstorming into a productive collaborative activity. Dennis and Valacich's study is particularly impressive because they were able to promote process gain on a task that typically incurs a net process loss, and they did it by systematically attempting to block sources of loss while enhancing sources of gain in participants' interactions. This provides an exemplary model of when and how scaffolds can be used to transform an otherwise individualistic task into a productive group task. The features of this experiment that seem particularly relevant here are that the task was productive for groups possibly because: 1) sharing multiple perspectives, 2) making partners' ideas accessible when individuals got stuck, and 3) partners working from their prior knowledge and attempting to build from their peers' ideas.

Providing evidence for process gain in a problem-solving domain, Laughlin et al. (2003) had university students solve letters-to-numbers math problems in three-person groups or alone. This letters-to-numbers task involved the letters A through J being mapped to a number 0 through 9. Participants proposed an equation, such as E + H, and the experimenter responded with what it equaled, say JB. If E and H were 4 and 8 respectively, then JB would be 12, where J = 1 and B = 2. After proposing an equation, the participants then made a guess as to the value of one of the 10 letters, and the

experimenter responded True or False, marking the end of a trial. The goal for the participants was to decode all 10 letters within a minimum of trials.

Laughlin et al. (2003) described an optimal strategy to solving these puzzles. If you know that J = 1, then J + JJ + JJJ = 123 = the letters given by the experimenter. By extension, if you add J + JJ + … + JJJJJJJJ = 123456789, you can solve the puzzle within one trial. In general, equations with more letters would allow participants to solve the problem more quickly than those with fewer letters. For example, A+B+C+D+E+F+G+H+I+J = 45 (because 0 through 9 sums to 45), so a second effective solution strategy would be to ask the experimenter what letters A through J sum to because you would learn two letters on the trial rather than just one. The point of describing these strategies is that the task involved multiple correct solution paths with some demonstrably better than others. In addition, the problem was complex enough that participants could gain insights into better strategies by seeing the contributions of their partners.

As they predicted, Laughlin et al. (2003) found that groups performed significantly better than even the best individuals on the letters-to-numbers task. (Participants, whether in groups or alone, did not typically perform optimally. No condition averaged better than 5-trials to solution on average. Over 35% of the lowest performing individuals did not solve the puzzle within the maximum of 10 trials versus 0% of groups). When greater parameters were included in the problem-instructions, such as to include at least four letters per equation, participants often performed better (though not always). Thus, Laughlin et al. found a task that afforded process gain even without elaborate social-structures. This task had multiple solution paths of measurably different quality, built upon participants' prior knowledge (of algebra and logic), and required innovation of equations for solving the problem within the parameters of mathematical possibility and group consensus.

In an attempt to characterize tasks that should promote benefits to groups over individuals, Laughlin et al. (2003) followed Steiner's (1972) suggestion and posited that tasks exist along a continuum of demonstrability ranging from "intellective" tasks that are highly demonstrable to "judgmental" tasks that are primarily based on preferences or attitudes. They suggested that the letters-to-numbers task promoted process gain because

it was an "intellective task," or one that involved their four conditions of demonstrability—agreed upon conceptualization, sufficient information, and participants who will recognize correct solutions and/or explain them to a partner. They hypothesized that greater demonstrability promoted greater process gain.

In a two-year longitudinal study, Phelps and Damon (1989) hypothesized that rote-learning math tasks and spatial tasks involving copying would show no benefit of collaborative learning whereas math and spatial tasks requiring greater reasoning (about proportions and about views from different perspectives) would show benefits from collaboration. Using these rote-learning and reasoning tasks as posttest measures, they compared the performance of four groups of fourth graders. One group collaborated on practice math problems. A second collaborated on spatial problems. The others were a trained control group and a no-treatment control group.

As predicted, the rote-learning measures showed no differences between conditions. On the reasoning measures, the group that collaborated on practice math problems did best on the math questions. Likewise, the group that collaborated on the spatial problems did best on the spatial questions.

On a transfer measure involving a math-related balance beam task, the math group did best. The spatial group was next highest. The trained control group was third, and the no treatment control did worst. That the mean of the spatial group was second highest (though not significantly higher than the trained control group) is an indicator that working collaboratively on an un-related task might have general educational benefits.

While this result provides interesting insight into potential long-term benefits of collaboration, the critical results for my purposes are that rote-learning and copying tasks did not yield productive collaborative interactions for learning and that more complex reasoning tasks may have benefited from collaboration. It is difficult to ascribe the latter results to collaboration because the only significant advantages observed were for groups that collaborated on the subject-matter tested. In other words, it may have been the extra training on the test-relevant materials that produced the advantage rather than collaboration per se.

Theoretical Framework

How might we characterize tasks that promote productive collaborative learning? Some researchers have suggested that more demonstrable tasks are best for successful collaboration (e.g. Laughlin et al. 2003; Steiner, 1972), while others have suggested that ill-structured tasks are best (e.g. Cohen, 1994; Hertz-Lazarowitz, 1989). In searching for a resolution to this paradox, a framework that was informative to a similar paradox in the field of transfer seemed relevant. Transfer involves the application of prior knowledge to a novel problem, and it is considered a measure of thorough understanding (Novick, 1988). Schwartz, Bransford, and Sears (2005) recently proposed two dimensions, innovation and efficiency, as a framework for reconciling contrasting findings in the field of transfer. While some researchers were suggesting that transfer was best promoted through learning activities with direct instruction, practice, and immediate feedback, others were suggesting that discovery learning, project-based activities, and other tasks allowing greater exploration were best (e.g. Anderson, Corbett, & Conrad, 1989; Tennyson, Park, & Christensen, 1985 versus Vollmeyer, Burns, & Holyoak, 1996). This contrast should sound familiar (i.e. demonstrability versus ill-structured).

The innovation and efficiency framework attempts to reconcile these contrasting views by suggesting that a trajectory toward deep understanding involves activities with both innovation and efficiency components. In other words, rather than seeing these as opposite poles on a single dimension, such that one must be antagonistic to learning while the other must be beneficial, they are posited as complementary, such that each type of activity contributes to an important aspect of understanding. For example, jazz pianists must have efficient mastery of scales and chords while also having creative ways of stringing them together to be successful.

While the key features of the innovation and efficiency dimensions still require empirical testing, a basic characterization is possible. Innovation tasks involve the use of prior knowledge to develop approximate solutions to novel problems whereas efficiency tasks involve receiving the expert solution to a novel problem and repeatedly applying it. For example, the Apollo 13 rescue mission presented an innovation task because familiar tools had to be configured in new ways to solve the square-peg and round-hole problem. By contrast, doing 50 addition problems related to borrowing after being shown an

example is an efficiency task because it is aimed at developing speed and accuracy on one specific procedure.

Instructionally, efficiency tasks involve lectures followed by drill-and-practice routines to help students become fluent in applying their knowledge. Innovation tasks, by contrast, involve phases of insight, solution attempts, and revision. As an example of how these dimensions can be combined, and the educational benefits possible from such a synthesis, Schwartz and Martin (2004) found that ninth graders who had an innovation experience and then efficient instruction were more prepared to solve a difficult transfer problem than students who experienced drill-and-practice on the same learning material. Specifically, Innovation students attempted to invent a standardized score while Efficiency students practiced calculating z-scores. Half of each group also later saw a worked example of a z-score calculation (an efficiency task). Thus, there were four conditions. On a posttest measure requiring a novel application of standardized scores to solve a difficult transfer problem, the Innovation students who received the worked example performed twice as well as all the other groups. In other words, the innovation experience made them more prepared to learn from an efficient example and adapt it to a difficult transfer problem.

Before examining why the innovation and efficiency dimensions might be relevant to collaborative learning, I will describe a novel feature of the last study that is relevant to the second experiment in this dissertation. Schwartz and Martin (2004) used a unique outcome measure to reveal otherwise hidden benefits of innovation instruction. The approach they used is known as a form of preparation for future learning (PFL) assessment because within the posttest they embedded a resource problem that, if students were prepared to learn from it, could inform their solutions to a subsequent target transfer problem (Bransford & Schwartz, 1999). The greater performance on the target transfer problem by students in the Innovation condition who received the resource problem could be attributed to their being more prepared to learn and transfer from that embedded resource. They could ensure it was due to learning from the resource problem because the half of the Innovation condition that did not receive the resource performed at the level of the other conditions. A simpler version of a PFL problem involves providing participants in all conditions the resource and seeing which students are able to

solve the target. The key feature of this approach to assessment is that it can reveal how well different instructional approaches prepare students for future learning.

Why might the innovation and efficiency dimensions be relevant to collaborative learning? First, informal observations of the study by Schwartz and Martin (2004) indicated that the students discussed and collaborated around the innovative activity quite productively. Second, the tasks in many of the studies described above that showed group performance benefits had many features in common with innovative tasks. In each case, prior knowledge could be built upon to progress toward new conceptualizations of the problem. For example, the letters-to-numbers task (Laughlin et al., 2003) allowed students to edit each other's math strategies until they found an acceptable approach, such as using formulas with more letters in it. The gears task (Schwartz, 1995), through students' dialogue, encouraged a switch in strategies from physical modeling to enumeration.

By contrast, efficiency tasks seem likely to yield little benefit from collaboration because such tasks call for less knowledge sharing, given that the solution is available from the start. For instance, if one already has the answers and merely needs to memorize them, why or how would one debate them with a partner? For this reason, I would not expect efficiency tasks to promote as much discussion or perspective sharing as innovation tasks. In addition, efficiency-oriented tasks, such as list-learning (e.g. Andersson & Rönnberg, 1995), have shown negative effects of collaboration.

Comparing the Innovation and Efficiency framework to other collaborative frameworks

In characterizing productive collaborative tasks, one might hypothesize demonstrability as the key feature. Steiner (1972) and Laughlin et al. (2003) both suggested that greater demonstrability should yield greater group productivity. While I agree that demonstrability is important for allowing a debate about different strategies, greater demonstrability may not be as important as the right amount of demonstrability. For example, if everyone within a group agrees on how to conceptualize a problem (one of the four conditions of demonstrability), then what is left to debate or learn? Complete demonstrability cannot yield process gain because it renders the group interaction meaningless by having an agreed upon conceptualization and members who are willing to

accept or imitate it upon seeing that it works. There is no push for conceptual growth, yet this push seems to be a key feature of productive collaborative tasks and well-designed innovation activities in general.

For example, from the demonstrability hypothesis, one would expect that on the letters-to-numbers task, being given the letter for 9 would help about as much as the letter for 1. Instead, 9 helped significantly more (Laughlin et al., 2003). So how do we explain this—is 9 somehow more demonstrable than 1? No. Instead, from the innovation-efficiency perspective, which suggests that good innovation tasks are those that help students connect prior knowledge to more expert concepts, this finding makes sense. Knowing A = 1 suggests a simple incremental strategy where a student can decode one new letter each turn (e.g. A+A = D = 2; A+D = G = 3; etc.). Knowing A = 9, on the other hand, is helpful in prompting a shift toward a more expert strategy because adding A yields two new letters each turn or even more (e.g. A+A = 18 = HJ; A+A+A = 27 = DB; AA+A = 108; etc.).[1] Thus, while the starting prior knowledge of adding known digits was the same, the resulting strategies and solution success could be quite different.

The innovation and efficiency dimensions have some similarity with the claims of Cohen (1994), who suggested that ill-structured tasks were best for collaborative learning. While ill-structured tasks are similar to innovation tasks in their requirement for invention, they do not suggest a need for efficient instruction, such as a subsequent lecture. Without such parameters, there may be some danger of groups spinning off a productive track. Periodic efficient instruction might help groups avoid generating errors in understanding or having vocal but wrong students dominate a discussion.

The hypotheses here are that tasks with an innovation component should present groups with greater chances to share insights and co-construct their prior knowledge into more-expert conceptualizations of a problem. These interactions should lead to a benefit from collaboration on measures of learning and transfer. Tasks that emphasize efficiency to the exclusion of innovation, by contrast, seem more likely to hinder groups because they are more like list-learning. Tasks, like list-learning, that do not push for progressive

[1] Only if I realized that A = 1 allows me to solve the problem in one trial (via A + AA + … + AAAAAAAAA = 123456789), would A = 1 promote my solution strategy as much or more than A = 9, but this insight seems well beyond what most people would spontaneously notice, even in collaboration with others (apparently).

re-conceptualization have shown poor results associated with collaboration. For these reasons, I expect people engaged in innovation tasks to show more discussion and greater knowledge sharing during the task than those engaged in efficiency tasks. Following expert feedback, I expect those doing the innovation version of a task to show greater recall and transfer of learning from the task. Below I describe two experiments to test these hypotheses. The first was an initial attempt to contrast innovation and efficiency tasks and examine their effects on learning and interaction in groups. It was an exploratory study and its primary contribution was to describe ways in which innovation versus efficiency activity affect interaction. The second experiment was more rigorous. It built upon the knowledge gained from the first study of what features are important for innovation and efficiency tasks, and it used a PFL assessment.

CHAPTER 2: THE CONCEPT-MAPPING STUDY

A first study for this dissertation examined the hypothesis that an innovation task would cause more productive collaborative interactions and better learning than an efficiency task. For this study, dyads read a one-page passage about cholesterol. They also worked with "Betty's Brain," a computer program known as a teachable agent because students teach an agent and learn in the process (Biswas, Schwartz, Bransford, & TAGV, 2001). Betty's Brain allows users to make a concept map where concepts are connected together by causal links. Students can then ask Betty questions about what she was taught, and this can help them reflect on Betty's knowledge (and their own). In the study, participants taught Betty based on the cholesterol passage (e.g., "Exercise *decreases* LDL").

When creating their concept maps, dyads in the innovation condition had to figure out all the links between concepts from the passage. Dyads in the efficiency condition received a list of all 23 pair-wise links. They had to use these links to make a completely connected map. Figure 2 shows an expert version of the map.

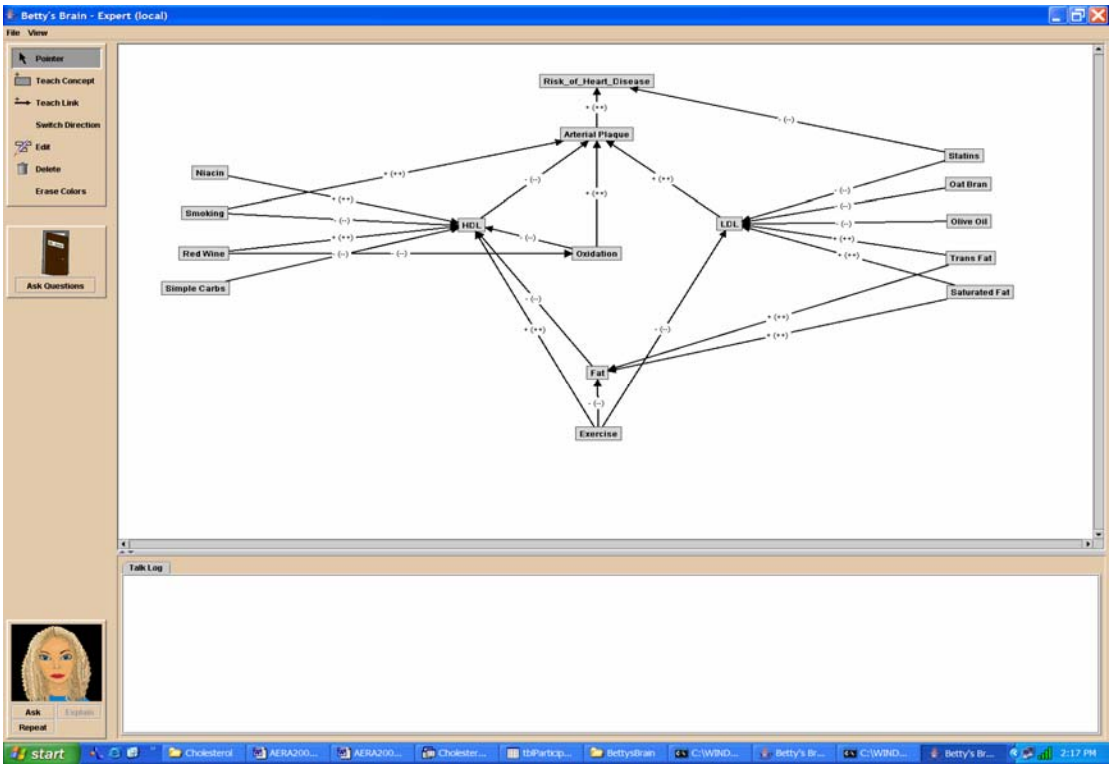

Figure 2. Expert Betty Map with 16 nodes (provided) and 23 correct links.

Methods

Participants—Thirty-eight participants, mostly undergraduates, were randomly paired and assigned to one of two conditions. Twenty-six women and 12 men participated. Ten dyads were assigned to the Innovation condition and nine to the Efficiency condition. Ten dyads were mixed gender, six in Innovation and four in Efficiency.

Materials and Measures—Test-1 and Test-2 assessed participants' knowledge of the cholesterol passage and the related concept map. These tests consisted of re-drawing the map from memory and answering 12 multiple-choice questions (e.g., "How does an increase in HDL affect Arterial Plaque?").

Procedure and Design— Table 1 summarizes the study design. The unusual component of this design is that students took two tests: one immediately after creating their initial map, and one after they had a chance to learn a second time by reviewing an expert map and the passage. This second test permitted an estimate of how well students were prepared to learn. During the test phases, all resources were removed and participants worked alone. The map-making phase was 20 minutes for both conditions, and the map-revision time depended on how long it took for participants to revise their Betty map until it matched the provided expert map (usually two to five minutes). Dyads were videotaped while collaborating.

Table 1
Design of Study 1

Step	**Context**	**Innovation**	**Efficiency**
1	Alone	Read Passage on Cholesterol	
2	Together	*Generate Map Links*	*Make Map from List of Links*
3	Together	Get corrective feedback on 3 possible links. Have Betty trace through map and study it if time.	
4	Alone	Test-1: Redraw the Map and Answer 12 Multiple-choice Questions	
5	Together	See an Expert Map and Revise Until Match	
6	Alone	Re-read Cholesterol Passage	
7	Alone	Test-2: Redraw the Map and Answer 12 new Multiple-choice Q's	

Results

Analyses of the learning outcome measures will be described first, followed by process measures obtained through transcript and video analysis, and finishing with the relationship between process and outcome measures. Although few significant learning differences were found, significant differences in process were observed. While some of the processes showed no relationship to learning, such as amount of turn-taking, others did. Explanations were important for Efficiency dyads, and degree of shared-knowledge was predictive of learning gains for both conditions, especially Innovation.

The outcome measures were participants' quiz scores and their map re-drawing scores. These measures were taken after initial creation of their Betty maps (Test-1) and after revision and re-reading (Test-2). Each of the 12 items on the quizzes, were scored as correct (1) or wrong (0). For the map re-drawings, Betty's Brain has a procedure that scores maps by comparing the answers obtained by tracing through a given map to the answer obtained from the expert map when comparing the effect of one concept on another. The advantage of this approach is that it weighs more central concepts, like HDL, more heavily than less central ones because they are connected to more nodes and are thus involved in more questions. A disadvantage of this approach is that it only attends to final answers, so participants could get the right answer for the wrong reason.

A total of 48 questions about connected concepts were assessed by the Betty program for the cholesterol map used in this study (48 is the total number possible, it is not arbitrary). If a participant's map gave the same answer as the expert map on 47 out of 48 of the tested relationships, then, they would score 98%. This scoring procedure was used on dyads' initial Betty Maps as well as each participants Test-1 map re-drawings (henceforth, Map1) and re-drawings at Test-2 (henceforth, Map2). Participants drew Map1 and Map2 on paper, so the experimenter traced them into Betty's Brain so the program could score them.

While neither condition significantly outperformed the other on either of these measures at Test-1 or Test-2, all $t(36)$'s < 1.5, $p > 0.15$ (two-tailed), the Innovation condition tended to improve more over time than Efficiency. A repeated-measures multivariate analysis of variance (MANOVA) indicated a marginally significant time-by-

condition interaction (*Wilks' Lambda* = .90, $F(1, 32) = 3.4$, $p < .10$). These results can be seen in Figure 3 separated by measure, time, and condition.[2]

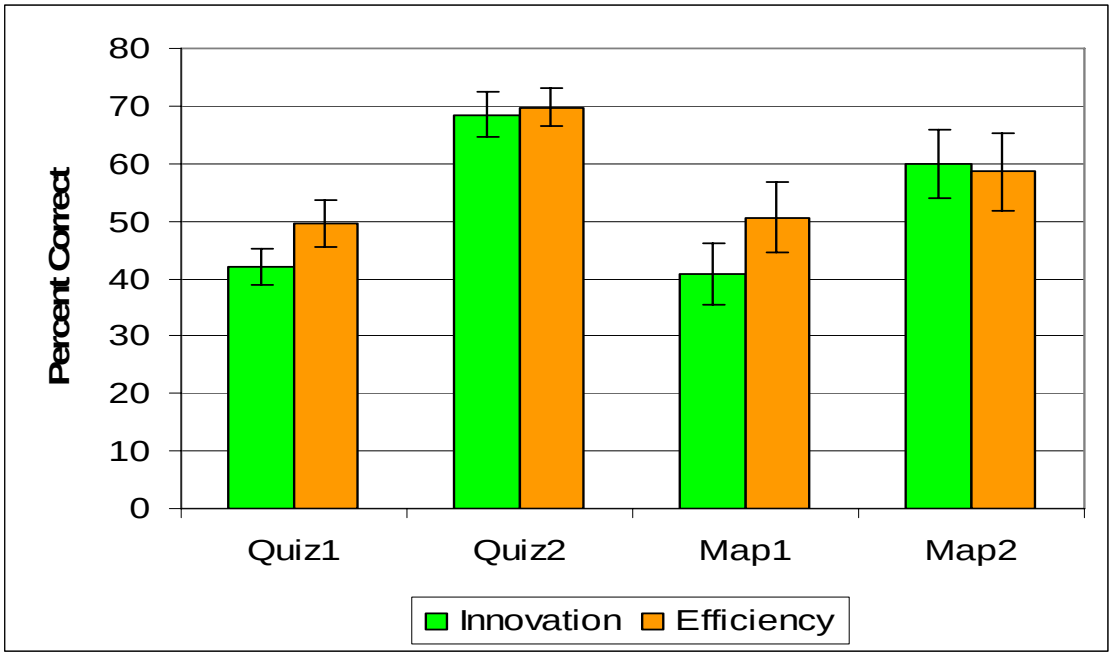

Figure 3. Individual (n = 36) performance on learning outcome measures by condition.

Quiz scores and map re-drawing scores were further analyzed for their reliability to check if they provided suitable measures. The quizzes did not show adequate reliability (alphas = 0.42 and 0.37 for Quiz1 and Quiz2, respectively). For the re-drawing scores, the current version of Betty does not provide the results for each of the 48 possible connections it assesses. Without these data, calculating their reliability is impossible. As a proxy, performance on each of the 23 correct links was recorded (analogous to a 23-item quiz) with participants receiving 1 point for each correct link in their map. Using this metric, the map re-drawings showed high reliability (alphas = 0.90 for both Map1 and Map2, and all alphas were at or above 0.84 when further separated by condition). I suspect the contrasting reliability results for the quizzes versus the maps were due to the limited domain knowledge that could be assessed via the 12-item quizzes versus the 23-item maps. For this reason, the quiz measure will not be described further,

[2] All indicators of variance around the mean in the figures and text are standard errors.

whereas the map measure will be analyzed in greater detail to determine where Innovation improved and how.

Given the high reliability of the maps, it may be little surprise that performance on each map predicted performance on the subsequent map task, suggesting that it was important for groups to start well. For Innovation dyads, the Betty Map score correlated moderately highly with their Map1 score ($r = 0.69$). (To obtain a Map1 score for each dyad, the Map1 score of each member of the dyad was averaged with their partner's score.) Similar results were obtained for the correlation between their Map1 and Map2 scores ($r = 0.65$). Separate from the total map score, I wondered whether the number of errors (i.e. wrong and missing links) in dyads' maps would be correlated over time, especially given that Innovation dyads did not receive corrective feedback until after Map1. As mentioned above, participants could receive credit from Betty on their map despite having an incorrect path in their diagram; therefore, analyzing errors separately from the total map score could provide different results. Correlations between number of errors in their maps were high (r's = 0.77 for Betty Map and Map1 and 0.69 for Map1 and Map2). Efficiency dyads' correlations between Map1 and Map2 were also high (r's = 0.92 and 0.91 for map scores and errors, respectively). Because only one group in the Efficiency condition created a Betty Map with errors, correlations between Betty Map and Map1 are not reported for Efficiency. Despite the high correlations between Innovation dyads' Betty Map scores and Map1 scores, there was only a modest correlation between their Betty Map scores and Map2 scores ($r = 0.26$), suggesting that errors on Betty were not helpful but did not mean imminent failure either.

Chi-square analyses indicated that the Innovation condition improved more than the Efficiency condition on map-links that both partners initially missed in Map1 ($\chi(1) = 8.3$, $p = .004$).[3] There was no significant difference between conditions on links where one partner was correct initially on a link ($\chi(1) = 0.47$, *ns*). On links where both partners got the link correct on Map1, Innovation showed marginally less forgetting ($\chi(1) = 2.7$, $p = .099$). These results are shown in Figure X.

[3] The linear-by-linear association chi-square procedure is reported here and on subsequent chi-squares using a two-tailed significance level.

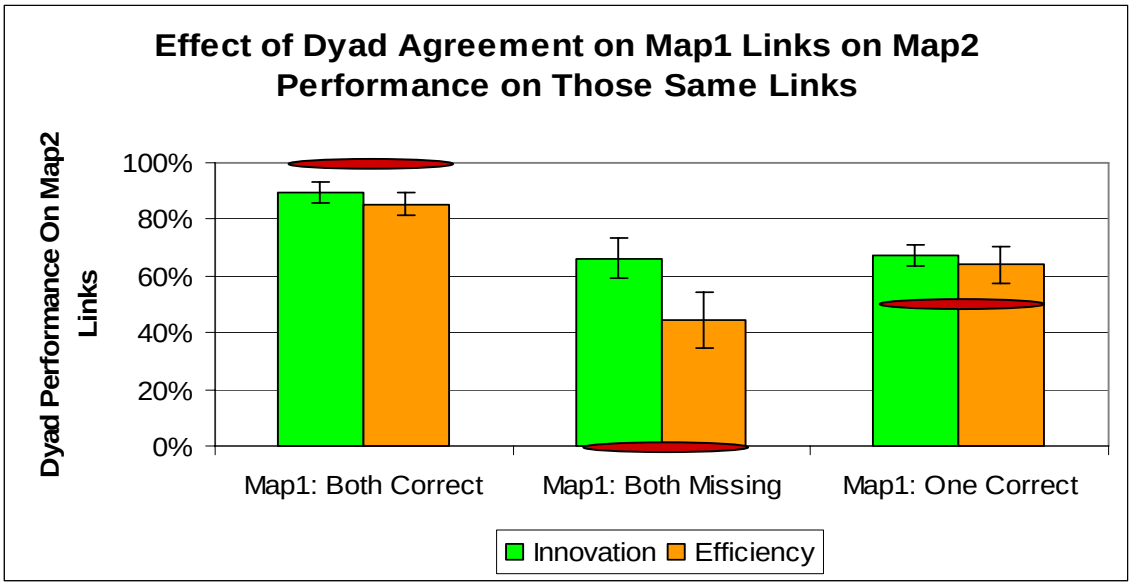

Figure 4. Map2 performance given Map1 performance. Oval lines show dyad performance on Map1.

To be sure these findings are not misinterpreted, I do not take them to mean that Innovation learned more than Efficiency (clearly their Map2 score was not significantly greater), however, they did improve more over time. This improvement appears to be due largely to their addition of items originally missing from their maps. Of the 80 links neither partner included on Map1, 41 of them were not in their original Betty Map. Only 5 of these 41 remained missing from both partners' maps on Map2.

A more detailed analysis of errors and error-correction on Map1 follows to provide further insight into how Innovation and Efficiency might differ. Individuals in the Innovation condition made more errors, but not significantly more than those in the Efficiency condition ($t(36) = 0.8$, *ns*). This pattern held true for errors of omission and commission.

Interestingly, the Innovation participants corrected a marginally higher percentage of their errors than the Efficiency participants ($t(36) = 1.5$, $p = .07$ (one-tailed)). These results are displayed by error type in Figure 5. They further support the findings of the Chi-square tests above by showing similar results when wrong links were included in the analyses with missing links.

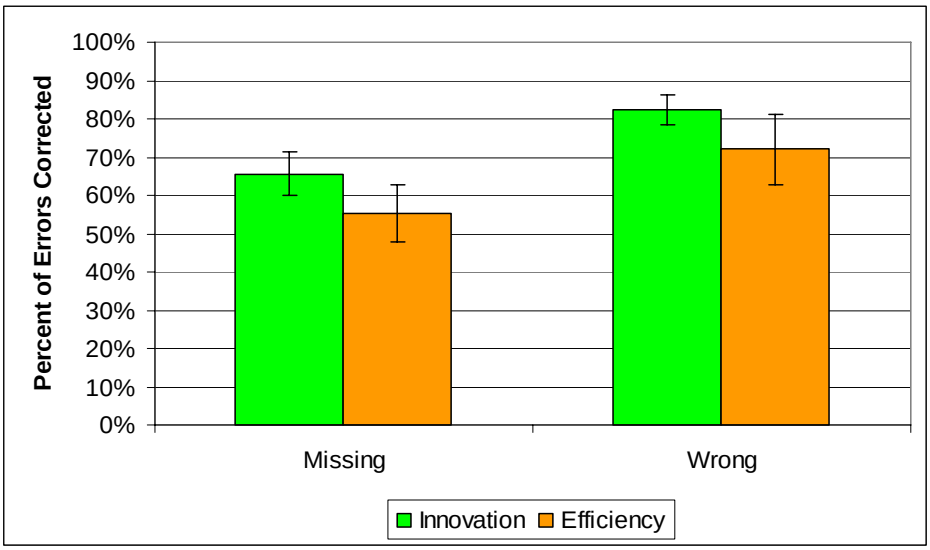

Figure 5. Percent of errors individuals corrected on Map2 by error-type and condition.

Less forgetting is another way in which Innovation could improve over time compared to Efficiency. Specifically, a link was counted as forgotten if it was correctly re-drawn in Map1 but not in Map2. Independent t-tests indicated that the Innovation participants forgot marginally fewer correct links ($t(36) = 1.3$, $p = .09$ (one-tailed)), as shown in Figure 6. This result was not due to Innovation participants re-drawing fewer correct links on Map1 and thus having fewer to forget on Map2 ($t(36) = 0.8$, *ns*, *M*'s = 11.2 versus 12.8 for Innovation and Efficiency, respectively).

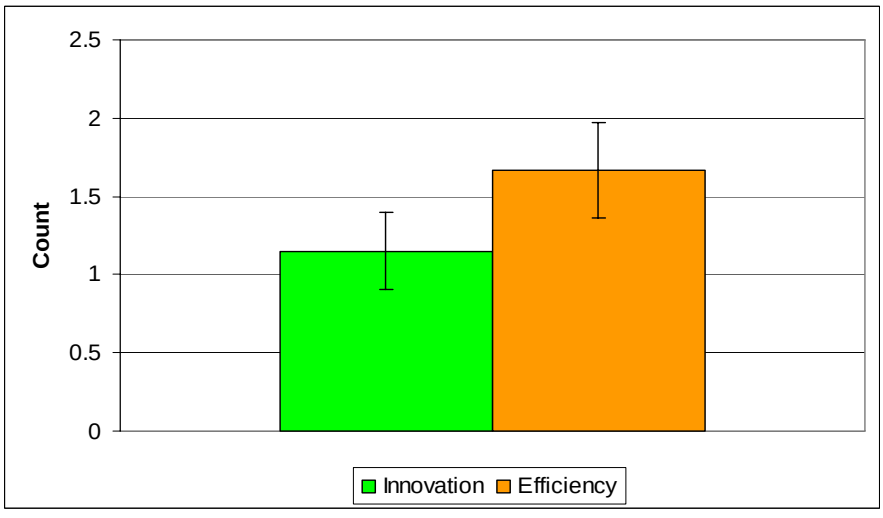

Figure 6. Innovation participants forgot fewer correct links from Map1 to Map2.

Thus, on learning outcome measures, Innovation and Efficiency did not differ tremendously except that Innovation improved more over time, primarily because it performed somewhat lower initially. The improvement was due primarily to greater error correction and, to a lesser degree, to less forgetting. What follows is an analysis of process measures to see how the conditions affected dyadic interaction and how these interaction patterns were associated with learning.

Process Measures Analysis

Informal observation and video analysis suggested that the conditions produced different overall interaction patterns characterized by divide and conquer strategies for Efficiency and active co-construction for Innovation, especially during the map-making phase as opposed to the Betty-checking and revision phases. Seven of the nine efficiency dyads partitioned the map-making task such that one student became the list-reader and the other became the link-maker. They also typically finished making the map in about 10 minutes and then studied or asked questions of Betty. Seven of the ten innovation dyads also tended to have a reader and a scribe; however, both participants typically searched through the text for answers. A cursory coding of text-reading behavior indicated that every participant in the Innovation condition checked the text at least three times regardless of whether they were the reader or scribe. They also remained highly engaged during the full 20-minute mapping phase, searching for links, double-checking each other, and periodically discussing what links could legitimately be inferred from the passage. Thus, as a general characterization, the innovation versus efficiency versions of the task produced distinct interaction patterns, though perhaps not distinct enough to produce robust learning differences.

Despite greater engagement, potential problems in the interaction of the Innovation dyads also existed. With no scripting of their process, these dyads had to determine their own path through the task. For one dyad, this meant starting from memory. This led to many errors and disagreements and ultimately the partners realized they needed to use the text. For others, it meant relatively independent searching through the text and announcing their findings or simply entering them into the map. Most groups were not so extreme, they typically used the map to help guide their search of the

text for links they thought might be needed (e.g. “Oh, we don’t have anything connected to oxidation yet; let’s find oxidation”). However, the relatively extreme cases indicate the danger on this innovative task of spinning off a productive track. Surprisingly, only one Innovation dyad proceeded systematically through the text from the start to the end to make their map. This group looked considerably like an Efficiency group in their map-making process and were the only Innovation dyad to construct a perfect Betty Map. They also were highly successful on average, but that was because their map-maker did exceptionally well on Map1 and Map2 while the reader performed near the average level.

The finding that the map-maker did well in the group mentioned above was not uncommon for the Innovation dyads. Of the seven Innovation dyads with a consistent map-maker, six of them showed greater performance by the map-maker on Map2 (across all seven dyads the average difference, in favor of the map-maker, was 28.0 ±15.3 percentage points). This was not the case for the seven Efficiency dyads with a consistent map-maker (-11.1 ±8.3 percentage points). One observation that might shed light on this finding was that the map-maker in Innovation dyads had time to study the map and see what links might be missing and which ones made sense. The reader in these groups may have been more concerned with finding and reporting the next link in the text, rather than considering how they fit into the larger structure of the whole map. In Efficiency, the map-maker was typically very busy making the next link their partner read from the list.

The following sections describe results of video and transcript analyses that fall into two general categories: resource use and knowledge-sharing. The resource use section includes analyses of dyads’ use of the Betty-Expert and their use of the Betty tracing feature. The Betty Expert gave corrective feedback on three paths in dyads’ Betty maps. The Betty tracing features allowed participants to ask Betty questions about how two nodes were associated in their map, to which Betty would respond by tracing through the path(s) between the nodes.

The second category of analysis, knowledge-sharing, describes three process measures: 1) the number of turns dyads took per link they produced while making their Betty maps together, 2) the number of explanations they gave during mapping, and 3) the degree to which they showed similar recall of their Betty map (as measured by similarity

of their Map1 drawings). The turns-per-link analyses indicated that Innovation dyads interacted more while making their maps. Analyses of explanations revealed that neither group explained more in general, but Innovation explained much more than Efficiency during the map-making phase and less during the Betty-Check phase. Included in this analysis is the difficulty of the links that both groups explained, which, contrary to what we might expect, were essentially the same. Finally, the degree of Map1 similarity concludes the section and suggests that Innovation dyads shared knowledge better.

Resource Use

The Betty's Brain Expert gave dyads feedback on three potential paths in their map. One path was a single link, the second involved multiple-links, and the third was feedback that no path existed between two disconnected nodes. This feedback was relevant to the Innovation condition because they could use it after the map-making phase to spot and potentially correct mistakes before submitting the Betty map for evaluation. Surprisingly, this feature may have done more harm than good. Four of the five groups that received feedback that they had a wrong answer created a wrong link in an attempt to correct their error. All four groups were missing one or more links from a multi-link path between two nodes. Rather than considering a multi-link solution, they created a direct path to solve the problem in the most immediate, but wrong, fashion.

This is not to say that the Betty Expert feedback was entirely harmful. One group that received error feedback did figure out the multi-path solution. One of the participants in this dyad was particularly vigilant about checking the text and insisting the other member also justify links via the text. In addition, one of the four groups, while making an incorrect direct link immediately after checking the Expert, also went to the text and found three correct links they had not put in their maps up to that point. While they did not realize that their direct link was wrong, the Expert seemed to prompt them to double-check the text for other errors they might have made. After making the new links, they checked the Expert again and found that everything was correct.[4]

[4] The Expert did not evaluate how they got their answer, just whether it was correct. Though participants were warned of this, they may not have fully understood what it meant.

After checking with the Betty Expert, the dyads also studied their map by using a second feature of Betty. They asked Betty questions about how two nodes in the map were related. For example, "If HDL increases, what happens to Arterial Plaque?" Efficiency dyads asked Betty many more questions than the Innovation dyads (8.3 (±1.0) versus 2.7 (±0.7), respectively). This resulted in Betty tracing through significantly more correct links in the Efficiency dyads' maps than in the Innovation dyads' maps, as shown in Figure 7.

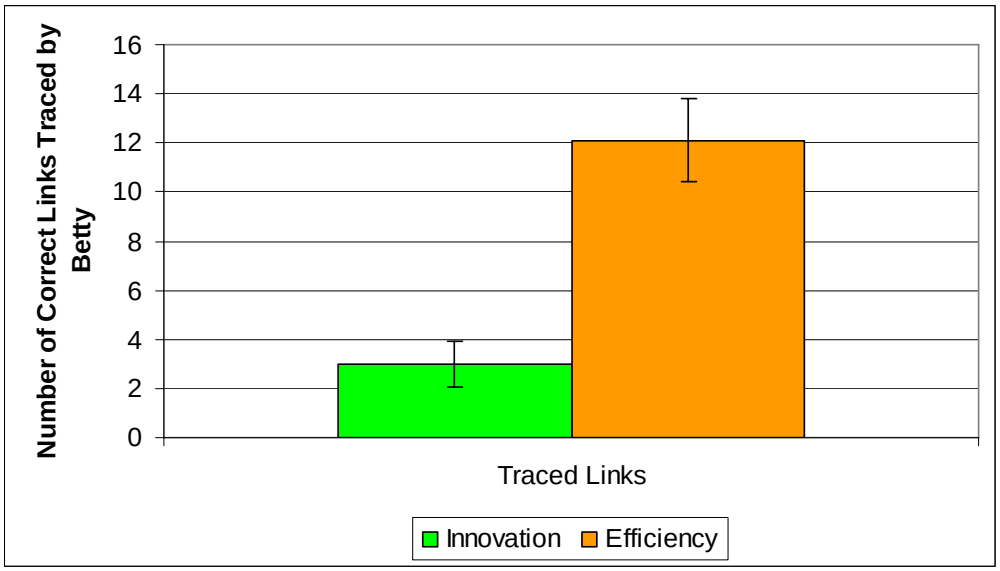

Figure 7. Number of correct links dyads asked Betty to trace through per condition.

Knowledge-Sharing

One of the critical differences between conditions was in the number of turns they took between each link they made in their Betty maps. This difference was critical because it is at the heart of the difference between what an efficiency activity versus an innovative activity should look like—the former should involve little discussion and considerable partitioning and practice; the latter should involve rich discussion, mutual responsiveness, and revision of ideas. If both conditions spent a minimum of turns between links, the experimental manipulation would have failed. As shown in Figure 8, this was not the case. As expected, Innovation dyads took significantly more turns per link they created in their Betty maps than Efficiency dyads ($t(16) = 2.9$, $p = .01$).

Efficiency dyads averaged less than three turns per link (2.9 ±0.2), indicating that a typical interaction during map-construction involved the list-reader stating the link and their partner entering it on the computer. Innovation dyads averaged over four turns per link (4.5 ±0.5). Table 2 shows sample turns from dyads in each condition to give a sense of the partitioning observed in the Efficiency dyads and the elaborated discussions (though not always correct!) in the Innovation dyads during map-making.

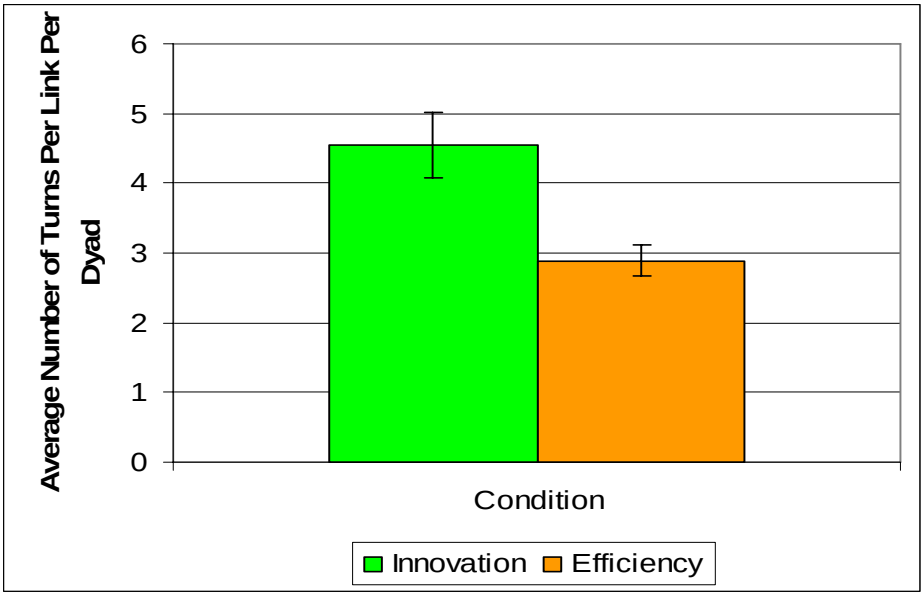

Figure 8. Turns taken per link made in Betty maps per condition.

Table 2

Sample Dialogue from Each Condition During Map-Making

Innovation	**Efficiency**
A: Also, **Smoking**, so it can increase the likelihood of plaque build up.	A: And **Statins** decrease **LDL**.
B: So that would go to **Arterial Plaque**. And that would be an increase in **Arterial Plaque**, right?	B: What are **Statins**?
A: Yeah. Or, is it causal, I don't know. It says it increases the likelihood. Do you think we should put that in?	A: Yeah, I don't know what **Statins** are.
B: Hmm, possibly. Do you think it is organized (like this, somewhat symmetrically—looking at the map) for any particular reason (laughs) (the participant seems to suggest maybe the link should not be added b/c it reduces symmetry).	B: Is that everything on the list?
A: Yeah, I think maybe we shouldn't then (both laugh) (they seem to be trying to take cues from map to help organize interpretation of the reading).	A: I think so, yeah.
B: Okay, never mind (that link, Sm + AP).	C: Next, increase in **Oxidation** decreases **HDL**. So increase, decrease.
A: So maybe we put **HDL** (to AP) though and **LDL** (to AP), right? B/c **LDL** begins plaque formation, and **HDL** removes it.	D: (makes link).
	C: Okay, next. Increase in **Niacin** increases **HDL**. So increase, increase.
	D: (makes link).
	E: **Oat Bran** and **LDL**. Decreases.
	F: (makes link).
	E: **Olive Oil** and **LDL**. Decreases.
	F: (makes link).

Note. From these interactions, we can see the handing off of answers characteristic of Efficiency dyads' map-making interactions. We also see an extended exchange in one Innovation dyad, with both participants using the map and text to guide their interpretation of what links to add or not add. In this case, their hypotheses about the map being symmetrical led them to a wrong conclusion, but they also reasoned correctly about HDL and LDL.

As shown in Table 3, both conditions took more turns during the Betty-check phase of the mapping process. Innovation averaged 5.5 turns per question they asked Betty while Efficiency averaged 6.3 turns. It appears that the teachable agent may have prompted greater discussion during this phase, especially for Efficiency dyads.

Table 3

Turns per Betty Question Contrasted with Turns per Link During Map-Making

	Turns Per Betty Question	**Turns Per Link**
Innovation	5.5 (±1.0)	4.5 (±0.5)
Efficiency	6.3 (±1.0)	2.9 (±0.2)

Turns-per-link is a relatively coarse code for quantifying discussion quality. Explanations provide a better measure of quality of discussion. In addition, previous research has found them to be associated with learning (e.g. Webb and colleagues). Explanations were coded in dyads' transcripts in terms of which turns included

explanations and also which links received an explanation at some point in the text even if that point came after the link was made (such as during the Betty-trace phase).[5]

The following analyses describe when explanations were given and the difficulty of the links explained. The number of explanations given by dyads in each condition did not differ significantly ($t(17) = 1.3$, *ns*). However, the timing of the explanations differed dramatically by condition. As shown in Figure 9, Innovation dyads did most of their explaining during the map-making phase where they were attempting to determine which nodes should be connected. Efficiency dyads did most of their explaining during the Betty-check phase. The number of explanations by phase differed significantly by condition ($t(16)$[6] $= 4.1$, $p < .001$, $t(17) = -2.8$, $p < .05$.) except in the revision phase ($t(17) = -0.7$, *ns*).

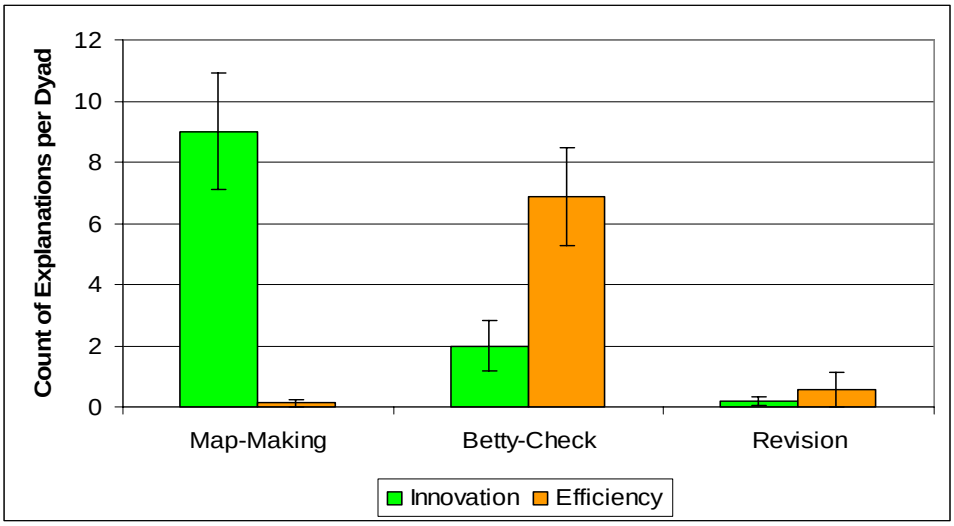

Figure 9. Where dyads' explaining occurred during the course of the experiment.

Given that both conditions produced a similar number of explanations, were those explanations about the same types of links? Specifically, were harder links explained more than easier links by one or both conditions? To answer this question, links were

[5] Because this work was primarily exploratory, a second rater was not recruited for inter-rater reliability purposes. However, many of the explanations were easy to code based on the use of the terms "because," "so," and "by." Other statements coded as explanations were: 1) those that questioned a partner's suggestion and offered evidence for why it was questioned, and 2) statements that provided text-referenced answers to a question posed by oneself or one's partner regarding a link or concept in the map.

[6] Video-taping of one Efficiency dyad was delayed until the end of map-making, though observationally, their process seemed similar to the others in that condition.

divided into three categories: easy, medium, and hard. These categories were defined by the average scores of Efficiency participants' performance on Map1. Using Efficiency to define the categories seemed more sensible than using or including Innovation because the Efficiency dyads worked with correct answers from the start and typically only read through the passage once. Using the Map1 link scores was appropriate because Map2 recall could be affected by the mistakes made on Map1. In any case, the hope was that any difference in recall between links for Efficiency dyads on Map1 would be due to their relative simplicity/complexity rather than extra processing or discussion. Easy links (5 of the 23 links) were those with average-scores greater than or equal to 75% recall. Hard links (9 links) were those with average-scores of less than 50% recall, and Medium links (9 links) were those that were left. As shown in Figure 10, neither condition differed on which links they explained, generally explaining about one-fourth of links in each category.

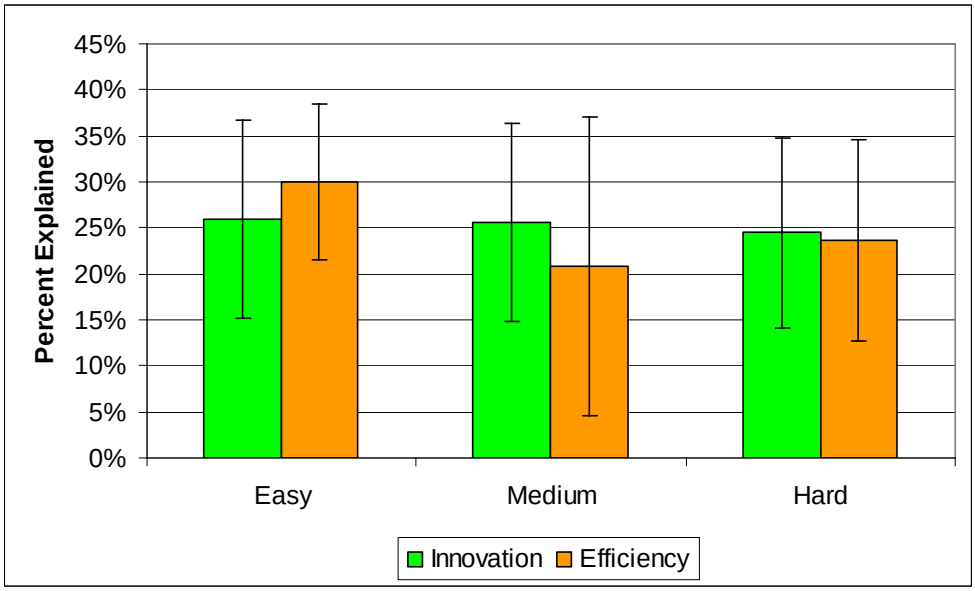

Figure 10. Percent of easy, medium, and hard links explained by condition.

The number of turns taken for explained links versus unexplained links is a relevant analysis for two reasons: 1) it provides a way of testing whether explanations implied greater dialogue, as we would expect; and 2) it suggests a possible mechanism whereby increased dialogue could have educational benefits (i.e. when said dialogue involves explanations). To conduct this analysis, the number of turns taken during the

construction of each link was counted (as in previous analyses). These counts were then separated for links that were explained versus unexplained and averaged across dyads in each condition. A repeated-measures MANOVA of turns taken per links-type indicated that, as one would expect, explained links received more discussion than unexplained links (*M*'s = 5.0 ±0.8 and 3.5 ±0.3 for explained and unexplained, respectively; *Wilks' Lamda* = 0.8, $F(1,15) = 4.7$, $p < .05$). This result was driven primarily by the Innovation condition as indicated by a marginally significant link-type by condition interaction (*Wilks' Lamda* = 0.8, $F(1,15) = 3.7$, $p = .07$). Given that Efficiency dyads did most of their explaining during the Betty-check phase (and not during the map-making phase), the extra turns taken during the explanation would not be counted in this analysis. For this reason, the interaction reported here should not be a surprise nor should it be interpreted to mean that Efficiency dyads' explanations were not as elaborate as Innovation dyads' explanations.

As a metric of knowledge-sharing, the interactions that dyads produce together can be informative. What they take away from the group interaction and apply on their own can also be valuable. One such measure from this study comes from the degree of similarity between partners' maps on the first map-redrawing test done alone (Map1). The total number of correct links, wrong links, and missing links in both partners' maps provided the score (TotalShared). As shown in Figure 11, Innovation dyads shared marginally more map elements than Efficiency dyads ($t(17) = 1.3$, $p < .10$ (one-tailed)). One potential consequence of dyads' explanations of the links they constructed was that it may have contributed to partners having more shared knowledge. Specifically, 58% of un-explained links were re-drawn correctly by both partners on Map1 whereas 69% of explained links were re-drawn correctly.

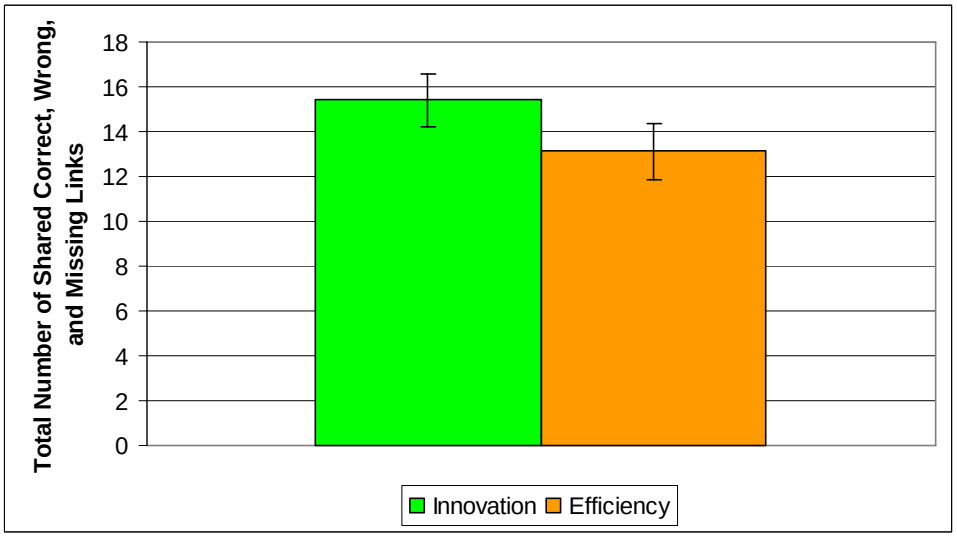

Figure 11. Map1 similarity (the number of shared links per dyad on Map1).

Relationship between Process and Outcome Measures

Following the report on process outcomes, this section describes the relationship between dyad processes and their learning. First, the relationship between learning and link-tracing with Betty will be examined. Second, turn-taking and learning will be analyzed. Third, potential effects of explaining on recall will be reported. Finally, the association of degree of shared-knowledge with future learning will be presented.

Efficiency dyads traced significantly more correct links with Betty than Innovation dyads. Was there a consequence to learning for this? A repeated-measures MANOVA indicated that correct links that were traced with Betty were recalled significantly more than links that were not traced, regardless of group (*Wilks' Lambda* = .62, $F(1, 16) = 9.9$, $p = .006$).[7] This effect was due primarily to benefits of Betty-tracing for Map1 scores (*Wilks' Lambda* = .72, $F(1, 16) = 6.2$, $p = .024$). As shown in Figure 12, results did not differ significantly by condition (MANOVA between subjects comparison: $F(1,16) = 0.9$, $MS = 0.11$, *ns*) nor were there any significant interactions between condition and time or Betty-trace (all *Wilks' Lambdas* > .89, $F(1, 16)$'s < 2.0, *ns*).

[7] One Innovation dyad was removed from the analysis because their question of Betty was about disconnected nodes. MANOVA results were nearly identical when including this dyad and substituting the Innovation mean Trace scores for their two missing scores.

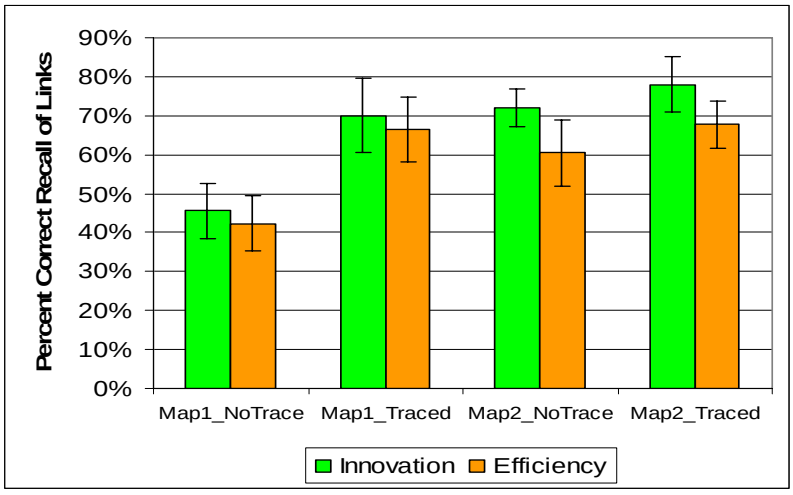

Figure 12. Dyad performance on correct links traced or not traced by Betty per condition.

While creating their maps, Innovation took significantly more turns discussing which links to include and why than Efficiency (4.5 versus 2.9). Did more turns per link imply greater recall of that link? Not exactly. Before examining these results in greater detail, a description of the transcript is appropriate. The transcript included turn-by-turn conversation. It also included which link was created on which turn. Because the map re-drawing assessment involved the same links that participants created in their maps, it was possible to connect turn-by-turn conversation around a link to learning and memory of that link. To my knowledge, this is the only study to use such a fine-grained coupling of interaction-process and learning-outcome in its analyses.

To answer the question above, the number of turns dyads took for each link they made in their Betty maps before revision (N = 360 links[8]) was correlated with their score on each link. Scores could range from 0 to 2 depending on if neither member of the dyad correctly re-drew the link, just one member, or both members. This procedure yielded no significant correlations (r = -0.01 and r = 0.02 on Map1 and Map2, respectively). In other words, it was not the quantity of discussion around a link that determined dyads' recall of it, and this was true regardless of condition (r's between ±0.06).

[8] 360 links is short of the total number of links ultimately created after the revision phase by all dyads (19 dyads * 23 links per dyad = 437 links). It is the number of correct links that were created before the revision phase.

While more turns per link during the map-making phase did not relate to learning, discussion during the Betty-trace phase did. Recall from Table 3 that both conditions took multiple turns per question they asked of Betty (between five and six turns)**.** The more turns Efficiency dyads took during this phase, the more they were likely to improve their map re-drawing scores from Map1 to Map2 ($r = 0.70$, $p < .05$). This finding did not hold for the Innovation group ($r = -0.12$, $p = .73$).

Explaining had important implications for learning, at least for Efficiency dyads. A repeated-measures MANOVA indicated that explained links were recalled more than unexplained links (*Wilks' Lamda* = 0.6, $F(1,15) = 8.7$, $p = .01$). While recall of both types of links improved over time (*Wilks' Lamda* = 0.5, $F(1,15) = 17.6$, $p = .001$), unexplained links gained more (*Wilks' Lamda* = 0.6, $F(1,15) = 8.8$, $p = .009$). Finally, a significant condition by link-type interaction indicated that Efficiency dyads performed better on explained links than Innovation dyads compared to their performance on unexplained links (*Wilks' Lamda* = 0.6, $F(1,15) = 11.1$, $p = .005$). As shown in Figure 13, the benefit of explaining is driven largely by Efficiency.

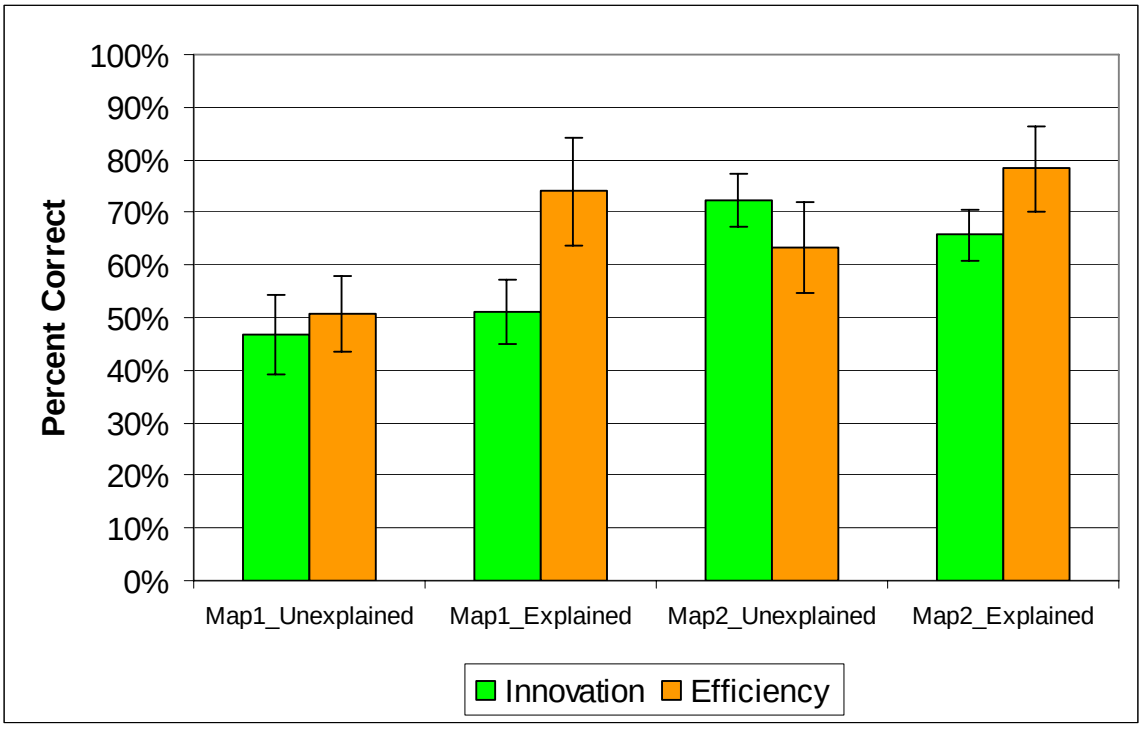

Figure 13. Dyad performance on explained and unexplained links by condition.

Recall that Innovation partners' maps were marginally more similar than Efficiency partners' maps on the Map1 re-drawing assessment. Across conditions, partial correlations controlling the association of Map1 score to Map2 score suggested that dyads that shared more knowledge gained more from Map1 to Map2 (*r's* = .80 and .72 and *p's* < .05 for Innovation and Efficiency, respectively). In other words, the more similar partners' maps were on the initial map re-drawing test, the more they were likely to improve after a re-learning opportunity. Interestingly, in a stepwise regression that compared the predictive value of Map1 score and number of shared correct and wrong links (Shared_RW), Map1 score best predicted Map2 score for Efficiency (R^2 = 0.85) while Shared_RW best predicted Map2 score for Innovation (R^2 = 0.74), and neither second predictor added significantly to the original model for either condition.[9]

Another way to examine potential benefits of shared-knowledge in a dyad is to measure performance on links that both partners missed on Map1. We can do this by seeing if their improvement on these links in Map2 is greater than expected. How much would we expect partners to increase on these missing links? On links where one partner missed it on Map1 while the other re-drew it correctly, participants went from an average score of 50% (by definition) to about 65%. In other words, they increased 15% on links that one partner missed initially. In order to show a 15% improvement, either both participants need to average 15% improvement or one partner needs to average 30% improvement while the other shows no forgetting. If the partners show 10% forgetting, as they did on links that both re-drew correctly on Map1, then a participant might need to average 40% improvement on these links to result in a 15% average improvement on Map1 links that one member missed initially. Figure 14 shows the minimum and maximum levels of improvement we would expect to see in situations where there was initial disagreement. The actual level of improvement where both partners were initially wrong exceeded these expected levels, and this was particularly true for Innovation dyads.

[9] Map1 score and the total number of shared correct, wrong, and missing links (TotalShared) added significantly to each other for both conditions when predicting Map2 score. This was the best model obtained for predicting Map2 score (R^2 = 0.79 and 0.93 for Innovation and Efficiency, respectively).

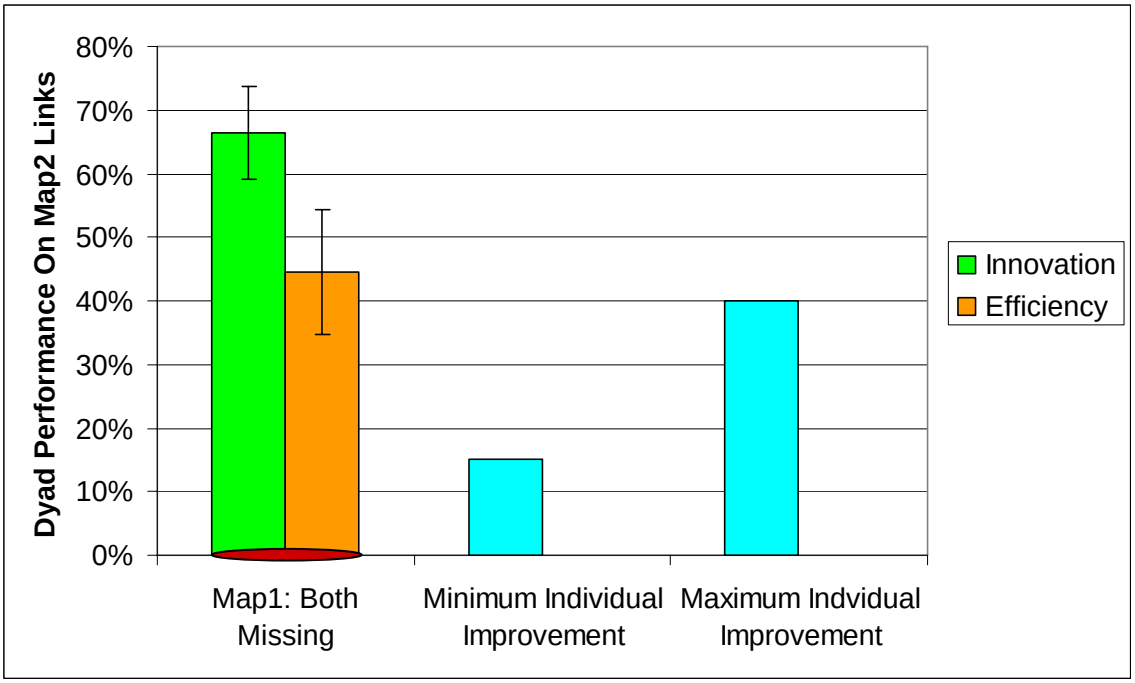

Figure 14. Dyad performance on Map2 on links that both missed on Map1. This is a measure of the benefits of agreement in a dyad. The blue bars represent the lower and upper-bound of how much we might expect individuals to improve.

Discussion

This first study revealed few learning differences between conditions. Neither Innovation dyads nor Efficiency dyads outperformed each other at either test phase. Despite this lack of learning differences some robust differences in patterns of interactions were found, many of which were in line with hypotheses. First, Innovation dyads showed more turn taking during the map-making process than Efficiency dyads. While Innovation dyads tended to discuss each link, Efficiency dyads showed greater partitioning with one student reading the answers and the other entering the information.

During the map-making phase, Innovation dyads showed much more explaining. When the map-checking phase was included, however, the Efficiency dyads caught up. Given that the Efficiency dyads did most of their explaining and turn-taking during this time it appears that Betty's Brain may have prompted them to do so via its map tracing feature. Interestingly, that map tracing feature also appears to have benefited learning because links that were traced were recalled significantly more than links that were not. While these findings might be exciting for researchers that study teachable agents, they may have undermined the treatment validity in this study.

A third variable was examined as a proxy for degree of agreement between partners, or degree of knowledge sharing. By counting the number of correct, wrong, and missing links that partners had in common on their initially redrawn maps (done alone on Test-1), we learned that Innovation dyads agreed marginally more than Efficiency dyads. In addition, we saw that explained links were more likely to be shared than unexplained links. Thus, on process measures the Innovation activity appeared to have many of the expected effects, prompting greater discussion, explaining, and knowledge sharing.

Why did these behaviors not result in greater learning for the Innovation dyads? One possibility is that they generated too many errors because the task did not provide enough feedback to keep them from spinning off. The more errors dyads made on their Betty maps, the more they made on Map1 and Map2.

Fortunately, a different method for relating process and learning outcomes was possible. This study introduced a novel method of relating turn-by-turn interaction to learning. Because the task and the assessment involved drawing the same concept map, it was possible to relate the interactions around each link as it was constructed to participants' later recall of that link. This method could provide the type of fine-grained analytic tool that could help the field achieve one of its widely accepted goals: understanding how specific interactions relate to specific learning. This tool could be even more informative if we could find a way to relate moment-to-moment interaction to subsequent performance on novel tasks.

The results of comparing moment-to-moment interactions around a link to learning of that link revealed interesting results. First, the number of turns taken per link had no relationship to recall of that link, regardless of condition. In other words, successful learning does not seem to involve simply more talk; more talk of a specific kind seems to be required. Specifically, more talk during the Betty-checking phase was associated with greater subsequent learning for Efficiency dyads. This was one of two measures that showed a relationship between process measures and recall after a second learning opportunity.

Interestingly, explaining had a positive benefit on recall of links on Map1, especially for Efficiency dyads. Links that were explained were more likely to be

recalled than links that were not. This benefit had decreased by Map2 and was not apparent for Innovation dyads on either map redrawing test.

One potential benefit of explaining that did carry over to performance on Map2 was its association with shared knowledge. Explained links were more likely to be recalled by both members of a dyad than unexplained links. The more partners agreed on the links in their initial re-drawn map, the more they gained on the second map test. Given that this was the only other process measure that showed an association to recall after feedback and a re-learning opportunity, it suggests that agreement within groups might be an important variable or proxy for other learning processes.

Elaborating on previous work, these results raise the possibility that it is not explaining per se that leads to robust understanding. Allowing for a large degree of speculation, perhaps it is the effort to explain something that leads one to construct a schema or framework for new information. The more explaining one tries to do, the more one's schema should be elaborated and refined. Given accurate information, schemas should improve with more explaining. If this was the case, then individual links might show some benefit from explanation. The largest benefit of explaining, however, would be seen for participants who spent considerable time developing a schema that coherently linked the various concepts together. These participants might include: 1) the map-makers in Innovation who had time to reflect on the map and try to determine what links were missing, 2) Innovation participants in general who generated few enough errors as to have developed a productive schema, not just one that would have to be abandoned upon corrective feedback, and 3) the Efficiency participants who spent most time discussing the interconnections in their map during the Betty-trace phase.

If this hypothesis was accurate, it might help us understand why Innovation dyads showed minimal or no benefit from explaining (correct links). The suggestion is that they were already attempting to explain (silently to themselves) which concepts to include and how they fit together, so those links they explained aloud might not provide a good indicator of the state of their internal schema. Again, this is very tentative speculation, but it might be an interesting hypothesis to pursue given the focus that explaining has received as an activity that supports productive collaborative learning.

CHAPTER 3: THE CHI-SQUARE STUDY

For the second study in this dissertation, a number of changes were made based on the results of the first study. In particular, the task was quite different. I will describe it in some detail to ground the discussion of the rationale for these changes that follows. In this second study, college-age participants learned about the chi-square formula, shown below.

$$\chi^2 = \Sigma\,(E - O)^2/E$$

In this formula, E stands for "expected value," or the value one would expect based on chance (or based on prior knowledge). This value can be calculated in two ways. First, if you know the outcomes are equiprobable, such that all categories in a distribution should have the same number of occurrences, then E is the total number of observations divided by the number of categories. For instance, we expect a six-sided die that is fair to have each side appear equally often on average. Second, if we do not have a prior belief about the outcome probabilities, we can estimate E by multiplying row and column totals and dividing by the total number of observations (i.e. the grand total).[10] For instance, if we are interested in children's versus adults' preferences for different kinds of food, we have no prior reason to expect them to have equal preference for all categories of foods. Some foods are probably preferred over others (such as ice cream over tomato juice). In this kind of situation, we should estimate E with the second formula.

Returning to the chi-square formula, O stands for observed occurrences. For example, if someone threw a six-sided die 60 times, we would expect, on average, that each number on the die would appear 10 times. If we observed that "2" appeared 55

[10] Mathematically, this estimated E works by asking what is the probability of an occurrence in a particular cell or category (i.e. the row total divided by the grand total is the probability of being in a given row. The column total divided by the grand total is the probability of being in a given column. Thus, the probability of an occurrence in any given cell is the multiple of those two probabilities. The expected value then multiplies by the grand total to convert from a probability to a number of expected occurrences.

times and the other sides only appeared once each, we would strongly suspect the die was weighted or unfair.

The chi-square formula can be used to give a number to indicate how unfair this die was. The more the number of observed occurrences differs from the expected the more a distribution differs from chance. Distributions that differ more from chance (or expectation) have a larger chi-square value. It is important to divide by E in this formula because otherwise having a difference of five occurrences between E and O would give the same chi-square value regardless of if there were ten observations or ten thousand.

This is the kind of information that participants in the second study attempted to learn. The key difference between the Innovation treatment and the Efficiency treatment was that Innovation participants had to try to invent a formula for distinguishing contrasting data sets (such as those for fair versus unfair dice). Then they would receive a lesson on the chi-square formula with a worked example and a subsequent practice problem. The Efficiency participants received the lesson first. Then they did the practice problems. All participants received a nine-page learning packet with three lessons on the chi-square formula. The first page seen by Innovation participants asked them to rank the fairness of three dice, one which was clearly unfair, and two which were fair but had different numbers of tosses. The second page was the chi-square formula and lesson. The third page was the subsequent practice problem. For Efficiency participants, the lesson came first, then the contrasting-cases practice problems, and then the final practice problem. This sequence was repeated three times for different lessons about the chi-square formula.

Some participants studied the learning packet with a partner while others worked alone. After finishing the learning packet, all participants took a posttest individually with three types of questions: 1) chi-square calculation questions, 2) comprehension questions about what they read or inferred about the chi-square (such as why do you divide by E), and 3) far transfer questions with a preparation for future learning (PFL) component that required participants to adapt their knowledge of the chi-square to invent a formula for inter-rater reliability.

It was hypothesized that all conditions would do well on the calculation questions. Comprehension questions were expected to show some advantage for the Innovation

conditions, and the far transfer questions were expected to show the greatest benefits of Innovation, especially for dyads. In other words, as the questions moved from calculations to greater conceptual transfer, it was expected that the benefits of Innovation would become more apparent, especially for dyads.

These hypotheses were based on previous findings that students doing innovative activity tended to share information well and engaged in the task very actively. The benefits of this engagement and knowledge sharing would not likely show up on the most basic material because students in all conditions should do well on these questions. However, differences should be apparent on more advanced questions that require understanding the material from multiple perspectives and in explicit fashion. Having the opportunity to reconcile different interpretations, share insights, and make their thinking explicit through explaining to a partner should help Innovation participants, and especially dyads, understand the material deeply. According to this reasoning, dyads in the Innovation treatment should do best on the most conceptually difficult measures because they should not only gain useful insights from the material, they should also learn from their partner's insights and from having an opportunity to explain their perspective.

Efficiency participants would have the correct formula to work with from the start, so they would have fewer (if any) different approaches to the problems to reconcile. Efficiency dyads were expected to partition the material, each partner doing a subset of the practice problems in order to complete them quickly. On chi-square calculation problems, these groups were expected to do well because they would just be repeating what they had learned and practiced. However, on problems requiring greater understanding of how the chi-square formula works, they were expected to do worse, perhaps especially badly if they had worked in a dyad that focused strictly on partitioning the problems for speed.

Rationale for the design of Study 2

Observations from the first study about the assessment, the materials, and the design informed three major revisions for the second study that will be detailed below. First, minimal learning differences between conditions were found in the first study using

a recall measure. Related studies suggested that assessments of deeper understanding might be more sensitive to learning differences between conditions, so measures of transfer and deep understanding were added to the second study. Second, the learning task and materials in the first study may not have been optimal in terms of construct and ecological validity. Significant changes included the use of contrasting cases in a domain encountered by most high-school or college students, statistics. Third, the design was expanded to include individuals and dyads such that the unique contributions of collaboration could be measured.

Regarding the assessment component, the first study focused on recall measures. Recall may not target the differences in understanding that result from Innovation versus Efficiency kinds of instruction. Instead, it may be in more advanced cognitive processes, such as transfer and adaptability, where differences due to the treatments will be observed. Reflecting on related work, such measures were important to finding differences (Schwartz & Martin, 2004; Sears, in press). For example, using a PFL measure, Schwartz and Martin (2004) were able to distinguish between students who received a lecture and opportunity to practice from those who had an opportunity to invent. Half of each group received a resource question, a worked example with a practice problem, which all students could answer. Later, a target transfer question required students to adapt that prior example to solve a related and more difficult problem. Participants in the innovation condition succeeded at this task twice as frequently as those in the efficiency condition. Those participants in both conditions who did not receive the resource performed poorly on the task. Thus, the Innovation treatment made participants more prepared to learn (from the resource) than the Efficiency treatment.

Regarding the materials, results of the first study suggested that the Innovation and Efficiency tasks needed greater construct and ecological validity. A key goal of innovation is to help students generate insights into the key features of a domain that generalize to other situations. Generating copious errors is not part of this goal, so the second study implemented cycles of innovation and efficiency to keep participants in the Innovation condition from spinning off. In particular, rather than having participants do all the inventing first and then receive feedback in the form of lessons and worked examples, the material was divided into three sub-units. Each sub-unit or lesson involved

starting with contrasting case examples and attempting to distinguish the cases, often by perception first and then by inventing a formula. This invention phase was followed by a lesson on the canonical solution to such problems. A lesson ended with a final practice example where the canonical approach was to be applied (and the solution was provided for reference). The Efficiency materials were similar. They contained the three lessons on the chi-square formula, but they avoided the innovation component by always putting the contrasting cases and final practice problem after the lesson that showed how to solve such problems.

Previous studies showing success with innovation activities used contrasting cases (e.g. Schwartz & Bransford, 1998; Schwartz & Martin 2004). While contrasting cases may not be necessary for successful innovation activities, recent theorizing suggests they can promote a process that may be central to productive innovation—working to reconcile incommensurables (Schwartz, Sears, & Chang, in press). Incommensurables occur when two or more items are defined by different units (or systems) that cannot be directly compared. In order to compare them, one must reconcile the incommensurable by putting the items in terms of a new unit (or system) that relates the other two. For example, 3/4 cannot be compared directly to 5/6 because they are in different units (fourths and sixths). Only by finding a common multiple (12) can we put the units on the same dimension and make them comparable: $3/4 = 9/12 < 10/12 = 5/6$.

A second example involves a balance scale. In order for balance to be achieved, two dimensions, weight and distance, must be compared. These dimensions are in different units (kilograms versus meters, for example). It is only by multiplying the weight and distance (from the fulcrum), that we obtain a new unit related to torque, that puts these previously incomparable dimensions together on a single scale. A less mathematical example also is possible. Students learning to read often face the incommensurable of translating groups of letters to words. They are learning to put two previously familiar systems (the alphabet, and the spoken language) into a common system, text. Translating that text to meaningful phrases and transforming those to a new form, such as a concept map (e.g. Sears, in press), would represent other incommensurables and their reconciliation.

Working to reconcile incommensurables in math and perhaps in general is thought to help students notice the key features of a problem and thus appreciate how the canonical solution relates those features on a single dimension. In a balance beam study, Schwartz, Martin, and Pfaffman (2005) found that if students were encouraged to use math, they were more likely to invent the weight times distance relationship. This allowed them to relate the two dimensions on one scale (i.e. reconcile the incommensurable) and predict balance beam outcomes as well or better than adults.

For the second study in this dissertation, participants were given the support of contrasting cases to help them notice and attempt to reconcile different dimensions or features of the chi-square formula. For instance, having to distinguish two distributions of numbers that both have the same difference from the expected value but a different number of observations should encourage students relate (E – O) to E and better prepare them to understand the importance of dividing by E in the chi-square formula. Participants in both conditions received the contrasting cases. Unlike Innovation participants, Efficiency participants were not expected to appreciate the contrasts because they would already have a formula to apply. They would lack key incommensurables to resolve because they already had a solution.

While attempting to reconcile incommensurables may be critical to a successful innovation task, it probably also requires that the resolution to the incommensurable be generalizable to future situations. Betty's Brain may be an ideal tool for helping young students learn about causal reasoning, but it may not be ideal for promoting older students' learning of concepts that can be applied to many different problems of a given form. For this reason, the second study abandoned the teachable agent component and turned to a domain that would permit the development of more generalizable knowledge.

Specifically, while the first study taught college students about cholesterol and heart disease (not a typical topic for most 18 to 22 year olds), the second taught a statistics procedure that is introduced in most if not all introductory statistics courses. In addition, rather than giving the Efficiency participants answers to copy into a concept map, this time they had the more typical activity of receiving a math lesson and then practicing what was taught. Thus, the second study aimed at having more clearly

distinguished innovation and efficiency treatments while also pushing for greater ecological validity.

Regarding the design, the first study only included dyads because it was primarily interested in the types of interactions that might underlie learning benefits of innovation versus efficiency. It was also an attempt to learn more about what features support productive innovation. For those purposes, focusing on dyads was fine. However, for the hypotheses about learning benefits of innovation that were at the heart of this dissertation, it was essential to include individuals and dyads in the second study. Without the 2x2 design with individuals and dyads receiving the Innovation and Efficiency treatments, we would not be able to conclude whether learning benefits were due to Innovation being a better educational approach in general, or whether it provided particular benefits for dyads, as was originally hypothesized. In other words, by including individuals, it would be possible to measure the relative benefit of being in a group for both treatments.

Methods

Participants—University students, mostly undergraduates, were recruited from a paid-subjects email-list and flyers around campus. The recruitment information requested students with a limited background in statistics. This meant that participants had either taken no statistics courses or an introductory level course, such as AP statistics in high school. A total of 76 students participated. Participants worked alone (40 total) or with a partner of the same gender (36 total) during the experiment. Participants were assigned to dyads based on who signed up for each time slot. In one dyad, the partners were friends, and in another they knew each other socially. These dyads were assigned to opposite treatments. Twenty-four women worked alone, 24 worked in pairs. For men, 16 worked alone and 12 worked in pairs. Individuals and dyads were randomly assigned to one of two conditions: an Innovation condition, or an Efficiency condition.

Materials—The study materials included a learning packet on the chi-square formula (See Appendix A) and a posttest (See Appendix B). The learning packet consisted of three units about different aspects of the chi-square formula. Each unit contained three pages: a Lesson page, a Problems page, and a Final Practice Example

page with answers provided at the bottom of the page. For Innovation, the sequence of pages for each unit was: 1) Problems, 2) Lesson, and 3) Final Practice Example. For Efficiency, the sequence was: 1) Lesson, 2) Problems, and 3) Final Practice Example. In other words, Innovation had to try to figure out a formula before getting the canonical solution, while Efficiency received the formula and had a chance to apply it. Table 4 shows the topic of each lesson in the packet. A key feature of these materials was how the Problems pages had contrasting cases designed to highlight key features of the formula(s).

Table 4
Learning Packet Summary (three pages per Lesson)

Lesson	Topic	Formula(s)	Key Points or Contrasts
1	The Chi-Square Formula	$\chi^2 = \Sigma (E - O)^2/E$ $E = \frac{\text{Total Observations}}{\text{\# of Cells}}$	1) Used to compare a frequency distribution to chance. 2) Why dividing by E is necessary.
2	Chi-Square when E must be estimated	$E = \frac{\text{Row Total*Column Total}}{\text{Grand Total}}$	1) The expected value (E) should be estimated when it is unknown. 2) When it is given (by fair dice or other situations assumed to have equal probabilities), the previous formula is appropriate.
3	Chi-Square when Sample-size changes	$\chi^2 = \Sigma (E - O)^2/E$ $\chi^2 = \Sigma E[1 - (O/E)]^2$	1) When (e-o) is constant, chi-sq. decreases with increasing sample size. 2) When (o/e) is constant, chi-sq. increases with increasing sample size.

A digital video camera and tapes recorded partners' interactions. The amount of time participants spent on each page of the nine page learning packet was also recorded during the learning phase. The outcome measures of interest consisted of three types of items on a seven-problem posttest. Two problems required calculations of the chi-square formula, three involved comprehension questions about where and how the formula works, and two involved a difficult transfer to the related statistics topic of inter-rater reliability. An important feature of these far transfer measures was that they were designed in PFL fashion (Bransford & Schwartz, 1999; Schwartz & Bransford, 1998; Schwartz & Martin, 2004). The first problem introduced the new type of problem, while the second provided a more difficult case in which those same principles applied. The

PFL idea is that only participants who were prepared to learn from the first problem would be able to answer the second one correctly. This approach allows one to estimate what kinds of instruction are better at preparing students for future learning.

Procedures—Participants worked in a small room with a couch and shared desk. Some individuals worked at a desk outside of the small room if the room was already filled. The experimenter (the author) sat in close proximity to the participants to answer their questions and make sure they progressed through the materials without looking ahead. All participants completed the learning packet. They were encouraged to understand what was on each page and to aim for about five minutes per page in order to complete the activities within the allotted 90 minutes. Dyads were instructed to work together to understand the material in the packet they had to share. The key difference between conditions was that participants in the Innovation conditions attempted to invent solutions to the Problems before receiving the Lesson for each of the three units in the learning packet. Participants in the Efficiency conditions read the Lesson before doing the Problems. For each unit, all participants did the Final Example problem after completing the lesson and practice pages. To keep participants from blurring the distinction between conditions, they were told to complete each page in the packet before going to the next page and to look back only if necessary (such as to recall the formula).

By attempting to construct solutions to the Problems before learning the canonical solution, the Innovation participants were meant to bump up against the challenges in each problem and the contrasts in the examples. By working to resolve these incommensurables, it was hypothesized that they would be more prepared to notice and understand how the canonical solution worked (Schwartz, Sears, & Chang, in press). For example, if participants used the previously taught version of the expected value calculation on the Problems in the second unit, they would obtain the same chi-square value for both distributions. This would push them to consider a different approach to the problem to distinguish the cases. Perhaps more importantly, it should help them realize the limits of the original formulation. Upon seeing the lesson for that unit, these participants should be more prepared to understand why that new way of calculating the expected value was important and when it was applicable.

For Efficiency participants, reading the Lesson and then doing the Problems would give them an opportunity to practice what they learned. Because they would have the canonical solution given to them, they were not expected to bump up against the incommensurables presented by the contrasting cases. For this reason, they were not expected to develop a deep appreciation for how each formula worked. Instead, by having an opportunity to practice the canonical solution, they were expected to develop fluency with its application.

To facilitate computational fluency for Efficiency and the noticing and attempting to reconcile incommensurables for Innovation, participants were asked to calculate a numerical answer for each problem using their formula. In some cases, Innovation participants struggled to do this, so text explanations were acceptable. Participants were allowed to use a calculator to assist them in their calculations.

After completing the learning packet, participants took a short break (about five minutes) before taking the posttest. The experimenter instructed them to spend 25 minutes on the posttest, that is was difficult, to try their best, and to work alone. Participants had access to a calculator but were told it was sufficient to set up the solution to the problem with all the numbers in it such that the experimenter could take the final step of calculating the answer. Some participants finished early, and none took longer than 30 minutes. The experimenter periodically notified participants of how much time remained for them to complete the test. Table 5 summarizes the phases of the study.

Table 5
Design of Study 2

Step	**Context**	**Innovation**	**Efficiency**	**Time**
1	Alone / Dyads	9-page Learning Packet on the Chi-Square Formula		35 to 65 min.
		1) Problems (invent) 2) Lesson (instruction) 3) Final Example (reinforce)	1) Lesson (instruction) 2) Problems (apply) 3) Final Example (reinforce)	
		Short Break		~5 min.
2	Alone	Posttest (7 problems: 2 calculations, 3 comprehension, 2 far transfer)		25 min.

Results

Before turning to the outcome measures, a brief characterization of participants' activities during the learning phase is appropriate. Dyads receiving the Innovation treatment discussed the material and shared their perspectives frequently, as expected. Discussion was often interspersed with periods of silent activity as partners tried to gain insight into the material. Some participants were able to develop formulas similar to the chi-square formula; however, most struggled. A somewhat surprising finding was that the Efficiency dyads also showed elaborate discussions on occasion. My impression was that they showed less time silently working to figure out how to approach a problem and more time checking each other's calculations, explaining them, or trying to understand what their partner was doing (sometimes while their partner sped along through the problems, seemingly quite comfortable with the formulas they were applying). Future video and transcript analyses should add rigor to these observations.

Time

A multivariate analysis of variance (MANOVA) revealed no significant differences in the amount of time taken for the learning packet or the posttest, though the means were slightly lower for Innovation as shown in Figure 15.

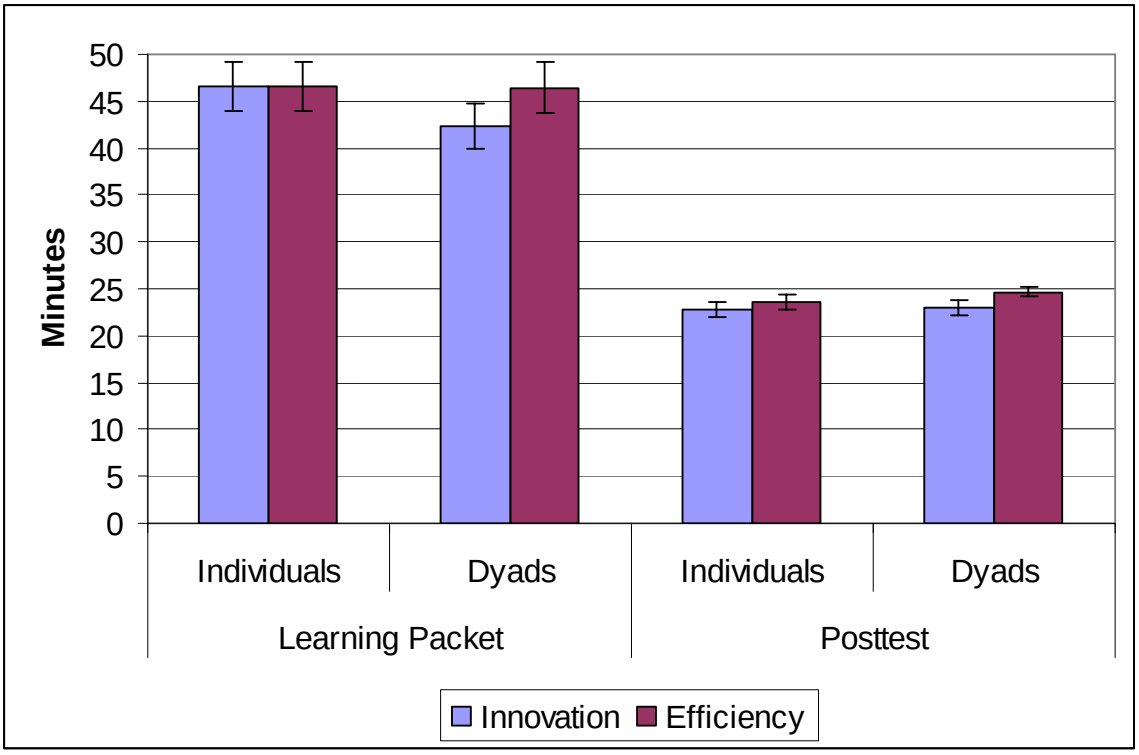

Figure 15. Mean time spent learning and completing the posttest. (Error bars represent standard error of the mean).[11]

Coding scheme

Table 6 shows the features of each problem on the posttest that were coded and how many points they could receive (or if they were just noted). As an example, on the first problem, participants were to calculate the chi-square statistic for a given data table showing the accuracy of two tests at predicting who had a disease. Because we do not have an *a priori* reason to expect these two tests to be equally accurate, we should not assume that all the cells in the data table would have the same expected value. It is more appropriate to calculate the expected value for each cell using the (row x column) / grand total approach described above.

[11] All subsequent error bars represent standard error of the mean unless otherwise noted.

Question 1: Drug test effectiveness.
Test A and Test B differ substantially in how expensive they are, so doctors wanted to find out whether they differed in how accurately they diagnosed patients with a particular disease. Do Test A and Test B differ significantly in their ability to predict who has the disease?

	Yes	No
Test A	10	10
Test B	6	14

Sample Solution 1: Using the less-appropriate expected value formula
E = (10 + 10 + 6 + 14)/4 = 10

Chi-square = $0/10 + 0/10 + (6-10)^2/10 + (14-10)^2/10 = 3.2$

Scoring:
1 point out of 2 for Expected Value component b/c it was a correct approach but not optimal (50%).
1 point out of 1 for the Chi-square component (100%).
Score on Question 1 is the average percent correct of the two components = (50% + 100%)/2 = **75%.**

Sample Solution 2: Using the correct expected value formula
Expected Value using Row*Column/Grand:

	Yes	No
Test A	(20*16)/40 = 8	(20*24)/40 = 12
Test B	(20*16)/40 = 8	(20*24)/40 = 12

Chi-square = $(8-10)^2/8 + (12-10)^2/12 + (8-6)^2/8 + (12-14)^2/12 = 4/8 + 4/12 + 4/8 + 4/12 = 1.667$

Scoring:
2 points out of 2 for Expected Value component b/c it was the optimal approach (100%).
1 point out of 1 for the Chi-square component (100%).
Score on Question 1 is the average percent correct of the two components = (100% + 100%)/2 = **100%.**

Figure 16. Example scoring of Question 1.

Figure 16 shows how two solutions for Question 1 would be scored. If participants calculated the expected value by simply dividing the total number of observations by four (the number of cells), then they would get one point. If they calculated the expected value by estimating what it should be for each cell based on row and column totals, they would earn two points. This was not an arbitrary system. Because Question 1 involved a situation where *a priori* expectations about test accuracy were not available, it was appropriate to use the expected value calculation that estimated a value of E for each cell using row and column totals. The other approach (shown in Sample Solution 1) to calculating E is appropriate for situations in which a known percentage of observations are expected to fall into each category (e.g. each side on a fair die should appear equally often, on average). Question 5 (the other calculation problem)

rewarded this version of the expected value calculation with two points and the other with only one, providing a counterbalance to the scoring of Question 1. Thus, students who simply applied one version of the expected value formula to the calculation problems, regardless of which version, would get one problem right and the other wrong.

Each question was divided into components or "Coded Features" that were weighted equally in that question's score. For instance, in the calculation questions, the expected value comprised half of the score and the chi-square calculation comprised the other half. To make these halves equally weighted, each component was transformed into a percent; then these percents were averaged to find the question score (e.g. 50% and 100% average to 75%). This approach was necessary because otherwise the expected value component, which was on a 0 to 2 scale, would be weighted more heavily in the Question 1 score than the chi-square component, which was on a 0 to 1 scale.

Because the Efficiency instructional method aims for increasing speed and *accuracy*, if a calculation error was made on a component, half a point was subtracted from the score for that component. If participants set up the equation correctly with all the numbers and then made an error in computing that value, that computational error was ignored. However, if participants entered a miscalculated number into their formula, such as the wrong grand total in the expected value formula or a miscounted number of agreements in the chi-square formula, they were penalized with a half-point deduction on that component. The idea behind this scoring was that statistical software programs can compute the chi-square statistic, but one still needs to enter the values correctly. Any calculations that were described with text only were given no credit except on the far transfer questions where a canonical solution had not been taught.[12]

[12] Two participants in the Innovation condition answered the PFL target problem using text. The first received no credit for her answer except for being penalized for suggesting an approach that involved negative transfer. The second received credit for the calculation of the expected value she described (and had computed successfully in the resource problem, Question 4). Despite successfully adapting the chi-square formula on the resource problem and suggesting a comparison between expected and observed agreements on the target problem, her description of that comparison was too general to receive any credit on that component.

Table 6

Coding Scheme for Posttest

Question #	Question Type	Question	Coded Feature / Answer	Points	Deductions
1	Calculation	Do these two tests for disease differ in their accuracy?	E = Total/#cells = *(negative transfer)* E = R*C/Total	1 *noted* 2	-.5 (calc. error)
			Chi-sq = (E – O)^2/E	1	-.5 (calc. error)
5	Calculation	Are these fair dice magnetized?	E = Total/#cells E = R*C/Total = *(negative transfer)*	2 1 *noted*	-.5 (calc. error)
			Chi-sq = (E – O)^2/E	1	-.5 (calc. error)
2	Compre-hension	Why divide by E in the Chi-square formula?	To account for sample size (i.e. make cells comparable)	1	
3	Compre-hension	How does sample size affect the chi-square value?	If (E – O) is constant, Chi-sq. decreases with increasing sample size.	1	-.5 if too specific [e.g. if (E – O) stays close to 0].
			If O/E is constant, Chi-sq. increases with increasing sample size.	1	-.5 if too specific [e.g. if (E/O) stays close to 1].
7	Compre-hension	When should someone use the chi-square formula?	To compare results to chance. Test for independence of variables.	1	
4	Far Transfer Resource Problem	How much beyond chance do two mechanics agree on their diagnoses of 50 cars (as shown in this reliability matrix)?	E = Total/#cells E = R*C/Total	1 2	-.5 (calc. error)
			Subtract? (Obs.– Exp.) Agrmnts.	1	
			Normalize?	1	
			Subtract Well?	1	
			Normalize Well?	1	
			Apply chi-sq. without regard to agreement or disagreement cells *(negative transfer)*	*noted*	
6	Far Transfer Target Problem	How much beyond chance do these chemists agree on the color that each element released when burned?	Total number of Agreements (or disagreements)	1	-.5 (calc. error)
			E = Total/#cells E = R*C/Total	1 2	-.5 (calc. error)
			Subtract?: (Obs. Agrmnts – Exp. Agrmnts)	1	
			Normalize?:	1	
			Subtract Well?	1	
			Normalize Well?	1	
			Apply chi-sq. without regard to agreement or disagreement cells *(negative transfer)*	*noted*	

Reliability and validity of the coding-scheme and the test

A random sample of 20 posttests, with three to four per condition, was selected for comparison for inter-rater reliability calculations. Across all 21 features, percent

agreement was 93%. On the Calculation questions, percent agreement was 90% (average Kappa = .828; minimum agreement of 85% on any given feature). On Comprehension questions, percent agreement was 93% (average Kappa = .819; minimum agreement of 80%). On Far Transfer questions, percent agreement was 95% (average Kappa = .826; minimum agreement of 85%). Thus, the coding scheme was reliable across raters.

The 19 scored features yielded a reliable overall assessment of students' understanding of the chi-square formula (alpha = .81; each item weighted equally). For the analyses that follow, the subset of the 19 features that composed a given question on the test was aggregated in equal-weighted fashion. For instance, the expected value component and the chi-square component were each weighted as 50% of the score for question one, as described previously. A similar procedure was used to aggregate scores by question-type (i.e. Calculation, Comprehension, Far Transfer). For example, question one and question five each composed 50% of the score on the Calculation question-type. The scores on each question-type could range from 0% correct to 100% correct. To support the validity of this grouping, correlations between the types of questions were analyzed. Only the calculation and comprehension question-types were correlated; $r = .316$, $p = .005$.

Overview of analyses and results

Recall that it was hypothesized that all conditions would do well on the calculation problems. As the problems moved away from the taught calculations and toward deeper conceptual understanding, it was expected that the Innovation conditions would show greater performance. By this logic, comprehension problems would show some advantage for Innovation and the far transfer problems would show the greatest advantage. In addition, it was expected that Innovation dyads would show the greatest advantages on the far transfer problems and outperform their peers who worked alone. Many of these hypotheses were supported.

To test the hypotheses, a repeated-measures multivariate analysis of variance (MANOVA) will serve as an umbrella test by which to compare individuals and dyads in Innovation versus Efficiency treatments on the three question-types (calculation, comprehension, and far transfer). Differences found with the MANOVA will be

examined further with independent t-tests. Next, special attention will be given to the far transfer questions and other measures of deep understanding because these were expected to reveal the greatest differences between conditions. Specifically, rates of negative transfer will be analyzed with chi-square tests while PFL measures from the far transfer questions will be examined with a repeated-measures MANOVA. Included in this portion of the analysis will be a comparison between the performance of real dyads and nominal dyads.

For the analyses that follow, each participant that studied in a dyad will be treated as a separate unit (i.e. independent) because despite studying together, they took the tests alone. This approach has precedent in previous well-known studies of collaborative learning (e.g. Barron, 2003; Phelps & Damon, 1989). In addition, practical reasons support this approach. First, counting the dyads as a single unit by averaging partner scores would decrease the sample by half. This conservative estimate may be unwarranted given that the average difference between partners' scores on each of the three question types was .33, .28, and .26 while the average difference between individuals under every possible dyad combination was .30, .34, and .24 for calculation, comprehension, and far transfer questions, respectively. In other words, the difference between real partners' scores was about the same as randomly paired individuals. A more appropriate approach would involve using a multivariate hierarchical linear model (MHLM); however, no statistical packages that I am aware of can do a three-level MHLM (Dr. A. Bryk, personal communication, May 30, 2006). Instead, I will note that univariate HLM tests produced similar outcomes as the repeated-measures MANOVA.

MANOVA Results

Figure 17 shows the performance of the Innovation versus Efficiency conditions on the different posttest question types. Figure 18 shows those results further subdivided by Individuals versus Dyads.

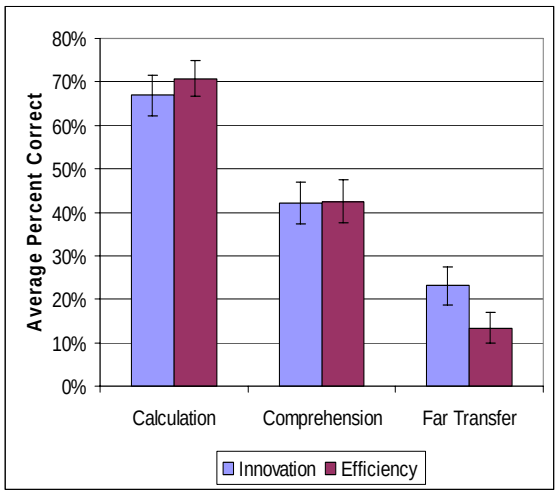

Figure 17. Performance under Innovation versus Efficiency learning conditions on learning and transfer measures.

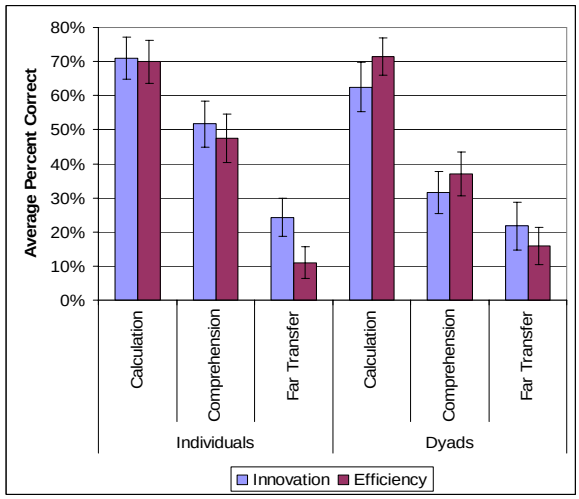

Figure 18. Individuals' versus dyads' performance on posttest measures by condition.

A repeated-measures MANOVA revealed no significant differences or interactions between conditions on the multivariate measures. However, taking advantage of *a priori* hypotheses, the results of planned within-subjects simple contrasts were examined. These contrasts compared performance on the Far Transfer measure to performance on the Comprehension measure and performance on the Calculation measure, and they revealed interesting differences between conditions. First, a marginally significant difference was found on patterns of performance on the Far Transfer versus Calculation measures for the Innovation versus Efficiency conditions; $F(1, 72) = 3.04$, $MS = 0.35$, $p = .086$. This result was due to Innovation significantly outperforming Efficiency on the Far Transfer measure; $t(74) = 1.75$, $p = .043$ (one-tailed), while scoring non-significantly lower on the Calculation measure; $t(74) = -.60$, $p = .274$ (one-tailed). [13]

Second, a marginally significant difference was found on patterns of performance on the Far Transfer versus Comprehension measures for Individuals versus Dyads; $F(1, 72) = 3.96$, $MS = 0.52$, $p = .050$. This result was due to Dyads scoring significantly lower on the comprehension measures than Individuals; $t(74) = 2.33$, $p = .023$, while

[13] All t-tests reported here use two-tailed criterion for significance unless stated otherwise. In this case, a one-tailed test was appropriate because this was a hypothesized result.

scoring non-significantly higher on the Far Transfer measures; $t(74) = -.22$, $p = .830$. This result was unexpected.

Measuring "Deep Understanding" with Negative Transfer and PFL Measures

For my purposes here, deep understanding will be defined as knowing what something is and what it is not. In cognitive terms, negative transfer indicates a failure to recognize what something is not. It is the application of prior knowledge to a problem that requires a different approach. We know from previous work that experts show less negative transfer than novices (Novick, 1988). In this experiment, negative transfer could occur on any of the questions that involved computation (Questions 1, 4, 5, and 6).

Students could show negative transfer on the calculation problems (Questions 1 and 5) by using the wrong version of the Expected Value calculation. Because these questions required different versions of the calculation, students who used one method consistently would show negative transfer on one (and only one) of these measures. The modal outcome across conditions was one negative transfer on these calculation problems.

On the far transfer questions (Questions 4 and 6), negative transfer involved applying the chi-square formula to the data without regard to agreements or disagreements. A better approach would involve applying the chi-square formula only to the agreement cells of a reliability matrix (or just to the disagreement cells). This approach would distinguish the source of any differences from chance as being due to more agreements (or more disagreements) than expected. The modal outcome across conditions was zero negative transfers on these far transfer questions.

Summing the negative transfer results from the questions described above, amount of negative transfer could range from zero to four. The distribution of responses was not normal, as shown in Figure 19. To characterize the distribution, scores were grouped into two categories: 1) those who showed "Zero to One" negative transfers, and 2) those who showed "Two to Four." As mentioned above, most students showed one negative transfer on the calculation problems, and most showed none on the Far-transfer questions, so dividing participants into those showing one or fewer negative transfers versus two or more seemed reasonable. (Similar results as those presented below were obtained when using three categories of "zero, one, and two or more").

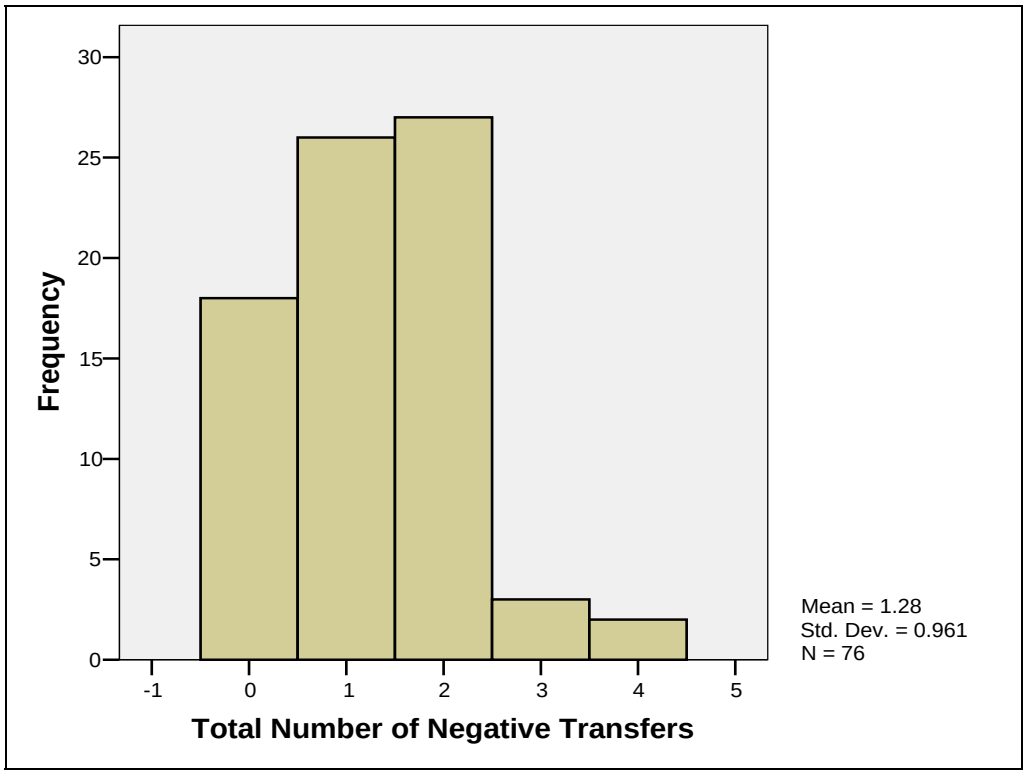

Figure 19. Number of negative transfers on the four questions that required mathematical computations. More negative transfer suggests less understanding of when a given formula does *not* apply.

Table 7 shows how each condition performed on the negative transfer measure. Chi-square analyses indicated that Innovation participants showed significantly less negative transfer than Efficiency participants; $\chi(1) = 5.40$, $p = .020$. This result was driven primarily by the Innovation Dyads relative to Efficiency Dyads; $\chi(1) = 5.6$, $p = .018$, rather than by Individuals; $\chi(1) = .92$, $p = .337$.

Table 7
Degree of Negative Transfer

	Number of Negative Transfers	Innovation	Efficiency
Individuals	Zero to One	13	10
	Two to Four	7	10
Those who studied in Dyads	Zero to One	14	7
	Two to Four	4	11

While negative transfer provided a measure of whether participants understood the chi-square formula well enough to know when or how it did *not* apply, it did not indicate whether participants understood the material well enough to know how to adapt it to difficult transfer problems. Adapting a formula to a new type of problem suggests a deep level of understanding. I would not expect most participants to achieve this within less than an hour of study. The surprising finding here is that some participants did. In particular, the Innovation participants, especially the dyads, adapted their prior knowledge of the chi-square formula to make an adequate measure of inter-rater reliability.

Before describing how such adaptation was scored, the Far Transfer problems will be described in greater detail. These two problems followed the preparation for future learning (PFL) design: 1) a resource problem was given, then 2) a related but more difficult target transfer problem (Bransford & Schwartz, 1999; Schwartz & Bransford, 1998; Schwartz & Martin, 2004). For each, participants had to calculate how much two raters agreed beyond chance. To do this successfully, participants had to transfer and adapt their knowledge of the chi-square procedure.

The first question was a resource question because it gave participants three tools to help them construct their answer, and these tools could also be used to help solve the target problem if participants recognized its similarity. The three tools were: 1) how many times the raters agreed, 2) a 3x3 reliability matrix (i.e. one that showed agreements along the diagonal), and 3) a prompt to calculate how much the raters should agree *by chance alone* before calculating how much they agreed *beyond chance*.

The second question was a target transfer problem because participants had to go beyond what they did in the resource problem. This time, they had to count how many agreements there were between two raters, transform raw data into a 5x5 reliability matrix or another representation for finding the expected values, and make the connection between expected agreements versus actual agreements. In other words, this problem lacked the resources of the former problem.

In order to solve these challenging problems adequately, participants had to calculate the expected number of agreements (or disagreements) and subtract this from the number of observed agreements (or disagreements). If they did this, they would

receive one point. If they also normalized the difference by dividing by the expected number of agreements (or disagreements) or the grand total, they would receive a second point. Thus, a key factor in solving these PFL problems adequately was whether participants used the expected value formula they learned for the chi-square and applied it selectively to the agreement cells (or disagreement cells). If they applied it to all of the cells and then calculated the chi-square (as they had learned to do previously), they would show negative transfer and would not have an adequate answer.

As described above, participants could score from zero to two points on each PFL problem. These scores were then transformed into a percent. A repeated-measures MANOVA revealed three marginally significant effects: 1) Innovation participants tended to score higher than Efficiency participants; $F(1, 72) = 3.15$, $MS = .422$, $p = .080$; 2) members of dyads tended to outperform individuals on the target question while doing less well on the resource question; *Wilks' Lambda* $= .963$, $F(1, 72) = 2.79$, $p = .099$; and 3) Innovation dyads scored higher than Innovation individuals on the target problem despite scoring lower on the resource problem while Efficiency dyads scored slightly higher on both than Efficiency individuals. This third difference was a 3-way interaction between question-type (resource vs. target), task-type (Innovation vs. Efficiency), and context (individual vs. dyad); *Wilks' Lambda* $= .959$, $F(1, 72) = 3.11$, $p = .082$.

These results can be seen in Figure 20. The Innovation dyads were the only condition to show improvement from resource problem to target problem. Notably, the two Innovation-dyad participants who answered the target problem well without having answered the resource problem well had partners who answered both problems well. For the rest of the participants, only those who did well on the resource problem did well on the target problem. Very few of them made that transfer successfully (5 of 10 for Innovation and 2 of 5 for Efficiency). Unlike participants in Efficiency, those in Innovation who made the transfer were not just students who aced the posttest. Innovation participants averaged .86 standard deviations above the mean on the posttest while those in Efficiency averaged 1.44 standard deviations above the mean. Thus, the target question built from the resource, as expected, and success on it was not limited to the tail-end of the distribution for Innovation participants. As will be discussed further below, this PFL measure showed the greatest benefits of Innovation on dyadic learning.

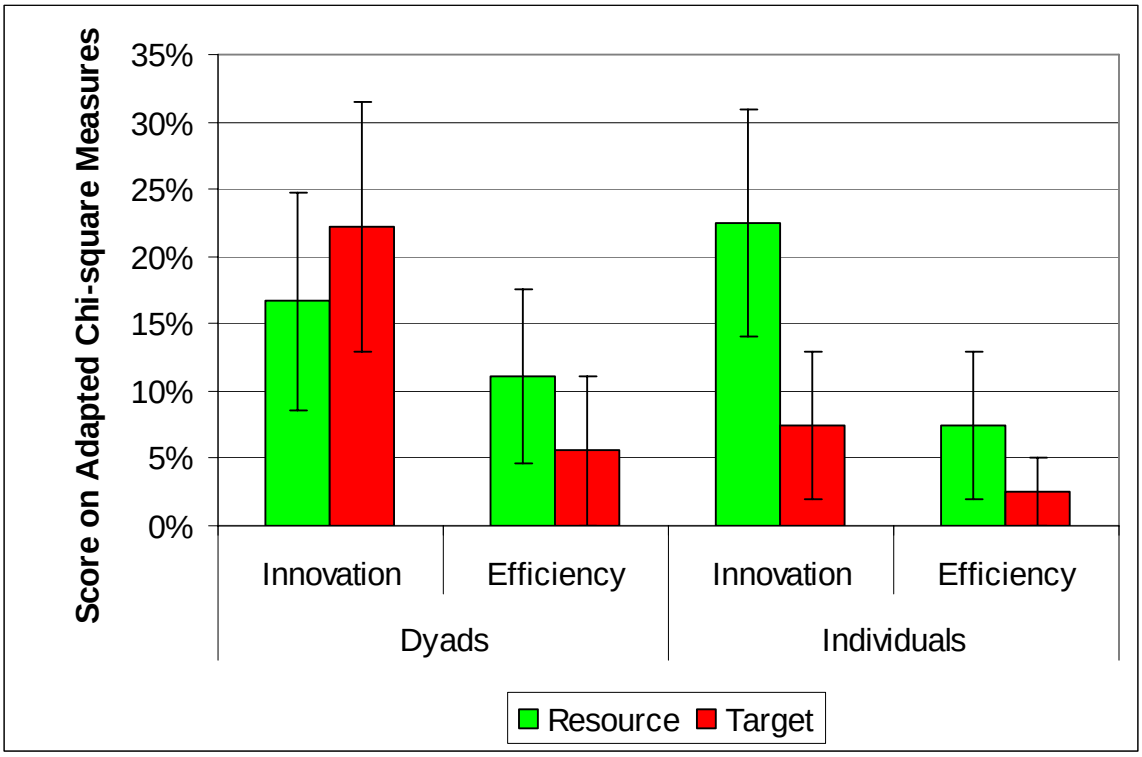

Figure 20. Participants in Innovation dyads were the only ones that gained from the resource problem to the target problem. Innovation scored marginally higher than Efficiency across the two far transfer problems.

Dyads' Success: Nominal versus Real Group Comparisons on the PFL Measures

As mentioned in the introduction, most comparisons between groups and individuals involve comparing averages—the average group to the average individual. Often this shows benefits of groups. The most stringent comparison involves comparing real groups to nominal groups or mathematically "grouped" individuals under truth-wins assumptions. As quick review, truth-wins means that if one member of the "group" answered the question correctly, everyone in that group would be assumed to answer it correctly. In other words, the right answer is always accepted by the group, or the truth always wins. In reference to these results, I will use "nominal dyads" and "truth-wins dyads" interchangeably. See the footnote below if interested in a more detailed explanation of the truth-wins calculations.[14] The value of using a comparison between

[14] The truth-wins to real dyads comparison was calculated as follows. The average score of each individual in a dyad was compared to the average score of each individual in a truth-wins dyad. The scores for the combined PFL measures and the target measure were calculated in similar fashion, so I will explain the

real dyads versus truth-wins dyads is that if real dyads exceed truth-wins dyads we know they have shown process gain. In other words, something in their interaction has allowed them to move beyond what they could have done as individuals, even with perfect knowledge sharing. As a result of interaction, they would have constructed some new understanding that neither one had before.

Figure 21 shows the performance of real dyads and nominal dyads under each task-type on the combined PFL measures described above. The nominal dyads' scores were modeled based on the scores of the participants in the individual conditions. Thus, Innovation nominal dyads were modeled on Innovation individuals' scores while Efficiency nominal dyads were modeled on Efficiency individuals' scores.

Although the Innovation dyads did not exceed the truth-wins Innovation dyads, they nearly doubled the success of the truth-wins Efficiency dyads. Notably, the real Efficiency dyads performed nearly as well as the nominal ones, suggesting that PFL

---

more complex one (the combined) first. For the Innovation individuals, 20 participants can be combined into 190 distinct dyads ("20 choose 2" = 20C2 = 20!/(20-2)!*2!). One way to think of this is that the top individual can combine with 19 other individuals to make 19 dyads. The next participant can combine with the remaining 18 individuals to make 18 dyads. The next can join the remaining 17 and so on.

The score for each truth-wins dyad then consisted of their joint performance on the four measures (each scored 0 or 1) on the combined PFL measures. If either one of the members got the correct answer on a measure, they would be scored as getting that measure correct. For instance, did one or both members of the dyad subtract expected agreements from observed agreements on the resource problem? If so, that pair would receive credit on the first of the four measures. Did one or both normalize adequately on the resource problem? Did either succeed on the similar measures on the target problem? Under the truth-wins assumptions if either member got credit for an item, the "dyad" would receive credit for that item.

The top performing individual in Innovation scored perfectly on the combined measure (4 points out of 4). This meant that all 19 dyads this person could participate in would receive a score of 4 out of 4, or 100%. The next best individual (who scored 2 out of 4) combined to make a score of 3 out of 4 with one other individual. With the 17 remaining individuals, he made a score of 2 out of 4, or 50%. The third best individual performed the same, so that person combined with the one to make 3 of 4 and with the 16 remaining to make 2 of 4. Of these first 54 possible dyads (out of 190), 19 scored 100%, two scored 75%, and 33 scored 50%. After computing the scores for the remaining 136 nominal dyads, all 190 scores were averaged. That process yielded the 27.5% shown in the figure. This same process was applied to the two items on the "target" measure, and for the individuals in the Efficiency condition.

For those who might be concerned that this process results in a comparison between the average performance of *individuals* who studied in dyads to nominal *dyads*, altering the computation to count the nominal dyads as "individuals who studied in nominal dyads" produced similar results (slightly worse performance for the nominal conditions, actually). The way this altering of the computation can be done is as follows. For the first individual (of 20), they can participate in 19 dyads, effectively teaching all 19 how to do the problem. That means 20 "nominal individuals" would score four points (rather than 19 dyads, as before). For the next individual, they could interact with 18 dyads, so 19 individuals overall would have their knowledge. Continuing in similar fashion, this process is equivalent to averaging the score of "21 choose 2" combinations of people, and it yields 210 "nominal individuals." Because the "nominal dyads" calculation gave the more stringent comparison, I chose it for these analyses.

measures reveal differences in learning that are not usually found with traditional measures both for individuals and especially for dyads. Figure 22 shows performance only on the target problem. Here the real Innovation dyads outperformed all others.

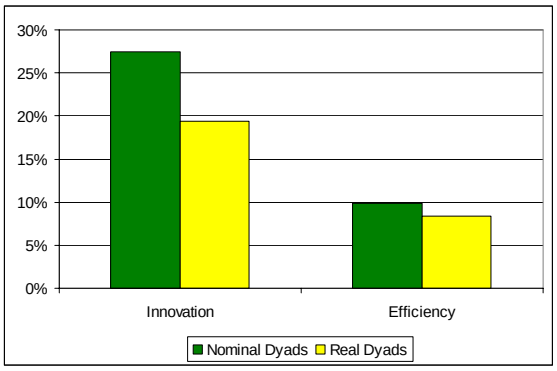

Figure 21. On the combined PFL measures, real Innovation dyads nearly double the performance of truth-wins Efficiency dyads.

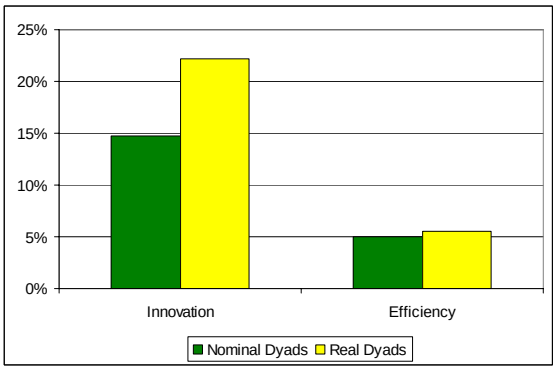

Figure 22. On the target question, real Innovation dyads exceed all others.

Re-analyzing the Results: The Necessity of Efficiency for Innovation

A certain amount of efficient knowledge may be necessary for innovation activities to be educationally beneficial. Pursuing this notion, we can re-analyze the data after removing participants who failed to show or acquire that efficient knowledge. In this study, the efficient knowledge was of how to compute the chi-square formula (i.e. the calculation questions). We can see in Figure 23 that a few students in each condition did not learn how to calculate the chi-square formula well. In particular, every condition showed a gap between participants scoring at or above 50% versus those scoring below 50% on the calculation questions.

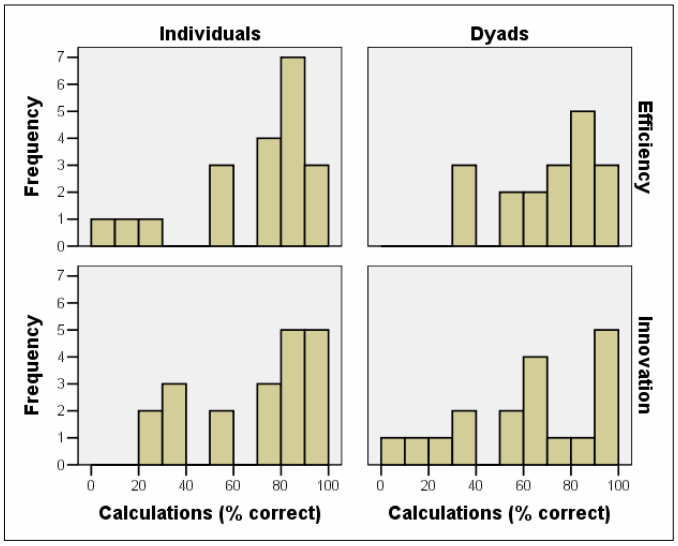

Figure 23. Scoring less than 50% suggests a failure to acquire the most basic understanding.

If it is the case that a certain amount of efficient knowledge is necessary to benefit from Innovation, then removing participants without that knowledge should reveal greater benefits of Innovation activity for learning. Below is a re-analysis of previous results separating from the rest those participants who scored below 50% on the calculation problems. This re-analysis strongly supports the conjecture.

In Figures 24, we can see that Innovation participants that failed to learn the calculation showed limited success on all of the other problems. More importantly, the pattern in Figure 25 suggests that learning the basic calculation allowed for significant benefits from Innovation. Innovation showed greater performance relative to Efficiency once only those who acquired the efficient knowledge were compared. This gain was particularly noticeable for the Innovation dyads relative to Efficiency dyads. Thus, Innovation did not cause a tradeoff between learning the basic material and learning the advanced material; instead, it appears to have built upon the basic material to help students deepen their understanding.

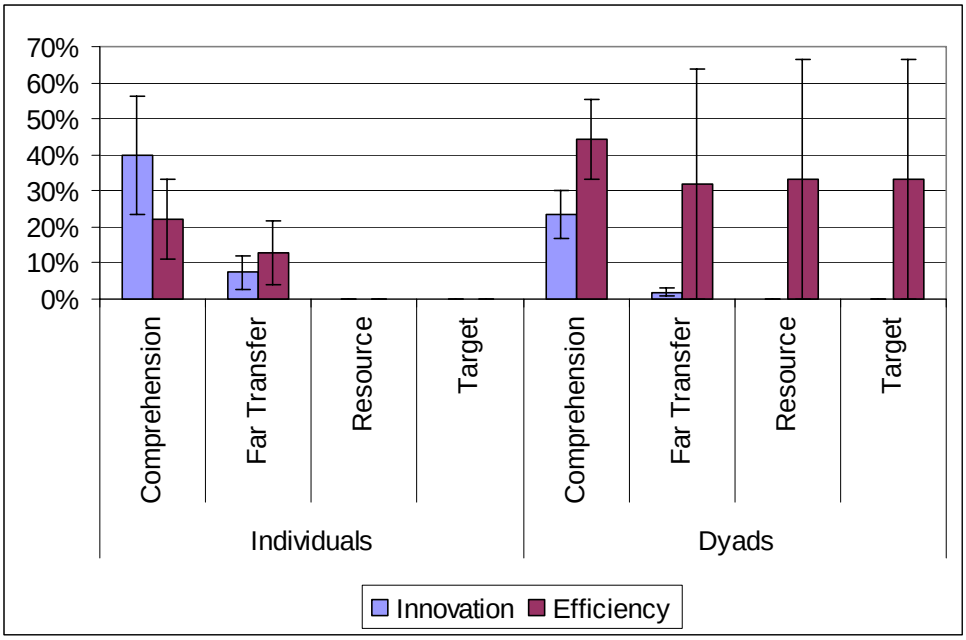

Figure 24. Innovation participants who failed to learn the basic calculation (i.e. scored below 50% on the calculation measures) showed low scores on all measures.

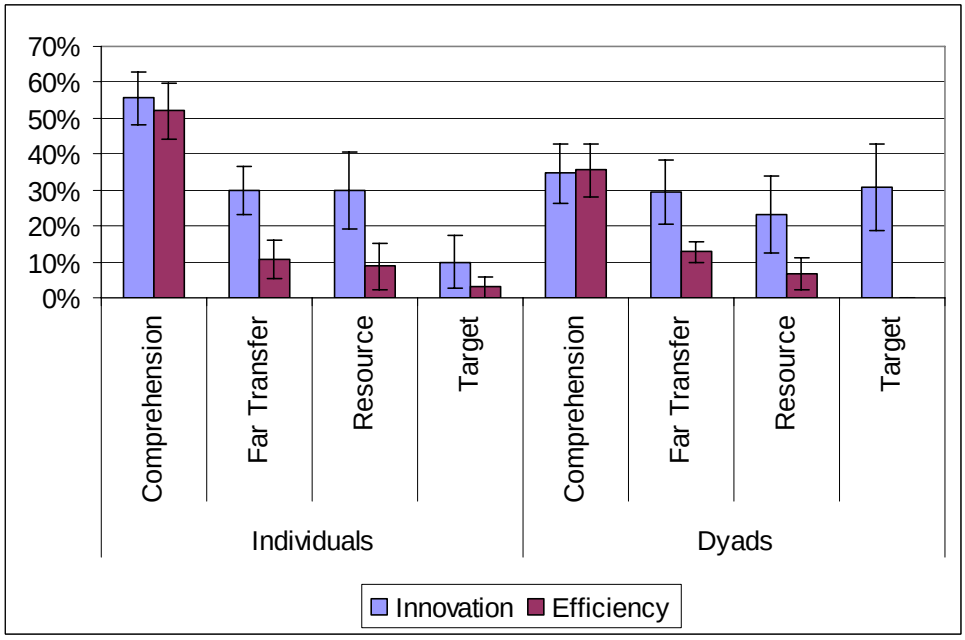

Figure 25. Innovation participants who acquired the relevant efficient knowledge (i.e. scored 50% or higher on the calculation measures) showed significant benefits to learning.

Discussion

The second study was designed to test the effects of Innovation versus Efficiency tasks on dyads' and individuals' learning. The general hypothesis was that as questions

went from basic calculations to conceptual understanding the benefits of Innovation for learning, especially in dyads, would become more apparent. This hypothesis was tested with a posttest that included three types of questions: 1) Calculation, 2) Comprehension, and 3) Far transfer. The far transfer questions followed the preparation for future learning (PFL) design in which a resource problem introduces new material, and a more difficult target problem builds upon it (Bransford & Schwartz, 1999; Schwartz & Bransford, 1998; and Schwartz & Martin, 2004). They examined how well participants were able to adapt their understanding of the chi-square formula to questions about inter-rater reliability. It was expected that all conditions would do well on calculations, Innovation would outperform Efficiency on comprehension questions, and Innovation would show the greatest advantage on the far transfer measures, especially Innovation dyads. As reviewed below, many of these hypotheses were supported.

Except where noted, the following results were in line with hypotheses. Participants in all conditions learned to calculate the chi-square statistic, averaging near 70% correct on the calculation measures. From these measures, Innovation did not appear any worse at the efficient application of knowledge even though they spent an equal time on the learning materials. This result goes against others' findings that tasks permitting greater exploration take longer to achieve the same educational outcomes (Anderson et al., 1989; Tennyson et al. 1985).

Innovation and Efficiency participants scored equally on the comprehension questions. Results for dyads on these questions were unexpected. Dyads performed significantly worse than individuals on these questions, regardless of condition. Video analyses should provide insight on this result in future work. For now, a tentative explanation comes from a post-hoc analysis of participants' time-on-task. The amount of time dyads spent reading the three Lessons pages was significantly less than the time individuals spent; 7.8 minutes versus 9.5 minutes.[15] Casual observation suggested that dyads were more intent on understanding, computing, and checking the calculations than reading the text carefully (perhaps because they were afraid to seem slow or because they did not want to lean over the text to study it closely and block the view of their partner).

[15] $t(71) = 2.36$, $p = .021$. The degrees of freedom here are less than previous analyses because times were not recorded in three instances due to experimenter error.

The transfer measures were hypothesized to show the greatest differences between conditions, and they did. Innovation showed less negative transfer than Efficiency, and this result was driven by dyads. Innovation dyads showed the least negative transfer while Efficiency dyads showed the most. This helps block one possible interpretation of the results. This interpretation is that the dyads did better on the problems requiring adapting the chi-square formula merely because they were more likely to slop around or try different things.

On the PFL target transfer problem, Innovation participants showed marginally greater ability to adapt their knowledge of the chi-square formula to make an adequate measure of inter-rater reliability. Innovation dyads also improved their performance from the resource problem to the target problem. This suggests the Innovation dyad condition best prepared students to continue learning. Taken together, the negative and positive transfer results support the notion that Innovation participants who worked together were the most flexible and adaptable with their knowledge *and* better understood its limits.

Comparisons between real groups and nominal groups provided the strongest support for the hypothesis that Innovation benefits learning most when done in groups. On the target transfer question, real Innovation dyads averaged 22.2% and Innovation nominal dyads averaged 14.7% under truth-wins assumptions. Real Efficiency dyads averaged 5.0% and Efficiency nominal dyads averaged 5.6%. Only one study that I am aware of has shown a task on which real groups outperformed nominal groups under truth-wins assumptions. Schwartz's (1995) study involved problem solving. I believe this dissertation study is the first to have found such an effect on learning.

<u>The Benefits of Efficient Knowledge for Learning from Innovation Activities</u>

While considerable support for the hypotheses was obtained, one pattern in the data ran counter to expectation. On every measure involving computation except the transfer measures, Innovation dyads performed worse than Innovation individuals while Efficiency dyads performed better than Efficiency individuals. This is important in conjunction with the finding that Innovation dyads did better on the transfer measures than Innovation individuals. It can clarify prevailing theories that suggest either Efficiency or Innovation tasks should be best for collaborative learning.

For example, some theorists like Steiner (1972) have suggested that highly demonstrable tasks should yield optimal results for groups relative to individuals. Others like Cohen (1994) have stated that ill-structured tasks should show the greatest advantages for groups. A simple answer to this paradox is that it depends on the measures. For basic procedural understanding, the Efficiency task appeared best for groups (and possibly individuals), while for deeper conceptual understanding, the Innovation task appeared best.

A better answer is possible, however; one that does not pit Innovation against Efficiency. In short, a certain amount of efficient knowledge is necessary to benefit from Innovative instruction, especially if working collaboratively. For instance, one can imagine that working well in a dyad requires some attention to coordinating activity with a partner. If the material is difficult for a given individual, then trying to coordinate with a partner may yield no benefit, or even worse, it may lead to errors and hinder learning. By contrast, for individuals who are comfortable with the material, coordinating activity with a partner can provide opportunities to see new approaches, make one's thinking visible through teaching, and thereby arrive at a deeper understanding. It seems likely that Efficiency tasks would make coordinating activity easier than Innovation activity because participants are given the common ground from which to work. Thus, for struggling students, Efficiency tasks might provide easier access to a most basic level of understanding and thereby an opportunity to benefit from working in a group. Innovation tasks (that are not supported by a teacher), by contrast, might simply overwhelm them.

According to these hypotheses, we would expect two additional results. First, the Innovation conditions should have more participants that failed to learn the most basic material (the calculations). Second, removing individuals that failed to acquire the efficient knowledge and re-analyzing the data would yield amplified benefits for Innovation because only those students with enough efficient background knowledge to gain from the Innovation experience would remain. Both of these predictions were supported. Ten Innovation participants, five individuals and five who studied in dyads, failed to score above 50% on the calculations. Only six Efficiency participants did likewise. The re-analysis of the data revealed that Innovation scores showed significant improvement relative to Efficiency scores after removing these individuals.

Given these results, an appropriate future experiment could present the chi-square materials as part of a high school or college statistics class to see whether and how these results might generalize from the laboratory. For example, participants could be randomly assigned to work in small groups or alone, and they could again be randomly assigned to Innovation or Efficiency conditions. To help ensure that all students had an opportunity to acquire the most basic understanding while avoiding potential experimenter effects, they could observe a pre-recorded lecture on the material after completing the course packet.

Finally, we gained two generative insights from this second study. First, the Innovation and Efficiency framework is not only useful for characterizing the transfer literature; it also appears to offer a tool for determining when different types of collaborative tasks are appropriate and what the benefits of each might be for learning. For example, efficiency tasks in groups might be particularly appropriate for students just beginning to gain exposure to a domain so that they can have a common ground from which to establish some understanding of basic tools with which to construct their future understanding. Innovation tasks, by contrast, seem appropriate for students who already have some efficient knowledge and need to be prepared for unknown tasks that build upon that knowledge.

These findings would not have been possible if appropriate measures were not available. Specifically, the second insight was that measures of deep understanding, especially the PFL measures, were sensitive to benefits of collaboration. For example, on the negative transfer measures, Innovation dyads did best while Efficiency dyads did worst across all four conditions. This suggests that perhaps Innovation encourages students to reflect on their solution as they proceed while Efficiency encourages them to proceed automatically. As a second example, the PFL measures revealed strong benefits of collaboration. On the target transfer problem, the real Innovation dyads exceeded the nominal ones (and scored four times higher than the Efficiency conditions). Surprisingly, with this measure, even the real Efficiency dyads outperformed the nominal Efficiency dyads (though this should be interpreted with caution, given the small sample from Efficiency that succeeded on the task). Perhaps previous studies of collaboration would show greater benefits for learning if they used PFL measures to assess their impact.

These unique results suggest that not only did Innovation support deep understanding for dyads; measures of deep understanding revealed otherwise hidden benefits of Innovation and collaboration.

CHAPTER 4: GENERAL DISCUSSION

This dissertation began with the question of what makes a good collaborative task for learning. Borrowing from recent developments in the transfer literature, two experiments with college students tested the effects of Innovation versus Efficiency tasks on participants' learning and collaboration. It was hypothesized that tasks with an Innovation component would naturally afford productive collaborations for learning whereas those with an Efficiency focus would not. Both studies provided support for this hypothesis while also revealing some surprising results. In addition, they highlighted methods of analysis that may have important roles in future studies of collaboration. Below, I will review these findings and then connect them to larger educational concerns, including how they relate to the development of adaptive expertise.

The first study examined dyadic interaction patterns resulting from Innovation and Efficiency versions of a concept-mapping task. While minimal learning differences were found between conditions, dyadic interactions looked quite different. As expected, the Efficiency version of the task led to partitioning of the task where one student became a reader and the other a scribe. The Innovation version led to greater turn-taking, explaining, and knowledge sharing during the map-making phase of the study. Interestingly, the concept-mapping teachable agent, Betty's Brain, may have prompted greater discussion and explaining by the Efficiency dyads during a map checking phase. This is possible because it had a feature that allowed participants to ask it to trace through the map and describe how different concepts in the map were related. This may have minimized some of the learning differences between conditions.

Using a novel method of analysis that allowed a fine-grained coupling of moment-to-moment interaction and learning, it was possible to see what types of behaviors were associated with successful recall initially and after a re-learning opportunity. Amount of turn-taking during map-making was not associated with learning for either condition. However, for Efficiency dyads, more turn-taking while questioning the teachable agent about their map was positively correlated with performance after the re-learning opportunity. Perhaps the most surprising finding was that explanations were associated with better recall for Efficiency participants but not for Innovation participants.

The greatest difference on explained versus unexplained links was observed on the initial recall test. The only measure that predicted performance after the re-learning opportunity for both conditions was the amount of agreement between partners on the first recall test (done alone). Greater agreement was associated with greater gains after re-learning.

The second experiment compared individuals' and dyads' learning of the chi-square formula. It found strong evidence for learning benefits of Innovation for individuals and especially for dyads. To my knowledge this is the only experiment to find evidence of real dyads surpassing nominal dyads modeled under truth-wins assumptions on a learning task. This effect was not caused by Innovation participants doing more random activity; they showed less negative transfer of the chi-square procedure than the Efficiency participants, especially the dyads.

These were the strongest results in favor of the hypothesis that Innovation tasks should naturally support productive collaboration. Interestingly, they would not have been found without the use of a preparation for future learning (PFL) assessment. The PFL assessment also revealed benefits of collaboration for Efficiency dyads compared to nominal dyads in that condition (although these results must be viewed with caution given that the numbers were very small). This suggests that studies of collaboration may have underestimated collaborative learning benefits by not assessing students' readiness to continue learning.

One surprising result from this second study was that Innovation dyads looked worse on non-PFL measures involving computation than their individual counterparts while Efficiency dyads looked better than their individual peers on these measures. By some accounts this would suggest that Innovation activities are best for dyads only when deep understanding is the goal. When procedural mastery is the goal, Efficiency activities might be best for dyads. These distinct outcomes could be reconciled in a better way, however. If we assume that a certain amount of efficient knowledge is necessary for productive innovation, then the dichotomy between deep understanding and mastery disappears. Removing those participants scoring below 50% on the efficient calculation measures, the benefits of Innovation were much stronger in general, and for Innovation dyads in particular.

This is an important point because it supports and extends previous theorizing about the use of the Innovation and Efficiency dimensions for promoting optimal learning. Specifically, Schwartz, Bransford, and Sears (2005) hypothesized that a combination of Innovation and Efficiency activities should place students on an optimal trajectory toward adaptive expertise (Hatano & Inagaki, 1986). Adaptive expertise might be considered the hallmark of a successful education. It moves beyond rapid execution of well-practiced procedures to include flexible adaptation of them to new contexts. It can also include changing the context to fit the tool. In other words, adaptive expertise is characterized by thorough understanding of both the problem and its solution, such that adaptations on either end can be made when desired. As an example, the NASA team and flight crew for the Apollo 13 mission that created a square peg for a round hole out of basic parts like tape and cardboard showed the ability to adapt different tools to an old problem (http://www.hq.nasa.gov/office/pao/History/SP-350/ch-13-4.html).

The studies in this dissertation add to the Innovation and Efficiency framework by indicating its relevance to collaborative learning. In particular, they suggested that Innovation tasks can promote productive collaborative behaviors. They also suggested a means by which these behaviors might be translated into learning benefits. I will address each of these points in turn.

First, why would Innovation tasks promote more collaboration? Without having the answer provided, participants must work together if they want to construct agreed-upon solutions. This would entail defining what constitutes an appropriate answer as well as how to reach it. For example, when asked to indicate which die is most fair, participants might first examine the distribution of data to see if they can estimate which die seems most fair. Then they could try to decide whether a high or low number for their solution should indicate greater fairness. Then they would need to try to find a formula that produces an acceptable result. In other words, at least two major elements, the outcome and the procedure, must be defined for Innovation tasks. This implies a need for considerable knowledge sharing. For efficiency tasks, the solution is given, so students only have to agree on whether they implemented it correctly or not. This allows for considerable variance in how much they decide to partition a task versus how much they decide to work together to make sure they understand how it works.

In addition to more knowledge sharing, how do Innovation tasks further promote learning for dyads compared to individuals? How could collaboration enhance Efficiency dyads' understanding of the basic material while promoting Innovation dyads' understanding of the more difficult concepts? Dyadic conditions should promote noticing of incommensurables. They could do this by each partner developing a different understanding of the material, and these would need to be reconciled. They could also notice the incommensurables in the contrasting cases more often than individuals because they would have twice as many people examining them. Innovation dyads should be more likely to notice the incommensurables built into the contrasting cases than Efficiency dyads because they work from the data rather than from the solution. These incommensurables are designed to highlight the key concepts in the solutions. Thus, dyads should notice incommensurables built into the materials more frequently, and they should generate and notice incommensurables dyad-members generate with respect to each other.

This analysis suggests a unit of analysis for future video coding. Perhaps through video analyses, the number of incommensurables that are noticed and reconciled between the Efficiency and Innovation dyads could be compared. I would expect the Innovation dyads to notice and reconcile more of the incommensurables built into the materials plus more incommensurables in their own perspectives. I would also expect that partners that worked to reconcile differences in perspective might also show more shared knowledge as revealed by more similar responses on the posttest. If this were the case, it might help explain why dyads with more shared knowledge in Study 1 showed greater gains on the second map re-drawing test. In essence, they would be the groups that worked to reconcile more incommensurables and had a more elaborated and stable schema to assist their recall.

Returning to the paradoxical performances of Innovation versus Efficiency dyads on the basic measures of understanding versus advanced measures, if we consider the role of efficient knowledge, we can reconcile these learning outcomes. If members of a dyad have acquired the relevant efficient knowledge and are willing to share it, then they should be more able to reconcile the incommensurables and come to a deeper understanding. If they do not have the efficient knowledge, then they might simply

confuse each other (e.g. by sharing misconceptions). The Innovation task does not provide a solution to work from; therefore, participants are more likely to fail to acquire the efficient knowledge, especially if a teacher is not available to help.

As mentioned before, one should not take this to mean that Innovation promotes understanding of difficult concepts at the expense of the most basic material. We learned from the analyses of participants who did not fail the calculation measures that they showed greater benefits from Innovation. If there were a tradeoff between learning the basic material and developing a more advanced understanding, then Innovation would show a decline or no gain. Instead, they showed a substantial gain suggesting that the Innovation materials led to a deeper understanding that built upon the more basic concepts. With teacher support, it should be possible for all students to benefit from those materials.

Taken together, one might consider these results to be relatively strong effects for an experiment that simply switched the order of the lessons pages and problems pages. On the face of it, that was the key manipulation, and a very simple one. Psychologically, however, the manipulation could promote vastly different views of the tasks. For Efficiency participants, receiving the answer and then a chance to practice it should push for speed and accuracy. This would be a very different mentality than the type of active sense-making we might expect from students receiving a page of contrasting cases and being told to try to find a solution that differentiates them in a way that matches intuition.

In line with this interpretation, a number of students in the Efficiency condition asked, "You want us just to write the chi-square formula?" upon seeing the first page of problems where it said, "What is your formula?" They seemed surprised, as if perhaps they had misinterpreted the task and were not simply meant to apply the formula they had just learned. This question did not fit a "drill-and-practice" schema for the activity. Innovation participants by contrast were more likely to say, "Oh, I get the pattern now. Each time we try to figure it out, then you give us the answer." They seemed more engaged in trying to understand not only the contents of the material, but also how it was structured.

In conclusion, the studies for this dissertation provided evidence that the combination of innovation activity with periodic efficient instruction (the lessons pages

and examples with solutions), led students to develop relatively thorough understanding of the chi-square formula in about 45 minutes. In addition, students who failed to acquire the efficient knowledge of how to calculate the chi-square procedure showed little benefit from Innovation activity whereas those who acquired that knowledge showed considerable advantage. An appropriate next step given this success on a single educational unit conducted in a laboratory is to consider how it might be applied in classrooms over extended periods of time. Such a test would allow us to see whether extended exposure to Innovation-and-Efficiency based activities moved students toward adaptive expertise at a compounding rate, a constant rate, or a decreasing rate.

An important point to consider leading up to this test would be what components are critical to a productive Innovation task. Not all inventing tasks are necessarily good for promoting adaptive expertise. From the first study to the second, we gained a better sense of what components might be important: 1) avoiding too many errors, 2) using contrasting cases to promote efforts to reconcile incommensurables, 3) choosing concepts for instruction that generalize, and 4) including PFL measures along with more standard ones.

While inventing appeared to be a key factor, something was missing in the first study. Periodic efficient instruction, corrective feedback, or other parameters were needed to keep participants from generating too many errors during the inventing phase. That is not to say that errors per se were harmful. We saw that when Innovation participants in the first study both missed a given link, they were very likely to recall it correctly after feedback. The danger of errors seems more likely to be due to the development of strong misconceptions or non-generative models that could interfere with subsequent instruction. They could also be an indicator that students lack the efficient knowledge to make an Innovation task educationally productive.

Providing contrasting cases that push students to notice and reconcile incommensurables seemed to promote learning benefits. Previous research suggested that working to reconcile incommensurables might be critical to learning from Innovation tasks (Schwartz, et al., in press; Sears, in press). This work extended these findings by showing their relevance in a collaborative context, as described above.

Finally, if a domain does not generalize, but is merely a list of disconnected facts, innovation seems unlikely to yield much benefit. This third point is related to the fourth, which is that PFL measures should be most sensitive to the learning benefits of innovation activities. For instance, if success on the calculation problems was the only outcome measure, the conditions would have appeared equivalent. By including measures of students' adaptability, we could see the benefits of Innovation. Without a domain that generalizes, PFL measures would be impossible.

Thus, the ability to create a PFL measure that clearly builds from component knowledge in the instructional unit and extends it to a new target domain could provide one litmus test for whether something has potential to be a productive innovation activity. For example, having instructed students in decimals and percentages, one could develop a resource problem that could lead students toward fractions or test their readiness to learn about them. This could be a type of PFL measure for Moss and Case's (1999) experimental math curriculum.

Future work should be able to precisely define the key features necessary for productive Innovation-and-Efficiency tasks. Through this knowledge, teachers should have one more resource for deciding when and how to use collaborative learning effectively in their classrooms. The ideal result would be if this particular tool promoted collaboration easily and naturally and with compounding benefits to learning when implemented over time.

Lesson 1 Practice Problems (page 1)

A Loaded Die?

Three dice are rolled, the first two were rolled 25 times, and the third was rolled 50 times. Find a number that ranks the dice in terms of how fair they are (or how much their distribution of scores differs from chance). Make sure your numbers support your rankings.

Case 1: Rolling a 5-sided Die. The number of times each result appears.

"1"	**"2"**	**"3"**	**"4"**	**"5"**
5	6	4	6	4

Rank: Most Fair Somewhat Fair Least Fair

Case 2: Rolling a 5-sided Die. The number of times each result appears.

"1"	**"2"**	**"3"**	**"4"**	**"5"**
4	4	10	3	4

Rank: Most Fair Somewhat Fair Least Fair

Case 3: Rolling a 5-sided Die. The number of times each result appears.

"1"	**"2"**	**"3"**	**"4"**	**"5"**
11	9	10	9	11

Rank: Most Fair Somewhat Fair Least Fair

What is your formula?

Lesson 1: The Chi-square Formula

People sometimes want to know whether two or more groups have different preferences, whether one drug is more effective than another at keeping people healthy, or whether a die is fair or not. In all of these cases, a distribution of numbers in a frequency matrix is being examined to determine whether it differs significantly from chance (or expectation) or not. A common formula used in statistics to address these questions is the chi-square formula. **The chi-square formula and its uses will be the focus of the three lessons in this packet.**

The Chi-square formula is:

$$\chi^2 = \Sigma\,(E - O)^2/E$$

E stands for the "expected value."
O stands for the "observed value."

We use the chi-square formula to determine how different a distribution is from chance. In cases of two (or more) variables, we use the chi-square to tell us if the variables are independent or if they interact. The larger the chi-square statistic, the less likely the variables are independent or the distribution is random.

EXAMPLE

If we have 3 months, each with the same number of days, we might wonder whether babies are born equally frequently during each month or whether there is a deviation from chance.

	April	**June**	**September**
Births	50	20	20

So, with this distribution, we can calculate the chi-square statistic.
First, find E:
We expect of the 90 births over the 3 months (i.e. 50 + 20 + 20 = 90), that each month would have 30 births if the distribution was random or fair. Thus, E = 30 for each cell in this distribution.

O = 50, 20, and 20 for each cell.

Thus,
$\chi^2 = [(30 - 50)^2/\,30] + [(30 - 20)^2/\,30] + [(30 - 20)^2/\,30] = 600/30 = 20.$
If there had been 30 births per month, our chi-square statistic would have been 0. In other words, larger values indicate distributions that are more different from chance.

Lesson 1 Practice Problem (page 2)

In this next problem, we have two dice (pretending that 3-sided dice exist). A casino owner rolls the dice and records how many times each combination of numbers appears because they want to see if the dice are fair or magnetized.

Case 1: Please provide a number to indicate how fair or magnetized the dice are.

	"1"	**"2"**	**"3"**
"1"	14	2	5
"2"	2	3	6
"3"	6	7	9

Circle one: Magnetized Fair

Your answer should be:
E = 6 for each cell (b/c 1/3 chance if dice are fair).
Chi-square = 64/6 + 16/6 + 1/6 + 16/6 + 9/6 + 0 + 0 + 1/6 + 9/6 = 116/6 = 19 and 1/3

Lesson 2 Practice Problem (page 1)

Do children differ from adults more than pigs differ from horses!?! Scientists recorded which food was chosen by each organism and their results are shown below. In each case, please indicate whether you think the preferences are different or not, and provide a number to indicate how different the preferences are.

Case 1: Sweets Preferences. Do children and adults differ in the sweets preferences?

	Candy	**Chocolate**
Children	6	14
Adults	16	4

Circle one: Yes No

Case 2: Barnyard fare. Do pigs and horses differ in their food preferences? Yes or No?

	Apples	**Oranges**
Pigs	14	6
Horses	16	4

Circle one: Yes No

Question: What happens if the Expected Value is not given to us by the variables in play?

Lesson 2: Chi-Square when E must be estimated

When E is not given by the variables in the problem (such as from fair dice where we would expect each number on a six-sided die to appear $1/6^{th}$ of the time), then it must be estimated. Also, the chi-square statistic measures interactions, not main effects, when two variables are involved.

We use the following formula to estimate the expected value of a cell:

$$E = (RowTotal * ColumnTotal)/GrandTotal$$

This formula makes sense because we are computing probabilities and then multiplying by the grand total to get a value. Specifically,

E = (RowTotal/GrandTotal)*(ColumnTotal/GrandTotal)*GrandTotal

In other words, the expected value of a given cell is estimated as the probability of being in that row and that column simultaneously (i.e. the probability of being in that cell) times the grand total (i.e. the number of observations).

EXAMPLE

Medical researchers might want to compare the effects of two drugs to see if one is better at keeping people healthy, so for each participant they recorded what drug they took and whether they stayed well or got sick.

	Drug A	**Drug B**
Stayed Well	24	6
Got Sick	1	9

Find E for each cell:
E for the top left cell is: 30*25/40 = 18.75
E for the top right cell is: 30*15/40 = 11.25
E for the bottom left is: 25*10/40 = 6.25
E for the bottom right is: 15*10/40 = 3.75

Thus, the chi-square is:
$\chi^2 = [(18.75 - 24)^2/ 18.75] + [(11.25 - 6)^2/ 11.25] + [(6.25 - 1)^2/ 6.25] + [(3.75 - 9)^2/ 3.75]$

In other words, the two drugs differed considerably in their effectiveness.

Lesson 2 Practice Problem (page 2)

Case 1: Researchers wanted to determine if best and worst grade assignments differed by era.

	1900's	**1920's**	**1940's**
Best	4	20	6
Worst	6	5	9

Your answer should be:
Chi-square = 4/6 + 25/15 + 9/9 + 4/4 + 25/10 + 9/6 = 8 and 1/3

Lesson 3 Practice Problem (page 1)

On three tasks, people were surveyed to see whether those with high levels of experience had a different preference from those with low levels of experience. On the second task, twice as many people answered the survey. On the third task, many more people answered the survey. Which task showed the greatest differentiation of people with high versus low levels of experience?

On the first task, the data looked like this:

Participant	Experience	Preference	Participant	Experience	Preference
1	High exp	A	11	Low exp	B
2	High exp	A	12	Low exp	B
3	High exp	A	13	Low exp	A
4	High exp	A	14	Low exp	B
5	High exp	B	15	Low exp	B
6	High exp	A	16	Low exp	B
7	High exp	A	17	Low exp	B
8	High exp	B	18	Low exp	B
9	High exp	A	19	Low exp	B
10	High exp	A	20	Low exp	A

Researchers sorted it into this matrix.

Case 1: How much does experience affect preference of A versus B?

	A	**B**
High	8	2
Low	2	8

Rank: Most Different Somewhat Different Least Different

Case 2: Holding the E to O ratio constant.

	A	**B**
High	16	4
Low	4	16

Rank: Most Different Somewhat Different Least Different

Case 3: Holding the difference between E and O constant.

	A	**B**
High	106	94
Low	94	106

Rank: Most Different Somewhat Different Least Different

Question: What happens to the chi-square statistic as the sample size changes?

Lesson 3

The chi-square formula can take different forms that are good for showing different aspects of it.

For example,

$$\chi^2 = \Sigma\ (E - O)^2/E$$

This form emphasizes that when the difference between the observed and expected is the same, the chi-square value decreases with increasing sample size.

Algebraically, another form of the equation can be derived:

$$\chi^2 = \Sigma\ E[1 - (O/E)]^2$$

This less common form of the chi-square formula is useful in highlighting that when the ratio of the observed to the expected is the same, the chi-square statistic increases as the sample size increases.

EXAMPLE

What is the effect of level of experience on preference for A or B?

	A	**B**
High	24	6
Low	6	24

E = 15 for each cell.
Thus, the chi-square is: 4*(81/15) = 4*3*(9/5) = 21.6

Lesson 3 Practice Problem (page 2)

What is the effect of level of experience on preference for A or B?

	A	**B**
High	315	285
Low	285	315

Your answer should be:
Chi-square = 4*(15^2/300) = 3

APPENDIX B: THE POSTTEST

Posttest

Question 1: Drug test effectiveness.

Test A and Test B differ substantially in how expensive they are, so doctors wanted to find out whether they differed in how accurately they diagnosed patients with a particular disease. Do Test A and Test B differ significantly in their ability to predict who has the disease?

	Yes	No
Test A	10	10
Test B	6	14

Question 2. Why do you divide by E in the chi-square formula?

Question 3. How does the sample size affect the chi-square statistic?

Question 4: Evaluating Mechanic's Diagnoses

Two mechanics looked at 50 cars and wrote down their diagnosis of the problem for each car. They calculated their agreement on their diagnoses as 38 agreements. From that, they reasoned that they matched on 38 of 50 cases, or 76%. Can you help them develop a better formula for how much their diagnoses matched? To do this, you should show: 1) how much they should match just from chance alone, and 2) how much they did match beyond chance?

Car Mechanics' diagnoses of car problems.

	Mechanic 1:		
Mechanic 2:	Engine	Alternator	Battery
Engine	17	3	0
Alternator	2	12	6
Battery	1	0	9

Question 5: Imagine you have two fair dice, one with 2 sides, the other with 3. Do you think these dice magnetized, and why?

	"1"	**"2"**	**"3"**
"1"	6	18	6
"2"	9	12	9

Circle one: Magnetized Not Magnetized

Question 6: Evaluating a coding scheme

A researcher would like a procedure for determining how well two raters agree. Specifically, find a formula to show how much beyond chance the raters **agree**.

Chemists' ratings of flame color for various unknown elements being burned.

Element	**Rater 1**	**Rater 2**
1	Red	Red
2	Red	Red
3	Red	Red
4	Red	Red
5	Red	Red
6	Red	Red
7	Red	Orange
8	Orange	Red
9	Orange	Red
10	Orange	Orange
11	Orange	Orange
12	Orange	Orange
13	Orange	Orange
14	Orange	Orange
15	Green	Green
16	Green	Orange
17	Green	Green
18	Green	Green
19	Green	Green
20	Green	Green
21	Blue	Green
22	Blue	Blue
23	Blue	Blue
24	Blue	Blue
25	Blue	Blue
26	Yellow	Yellow
27	Yellow	Yellow
28	Yellow	Yellow
29	Yellow	Yellow
30	Yellow	Yellow

Question 7. When should someone use the chi-square formula?

END. Thank You!

REFERENCES

Anderson, J., Conrad, F., & Corbett, A. (1989). Skill acquisition and the LISP tutor. Cognitive Science, 13, 467-505.

Andersson, J., & Rönnberg, J. (1995). Recall suffers from collaboration: Joint recall effects of friendship and task complexity. Applied Cognitive Psychology, 9, 199-211.

Antil, L., Jenkins, J., Wayne, S., & Vadasy, P. (1998). Cooperative learning: Prevalence, conceptualizations, and the relation between research and practice. American Educational Research Journal, 35, 419-454.

Azmitia, M. (1996). Peer interactive minds: Developmental, theoretical, and methodological issues. In P.B. Baltes & U. M. Staudinger (Eds.), Interactive minds: Life-span perspectives on the social foundation of cognition. Cambridge: Cambridge University Press.

Barron, B. (2000). Problem solving in video-based microworlds: Collaborative and individual outcomes of high-achieving sixth-grade students. Journal of Educational Psychology, 92, 391-398.

Barron, B. (2003). When smart groups fail. The Journal of the Learning Sciences, 12, 307-359.

Biswas, G., Schwartz, D., Bransford, J., & the Teachable Agents Group at Vanderbilt (2001). Technology support for complex problem solving: From SAD environemtns to AI. In K. Forbus and P. Feltovich (Eds.), Smart machines in education: The coming revolution in educational technology.

Bransford, J.D., & Schwartz, D.L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad and P. D. Pearson (Eds.), Review of Research in Education, 24, 61-100. Washington, D.C.: American Educational Research Association.

Chiang, C.-L., & Guo, C.-J. (1999). Different ways to reach agreement and to handle disagreement in science group discourse. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Boston, MA, March 1999).

Cohen, E.G. (1994). Restructuring the classroom: Conditions for productive small groups. Review of Educational Research, 64, 1-35.

Coleman, E. (1998). Using explanatory knowledge during collaborative problem solving in science. The Journal of the Learning Sciences, 7, 387-427.

Dennis, A., & Valacich, J. (1993). Computer brainstorms: More heads are better than one. Journal of Applied Psychology, 78, 531-537.

Hatano, G., & Inagaki, K. (1986). Two courses of expertise. In H. Stevenson, H. Azuma, & K. Hakuta (Eds.), Child development and education in Japan (pp. 262-272). New York, NY: Freeman.

Hertz-Lazarowitz, R. (1989). Cooperation and helping in the classroom: A contextual approach. International Journal of Educational Research, 13, 113-119.

Johnson, D., & Johnson, R. (1999). Making cooperative learning work. Theory into Practice, 38, 67-73.

King, A. (1999). Discourse Patterns for Mediating Peer Learning. In A. O'Donnell and A. King (Eds.), Cognitive perspectives on peer learning (pp. 87-115). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Lamm, H., & Trommsdorf, G. (1973). Group versus individual performance on tasks requiring ideational proficiency (brainstorming): A review. European Journal of Social Psychology, 3, 361-388.

Laughlin, P., Zander, M., Knievel, E., & Tan, T. (2003). Groups perform better than the best individuals on letters-to-numbers problems: Informative equations and effective strategies. Journal of Personality and Social Psychology, 85, 684-694.

Moss, J. & Case, R. (1999). Developing children's understanding of the rational numbers: A new model and an experimental curriculum. Journal for Research in Mathematics Education, 30, 122-147.

Novick, L.R. (1988). Analogical transfer, problem similarity, and expertise. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 510-520.

O'Donnell, A. (1999). Structuring dyadic interaction through scripted cooperation. In A. O'Donnell and A. King (Eds.), Cognitive perspectives on peer learning (pp. 179-196). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Olivera, F., & Straus, S. (2004). Group-to-individual transfer of learning: Cognitive and social factors. Small Group Research, 35, 440-465.

Phelps, E., & Damon, W. (1989). Problem solving with equals: Peer collaboration as a context for learning mathematics and spatial concepts. Journal of Educational Psychology, 81, 639-646.

Schwartz, D. (1995). The emergence of abstract representations in dyad problem solving. Journal of the Learning Sciences, 4, 321-354.

Schwartz, D., Bransford, J. & Sears, D. (2005). Efficiency and innovation in transfer. In J. Mestre (Ed.), Transfer of learning from a modern multidisciplinary perspective. CT: Information Age Publishing.

Schwartz, D., & Martin, T. (2004). Inventing to prepare for learning: The hidden efficiency of original student production in statistics instruction. Cognition & Instruction, 22, 129-184.

Schwartz, D., Martin, T., & Pfaffman, J. (2005). How mathematics propels the development of physical knowledge. Journal of Cognition and Development, 6, 65-88.

Schwartz, D., & Sears, D. Deepening our understanding of processes and outcomes of collaborative learning: Defining and investigating more and less productive efforts. Symposium presented at the American Educational Research Association (AERA) 2004 Conference, San Diego, CA, April 2004.

Schwartz, D., Sears, D., & Chang, J. (in press). Reconsidering prior knowledge. To appear in M. Lovett and P. Shah (Eds.), Thinking with Data. Mahwah, NJ: Erlbaum.

Sears, D.A. (in press). Effects of innovation versus efficiency tasks on recall and transfer in individual and collaborative learning contexts. To appear in Proceedings of the International Conference of the Learning Sciences (ICLS) 2006.

Slavin, R. (1990). Cooperative learning: Theory, research, and practice (1st ed.). Englewood Cliffs, New Jersey: Prentice Hall.

Slavin, R. (1996). Research for the future: Research on cooperative learning and achievement: What we know, what we need to know. Contemporary Educational Psychology, 21, 43-69.

Steiner, I. (1972). Group process and productivity. New York: Academic Press.

Tennyson, R., Park, O., & Christensen, D. (1985). Adaptive control of learning time and content sequence in concept learning using computer-based instruction. Journal of Educational Psychology, 77, 481-491.

Vollmeyer, R., Burns, B., & Holyoak, K. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. Cognitive Science, 20, 75-100.

Vygotsky, L. (1978/1932). Mind in society: The development of higher psychological processes. (M. Cole, V. John-Steiner, S. Scribner, and E. Souberman, Eds.). Cambridge, MA: Harvard University Press.

Webb, N. (1983). Predicting Learning from student interaction: Defining the interaction variables. Educational Psychologist, 18, 33-41.